Revisão de "Automatic Acquisition and Expansion of Hypernym Links" Emmanuel Morin e Christian Jacquemin

Por Luís Sarmento
Linguateca – Pólo do Porto

Introdução

- Objectivo
  - facilitar a construção automática de tesauros a partir de texto "livre"
  - descobrir relações entre Unidades Terminológicas
- Extracção de informação / Conhecimento
  - Extracção de Relações Semânticas
    - Principalmente Hiperonímia/Hiponímia
    - Relações "Fusão" e "Produz"
- Envolve
  - Extracção de Terminologia (indirectamente)
  - Pesquisas/Inferência de PLS
  - Aproximação Linguística mas também estatística (-)

Introdução

- Muitas semelhanças com o Corpógrafo!
  - Ponto de vista técnico
  - Filosofia de Pesquisa
- Mais sofisticado que o Corpógrafo na:
  - Geração de variantes terminológicos
  - Obtenção de PLS: Padrões Léxico-Sintácticos
- Corpógrafo uma ferramenta de utilizador:
  - Potencial de produção massiva de tesauros

Arquitectura

- Sistema envolve 3 ferramentas:
  1. ACABIT (Daille 1996):
     - extracção de UT multi-palavra
  2. FASTR (Jacquemin, 1996):
     - obtenção de variantes UT
     - muito interessante -> podia ser desenvolvido para PT
  3. Prométhée (Morin, 1999):
     - Estruturação das UT (i.e.: obtenção de relações)
     - Pesquisa de evidências usando PLS
     - O artigo foca essencialmente esta ferramenta
- Corpora -> hierarquia de UT

O Prométhée

- O Prométhée é:
  - sistema para a extracção de informação de relações semânticas entre UT a partir de corpora, usando padrões léxico-sintácticos (trad.)
- 2 funcionalidades base:
  - Aquisição de PLS (++)
  - Extracção de pares relacionados usando bancos de LSP

A arquitectura

- 3 Módulos:
  1. Pré-processador lexical:
     - Atomização, análise morfo-sintáctica, lematização
     - NP, acrónimos e sequências de NP são identificadas
  2. Shallow Parser e Classificador:
     - Responsável pela extracção de PLS
  3. Extractor de Informação
     - Responsável pela aquisição de novos pares conceptualmente ligados (por uma dada relação)

A descoberta de PLS

- Esta é a parte mais interessante do sistema
- Tem 7 passos, mas resumidamente:
  - a partir de um conjunto de UT que conhecemos e que estão relacionadas por uma dada relação, procurar os contextos em que essas UT ocorrem e tentar generalizar
- Ex. arquétipo:
  - Entrada: p1(banana, fruta) e p2(carro, veículo)
  - Pesquisa em corpora de contextos, seguida de generalização
  - Saída: padrão "NP1 é um tipo de NP2".

Os 7 passos

1. Seleccionar uma relação semântica representativa. Ex: Hiperonímia
2. Criar uma lista com os vários pares representativos da relação. Construção manual a partir de corpora ou usando um recurso já existente (tesauro base).
   - Ex: p(neocortex, "área vulnerável") *
3. Procurar frases onde os pares (lematizados) ocorram e recolher padrões:
   - Ex: "Foram encontrados danos neuronais em certas *áreas vulneráveis* tal como o *neocortex*" **
   - "Encontrar NP em certas NP tal como NP"

