A expansão de conjuntos de co-hipónimos a partir de colecções de grandes dimensões de texto em Português

Luís Sarmento

Objectivo

- Pesquisa de co-hipónimos. Ou seja:
  - elementos que possuem um hiperónimo comum
  - instâncias de uma mesma “classe”
  - elementos “semelhantes”
- A partir de um ***conjunto*** de exemplos / *sementes*
  - [maçã, laranja] -> [papaia, banana, amora…]
  - [azul, laranja] -> [verde, preto, carmim, …]
- Extraindo informação de quantidades massivas de texto não anotado morfo-sintacticamente

Motivações

- Práticas:
  - auxílio à construção de recursos léxico-semânticos para a língua Portuguesa do tipo WordNet.
- Teóricas:
  - estudo de medidas e problemas associados à semelhança e à analogia semântica
- Outras (secundárias para a presente investigação):
  - pesquisa de hiperónimos por fusão de dados
  - construção de *topic maps* para navegação em colecções
  - expansão que expressões de pesquisa em motores

Proposta de dois métodos simples

- Método 1:
  - pesquisa de contextos semelhantes
  - segue uma aproximação mais tradicional
  - segue definições teóricas
  - aprendizagem levemente-supervisionada
- Método 2:
  - pesquisa de elementos coordenados
  - segue uma aproximação mais prática
  - envolve preocupações de performance
  - método simples, “clusterting” (ou nem isso…)

A nossa fonte de texto: BACO

- BAse de Co-Ocorrências:
  - Base de dados gigante gerada a partir de uma recolha Web: 6 GB de texto / 1000M átomos
- Várias tabelas:
  - Frases: 35M
  - n-gramas (sequências 1, 2, 3, *4* palavras): *273*M
  - co-ocorrências: 780M
- Muita redundância tentando obter representações que permitam pesquisas mais rápidas (?)

A tabela de 4-gramas

- Pesquisar directamente sobre 35M frases / 6GB de texto pode ser muito lento.
  - limitações do próprio índice, muitas leituras em disco…
- Tabela de 4-gramas: (p1, p2, p3, p4, f, d)
  - permite pesquisar texto no contexto de uma janela de 4 palavras de uma forma mais eficiente: há agregação!
- Vamos assumimos a limitação do contexto e utilizar esta tabela como base para as pesquisas
- Vamos pesquisar candidatos com 1 palavra apenas

Método 1: Contextos 3 Palavras

- Obervação:
  - Elementos semelhantes ocorrem em contextos léxicais semelhantes
- “a compota de morango/laranja”:
  - “a compota de X” --> X? Fruto? Vegetal?
- “o gelado de morango/baunilha”:
  - “o gelado de X” --> X? Fruto? Alimento?

Método 1: Algoritmo

- Para os elementos do conjunto S = {s_1,s_2,..s_n}
  - Procurar contextos de 3 palavras: c_1,c_2, c_3, s_i
  - contexto anterior: “bias” para pesquisa nomes (?)
- Para contextos “representativos” c_{i1},c_{i2}, c_{i3}, X:
  - Procurar elementos X_i que ocorram no mesmo contexto
- Contextos “representativos” co-ocorrem com:
  - L sementes: garante especificidade
  - menos de M_{max} elementos do léxico (250): impede sobre-generalização
- O algoritmo “aprende” contextos “representativos”.

Método 1: Resultados

#	Conjunto inicial	L	# C	Resultado
1	amarelo, vermelho, azul	3	4 4	verde (26), branco (22), preto (19), cinza (14), castanho (14), ... violeta (6), prata (5), *escuro* (4), dourado (4), *fe* (4), ... *pele* (4), *cores* (3), ... *liso* (2), *carvalho* (2), marrom (2), ... *terra* (2), *iluminado* (2), *54* (2), *brasil* (2), *pobre* (2)
2	granito, mármore, basalto	3	3	*betão* (2), *vidro* (2), *papel* (2), *pedra* (2), *madeira* (2), *material* (2)
3	whiskey, rum, gin	2	6	vodka (4), vinho (3), tequila (3), porto (3), cerveja (2), licor (2), sumo (2), coca-cola (2), verdelho (2), whiskie (2), tinto (2), aguardente (2), conhaque (2), *jack* (2), *neoplast* (2), uisque (2), água (2), *coca* (2), *plástico* (2), champanhe (2), cachaça (2), champagne (2)
4	porto, braga, aveiro	3	2 1 0	coimbra (141), lisboa (137), *vila* (126), *castelo* (115), leiria (110), viseu (110),... almada (51), guimarães (49),... *cidade* (5) ... régua (2), *avaliação* (2), *recrutamento* (2), *municípios* (2), *editorial* (2), gorazde (2), *gás* (2), *coliseu* (2), alvor (2), inhambane (2)

Método 1: Comentários

- O topo da lista é quase sempre constituido por candidatos verdadeiros (+)
  - excepção para Conj. 2: palavras “raras”
- Alguns candidatos estão incompletos (-)
  - “coca” “jack”…: limitação intrínsica
- Sensibilidade a certas ambiguidades (-)
  - “uma garrafa de X”: X = “vodka” V X = “vidro”
- Algoritmo elegante mas pouco eficiente (-)
  - Tempo de execução: 1m até 25m iBook G4

Método 2: Elementos Coordenados

- Observação:
  - Elementos semelhantes ocorrem no contexto de coordenações, isto é são enumeráveis / listáveis
  - “gelado de morango, laranja e limão”
  - “prova de matemática e física ou química”
- A obtenção de elementos semelhantes pode ser feita procurando elementos que se coordenam com os elementos do conjunto semente.