Os 7 passos

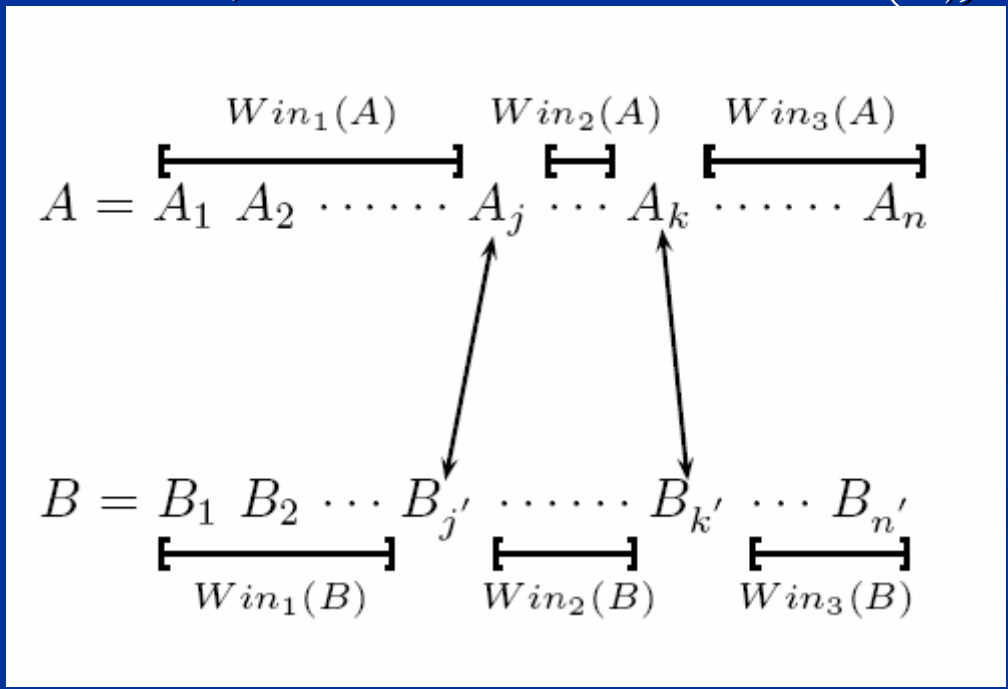
4. Generalização dos padrões.
   - Para todos os padrões recolhidos, encontrar generalizações (vamos ver com mais cuidado!!!)
5. Validação dos padrões por um perito
   - O papel do perito pode ser essencial em certas relações / domínios
6. Pesquisa de novos pares relacionados usando os padrões
7. Validação por um perito dos pares obtidos

A generalização de padrões

- Par 1: HIPER(vulnerable area,neocortex):
  - "Neuronal damage was found in the selectively vulnerable areas such as neocortex, striatum, hippocampus and thalamus"
  - PLS gerado: NP find in NP such as LIST
- Par 2: HIPER(complication, infection)
  - Therapeutic complications such as infection, recurrence, and loss of support of the articular surface have continued to plague the treatment of giant cell tumor
  - PLS gerado: NP such as LIST continue to plague NP

A generalização de padrões

- Consideraremos v. abstractas dos PLS:
  - $A = A_1A_2...A_j...A_k...A_n$ com $HIPER(A_j,A_k)$
  - $B = B_1B_2...B_j...B_k...B_n$ com $HIPER(B_j,B_k)$

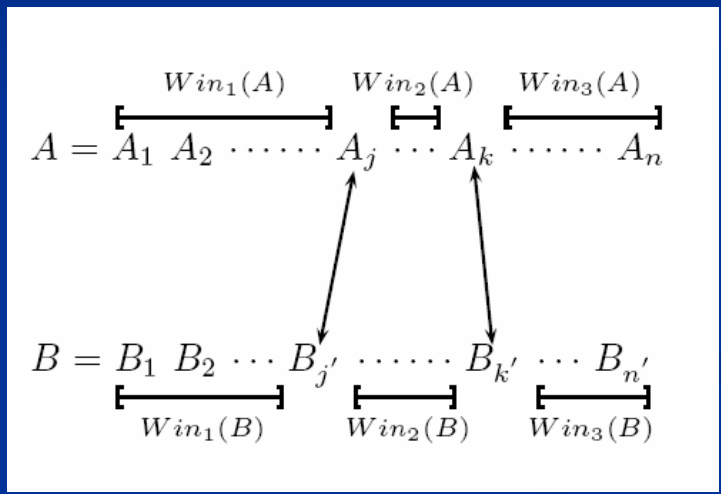


A generalização de padrões

- Função de semelhança PLS SIM(A,B)
  - Hipótese (Isomorfia Sintáctica): Se duas expressões léxico-sintácticas A e B representam o mesmo padrão, os itens A_j e B_j e os itens A_k e os itens B_k (os "pontos de fixação") têm a mesmo função na frase.
  - Assim, podemos concentrar-nos apenas nas janelas que estes pontos de fixação criam para testar as semelhanças dos PLS (Corolário?)

Função de Semelhança

- Sim(A;B) = $\Sigma_{i=(1,3)}$ Sim(Win$_i$(A);Win$_i$(B))



  - Em que Sim(Win$_i$(A);Win$_i$(B)) foi definido como uma função da maior sub-sequência comum (MSC).
    - Formalismo um pouco confuso pela recursividade

A maior sub-sequência comum

- Consideremos duas strings X e Y:
  - X[1...m] = $X_1X_2...X_m$, comprimento m
  - Y[1...n] = $Y_1Y_2...Y_L$, comprimento n
- Sejam:
  - X[1...k] e Y[1..L] os prefixos de comprimento k e L das strings X e Y respectivamente
  - c[k,L] o comp. da MSC entre X[1...k] e Y[1..L]
- Então:

$$c[k,l] = \begin{cases} 0 & \text{if } k = 0 \text{ or } l = 0 \\ c[k-1,l-1]+1 & \text{if } k,l > 0 \text{ and } X_k = Y_l \\ \max(c[k,l-1], c[k-1,l]) & \text{if } k,l > 0 \text{ and } X_k \neq Y_l \end{cases}$$

A generalização de padrões (cont.)

1. Comparam-se todos os padrões dois a dois usando a função de semelhança
2. Cria-se uma matriz de semelhanças PxP
   - $M[i,j] = Sim(P_i,P_j)$
3. É executado o agrupamento das expressões semelhantes
4. Para cada grupo é escolhida a expressão que possui o menor desvio padrão com as restantes do grupo
5. Para o exemplo:

NP find in NP such as LIST

NP such as LIST continue to plague NP

NP such as LIST

Avaliação do processo de descoberta

- Corpus usado Agro-Alim (Fr: ?? àtomos)
- 40 pares hiper-hipo definidos manualmente. Ex:
  - fruits tropicaux : bananes
  - cations : sodium
  - arbres : chênes
  - cereales : ble
  - fruits : orange
  - fruits : kiwi
  - legume : carotte
  - legume : concombre
  - sucre : saccharose
  - huiles : huile de soja
- Nesta fase os autores usaram UT de uma palavra apenas (acelerar a convergência?)
- Podia ter-se recorrido a um tesauro para obter os pares

O processo de descoberta

- 40 instâncias permitiram gerar 11 PLS:
  - {deux|trois...|2|3|4...} NP1 (LIST2)
  - {certain|quelque|de autre...} NP1 (LIST2)
  - {deux|trois...|2|3|4...} NP1 : LIST2
  - {certain|quelque|de autre...} NP1 : LIST2
  - {de autre}? NP1 tel que LIST2
  - NP1, particulièrement NP2,
  - {de autre}? NP1 comme LIST2
  - NP2 {et|ou} de autre NP1
  - NP1 et notamment NP2
- E agora vamos "pescar" novos pares!