Método 2: Preparação dos dados

- Pré-selecção de tuplos para ***alguns*** padrões
- Geração de tabela auxiliar par(x,y)
- Tabelas mais pequena
- Pesquisas mais rápidas

Padrão	Tuplos recolhidos
, X e Y	179415
, X ou Y	25.203
, X , Y	399.013
X , Y ,	428.746
X , Y e	202.619
X, Y ou	28.941
X e o Y	112.746
X e a Y	153.477
X ou o Y	6.824
X ou a Y	13.083
X , o Y	207.068
X , a Y	271.152
Total	**2.028.287**

Método 2: Algoritmo

- Para os elementos do conjunto de sementes S = $\{s_1, s_2, ..s_n\}$, procurar em par(X,Y):
  - X para os quais Y = s_i
  - Y para os quais X = s_i
  - … contabilizando número de contextos distintos / pares para os quais co-ocorrem
- Eliminar elementos “ruidosos” da lista de candidatos
- Algoritmo simples

Método 2: Resultados

#	Conjunto inicial	Resultado
1	amarelo, vermelho, azul	verde (48), preto (39), branco (38), laranja (28), rosa (23), cinza (18), castanho (18), violeta (13), cinzento (11), negro (11), lilás (11), *cor* (11), … cores (6), … transparente (4), azulão (4), *champanhe* (4), *sol* (4), *céu* (3), castanha (3), mediterrâneo (3), alaranjado (3), *camisa* (3), *claro* (3), púrpura (3), âmbar (3)…
2	Granito, mármore, basalto	*madeira* (9), pedra (8), calcário (7), *bronze* (7), *cimento* (6), *vidro* (5), xisto (5), *cantaria* (4), *tijoleira* (4), ardósia (3), arenito (3), barro (3), gesso (3), calcários (3), travertino (3), *tabaco* (2), *ouro* (2), *cellano* (2), (2), quartzo (2),…
3	whiskey, rum, gin	Vodka (8), conhaque (3), tequila (3), rumpi (2), *tabaco* (2), *limão* (2), calvados (2), *creme* (2), bourbon (2), *curaçau* (2), *anseios* (2), *açúcar* (2), brandy (2),…
4	Porto, braga, aveiro	lisboa (31), coimbra (28), leiria (24), gaia (23), viseu (22), setúbal (22), évora (21), guimarães (21), guarda (19), *minho* (19),… *algarve* (16),… *madeira* (14), …*portugal* (13), … cidade (13), … *sporting* (12), *benfica* (12) … (várias centenas de candidatos)

Método 2: Comentários

- Mais profícuo: são gerados muitos mais candidatos (+/-?)
- O topo da lista é mais uma vez povoado por elementos "verdadeiros" (+)
  - excepção para Conj. 2: palavras "raras" mas melhor que o anterior
- Muito rápido (+):
  - Após criação de tabela auxiliar (~1h) pesquisas 5-10s
- Ruidoso: obtem elementos relacionados mas por vezes a relação é apenas contextual (sintagmática) (-)
  - "whiskey" e "tabaco"

Breve Discussão

- Ambos os métodos:
  - possuem diferentes problemas: possibilidade de validação cruzada;
  - apresentam problemas quando os dados são esparsos - problema típico;
  - usam medidas de mérito muito simples baseadas no número de co-ocorrências distintas;
  - podem melhorar a sua precisão com filtros relativamente simples, quer estatísticos quer linguísticos
  - podem desde já auxiliar a construção semi-automática de recursos léxico-semânticos.

Questões incómodas…

- O que são de facto elementos “semelhantes”?
  - Forte dependência do domínio / contexto
    - Granito, mármore: em construção civil? em escultura? em geologia?…
  - Forte dependência do âmbito semântico
    - Azul, Amarelo: são “cores” ou são “cores primárias”?
- Como detectar diferentes domínios / âmbitos?
  - E pesquisar elementos “semelhantes” e conformidade?

Comentários Finais

- É possível extrair alguma semântica em português com métodos simples e “ingénuos”
  - Há muita margem para evolução.
- O problema dos dados esparsos necessita de emprego de métodos de suavização
- A descoberta / determinação automática de contextos / âmbitos é um problema interligado mas fundamental
- É possível a aplicação prática dos métodos apresentados com optimizações simples

DEMONSTRAÇÃO