Avaliação do processo de descoberta

- Avaliar os pares relação descobertos pelo Prométhée:
- Como habitualmente P, R e F:
  - N_{tot}: Nº pares existentes (manual)
  - N_{cor}: Nº pares correctos
  - N_{err}: Nº pares correctos
  - $P = (N_{cor})/(N_{cor}+N_{err})$
  - $R = N_{cor}/N_{tot}$
  - $F = 2\ (P \times R)\ /\ (P+R)$
- Calculados estes valores para todos os 11 PLS

Avaliação do processo de descoberta

- Resultados:
  - Pmédia = 82% (média-alta)
  - Rmédia = 56 % (média-baixa)

Pattern	# pairs of terms	P.	R.	F-M.
{deux|trois...|2|3|4...} NP_1 ($LIST_2$)	270	84%	56%	68%
{certain|quelque|de autre...} NP_1 ($LIST_2$)	212	87%	52%	65%
{deux|trois...|2|3|4...} NP_1 : $LIST_2$	241	79%	51%	62%
{certain|quelque|de autre...} NP_1 : $LIST_2$	116	84%	47%	60%
{de autre}? NP_1 tel que $LIST_2$	210	86%	70%	76%
NP_1, particulièrement NP_2,	4	100%	36%	53%
{de autre}? NP_1 comme $LIST_2$	90	69%	64%	67%
NP_1 tel $LIST_2$	36	90%	67%	76%
NP_2 {et|ou} de autre NP_1	17	59%	65%	62%
NP_1 et notamment NP_2	6	70%	43%	53%
chez le NP_2, NP_1,	14	62%	66%	64%
Total	1216	82%	56%	66%

Avaliação do processo de descoberta

- Detecção de relações sub-especificadas:
  - (“característica”,”dureza”)
  - “característica” é demasiado genérico. Que “característica”?
  - “elemento”, “espécie”,”factor”, etc... (Termos Secundários?)
- Pares muito específicos do assunto/amostra:
  - (“ambiente ideal”, “Malásia”)
  - “...crescem num *ambiente ideal* tal como na *Malásia*”
- Confusão hiperonímia/meronímia
  - (“tronco”,”membro”)
  - “...parte do *tronco* em particular os *membros*...”
- Possíveis causas:
  - Ambiguidade morfo-sintáctica entre Adj e PP
  - Dificuldade de detecção de NP (sobreposição, inclusão, coordenação...)

Comentários Rápidos

- Valores de P e R nada surpreendentes
- Foram "apenas" obtidos os padrões "triviais"
  - Manualmente talvez fosse melhor
  - O Corpógrafo tem padrões menos triviais!
- Contudo, a hiperonímia não é habitualmente expressa só por estas construções...
  - ao contrário de certas relações funcionais (ex: causa-efeito)
- ... estando grande parte da informação da hiperonímia em domínios técnicos implícita na morfologia.
  - "célula" : "neurónio" vs. "célula" : "célula da glia"
- Em todo o caso, é um método automático e potencialmente "portável" para outras relações!

Outras experiências

- Teste com a "relação" "fusão":
  - "Dixons Group Plc said shareholders at a special meeting of Cyclops Corp approve the previously announced *merger of Cyclops with Dixons*."
- Processo ligeiramente diferente (Reuters Corpus!):
  - Carregamento do Prométhée com 2 PLS base "arquétipos"
  - Extracção de pares usando os padrões base
  - Pesquisa de mais PLS usando os pares anteriores: +5
  - Pesquisa de novos pares usando os 7 padrões
  - 101 novos pares P = 92% mas R=??
- Claro que é uma relação muito do domínio e o corpus é de domínio específico e linguagem (hiper?) controlada

Outras experiências ainda...

- Teste com a relação "produz":
  - produz("nome_da_empresa","produto")
  - Permite "distribui", "vende", "fornece", etc.
- A aproximação foi a mesma que a anterior
- Mas não houve convergência:
  - A explicação foi que os pares gerados não chegam para garantir a convergência
  - Claro que o domínio é *muito maior* do que no caso anterior
  - Talvez um catálogo de empresas permitisse o "boot-strap"

A expansão dos Hiperónimos

- Usando os PLS 11 anteriores
- 1216 tuplos hiperonímia
  - 26.2% entre UT's multi-palavra
  - 23.5% entre UT's simples
  - 50.3% entre uma UT multi-palavra e outra simples
- O tuplos obtidos referem-se aos possíveis de detectar no contexto de uma frase
- Representam apenas uma fracção dos existentes
- Solução: tentar propagar as relações!

A expansão dos Hiperónimos

- Ideia: encontrar relações entre UT's compostas, a partir das relações conhecidas para UT's de 1 palavra
- Supondo que conheço:
  1. hiper(fruta, maça)
  2. UT(sumo de fruta) e UT(sumo de maça)
  3. Relacionados(nectar,sumo)

  Podemos gerar por propagação (ou não):
  1. hiper(sumo de fruta, sumo de maça)
  2. Relacionados (nectar de fruta, sumo de maça)
  3. REL (N1 de fruta, N2 de maça)
     - com N1 e N2 semanticamente ligados
- Verifica-se ser possível propagar relações quando as UT compostas são UT Variantes

O que são UT Variantes?

- São UT que partilham semelhanças estruturais e semânticas, e que por isso permitem propagar relações
- 3 tipos de UT's Variantes a considerar:
  - Sintácticas
  - Morfo-sintácticas
  - Semânticas
- Há algoritmos que permitem detectar as UT variantes!
- Detecção (ou geração) de variantes:
  - FASTR (Jacquemin 99)
  - Bastante sofisticado!

Variantes Sintácticas

- As palavras da UT original mantêm-se na UT variante, mas a estrutura relativamente à UT original é diferente.
  1. **Coordenação:** combinação de dois termos com a mesma "cabeça": "frutos frescos ou secos"
  2. **Modificação:** inserção de um modificador: "resistência [mecânica] do manípulo"
  3. **Sinapse:** remoção de algumas palavras função: cultivo de bananas / cultivo das bananas

Variantes Morfo-Sintácticas

- O conteúdo da UT original ou de uma sua variante morfológica é encontrado na UT variante. A estrutura sintáctica também é alterada.
  1. **Nome-Nome:** "semente de alfarrobeira"/ "semente de alfarroba"
  2. **Nome-Verbo:** "produção de enzimas" / "produzir enzimas"
  3. **Nome-Adjectivo:** "produção de fruta", "produção frutícola"

Variantes Semânticas

- Relações semânticas (sinonímia, hiperonímia) encontradas entre palavras da UT original e da UT variante:
  1. **Semânticas (Puras):** "farinha de trigo" / "farinha de milho" (co-hiperonímia trigo/milho)
  2. **Sintáctico-Semânticas:** (1) mais possibilidade de Modificação/Sinapse: "grãos duros de milho" / "grãos de trigo"
  3. **Morfo-sintáctico-semânticas:** (2) + aplicar transformações também às palavras morfologicamente relacionadas: "açúcar residual" / "resíduo de glicose"

Restrições à produção de Variantes Semânticas

- Consideremos duas UT's possivelmente variantes:
  - UT = $w_1 w_2$ e UT' = $w_1' w_2'$
- Observação:
  - w1 e w1' podem estar semanticamente relacionados
  - w2 e w2' podem estar semanticamente relacionados
  - Não implica UT e UT' semanticamente próximos!!
- É necessário impor algumas restrições:
  1. Isomorfia sintáctica
  2. Isomorfia semântica
  3. Relação semântica holística

Isomorfia sintáctica

- As palavras relacionadas têm de ocupar posições similares nas UT, como:
  - Cabeças
  - Argumentos
  - Modificadores
- Ex:

  processo de elaboração !ISO_SIN elaboração de um método

  "processo" e "método" são sinónimos mas

  "elaboração" não se encontra na mesma posição sintáctica

Isomorfia semântica

- As palavras relacionadas têm de ter significados semelhantes em ambas as UT's.
- EX.:

  análise da distribuição estatística

  !ISO_SEM

  análise de divisão estatística

  Apesar de haver ISO_SIN
- É necessário lidar com polissemia
- Parece bem mais difícil (outras colocações ajudariam?)

Relação semântica holística

- Temos de verificar que as UT's completas (i.e. os holons) são semanticamente equivalentes.
- EX:
  - inspecção alimentar
  - não é semanticamente equivalente a
  - controlo alimentar
  - apesar de inspecção e controlo serem próximos
- Parece-me ainda mais difícil e dependente do domínio.

Variantes semânticos: definição

- Duas UT multi-palavra w1w2 e w1'w2' são variantes semânticos se 3 condições se verificarem:
  1. Existe uma relação semântica ***S*** entre w1 e w1' e/ou w2 e w2'. O elemento não relacionado é idêntico ou é morfologicamente relacionado
  2. w1 e w1' são cabeças enquanto que w2 e w2' possuem papéis temáticos semelhantes
  3. w1w2 e w1'w2' possuem **a mesma** relação semântica ***S***

Variantes semânticos: Corolário

- Se
  - duas UT compostas T e T' forem consideradas (hipoteticamente) variantes semânticos (usando as regras)
  - e T e T' são estruturados a partir das UT simples w1/w1' ou w2/w2'
  - e w1/w1' e w2/w2' que verificam (1) e (2)
- Então
  - podemos assumir que as UT Variantes T e T' partilham a mesma relação semântica S que um dos seus constituintes (w1/w1' ou w2/w2'), e por isso poderemos propagar a relação
- Algoritmo:
  1. procurar UT variantes
  2. propagar entre as UT variantes as relações semânticas dos seus constituintes (se as conhecermos)

Pesquisa de Variantes

- Baseada nas regras e restrições anteriores foi criada uma meta-gramática de produção de variações
- 110 meta-regras!
  - 16 para termos Adj N
  - 22 para termos N Adj
  - 72 para N Prep N
- Ex:

  N1 Prep N2 -> M(N1; N) Adv? Adj? Prep Art? Adj? S(N2)

  composição do fruto -> compostos químicos da semente
- Usada no FASTR
- Versões para EN, DE, JP, SP. E ... PT? Bom projecto!

Pesquisa de Variantes

- Recapitulando a ideia base:
- Conhecemos a relação
  - "semente" é parte do "fruto"
- Conhecemos e somos capazes detectar UT variantes:
  - composição do fruto
  - compostos [químicos] da semente
  - Usando: N1 Prep N2 -> M(N1; N) Adv? Adj? Prep Art? Adj? S(N2)
- Pela definição de UT variantes podemos assumir:
  - "compostos químicos da semente" é parte do "composição do fruto"

Resultados da pesquisa de variantes

- Foram encontrados 1.143 variantes
  - Sintácticos: 495
  - Semânticos: 584
  - Morfo-sintácticos: 64
- 981 Variantes foram considerados correctos – 85.5%
- Resultados parciais variáveis:
  1. Sintácticos – P = 93.9 %
  2. Semânticos "puros" – P = 86.2 %
  3. Morfo-sintácticos – P = 71.2 %
  4. Semânticos com variações sintácticas/morfológicas – P = 73.8 %

Para (3) e (4) muitas UT variantes são semanticamente diferentes das originais

Propagação de Relações

- Pretende-se projectar hierarquias de UT 1 de palavra em hierarquias de UT multi-palavra.
- Recordemos que o FASTR, precisou de poder testar relações semânticas entre palavras simples
- **2 opções** disponíveis para este efeito:
  1. Tesauros [AGROVAC]
  2. Conjuntos de hiperónimo e os seus co-hipónimos gerados pelo Prométhée a partir do corpus (tipo synsets do wordnet) organizados hierarquicamente
     1. Ex: fruta – maçã – Cartland

Propagação de Relações

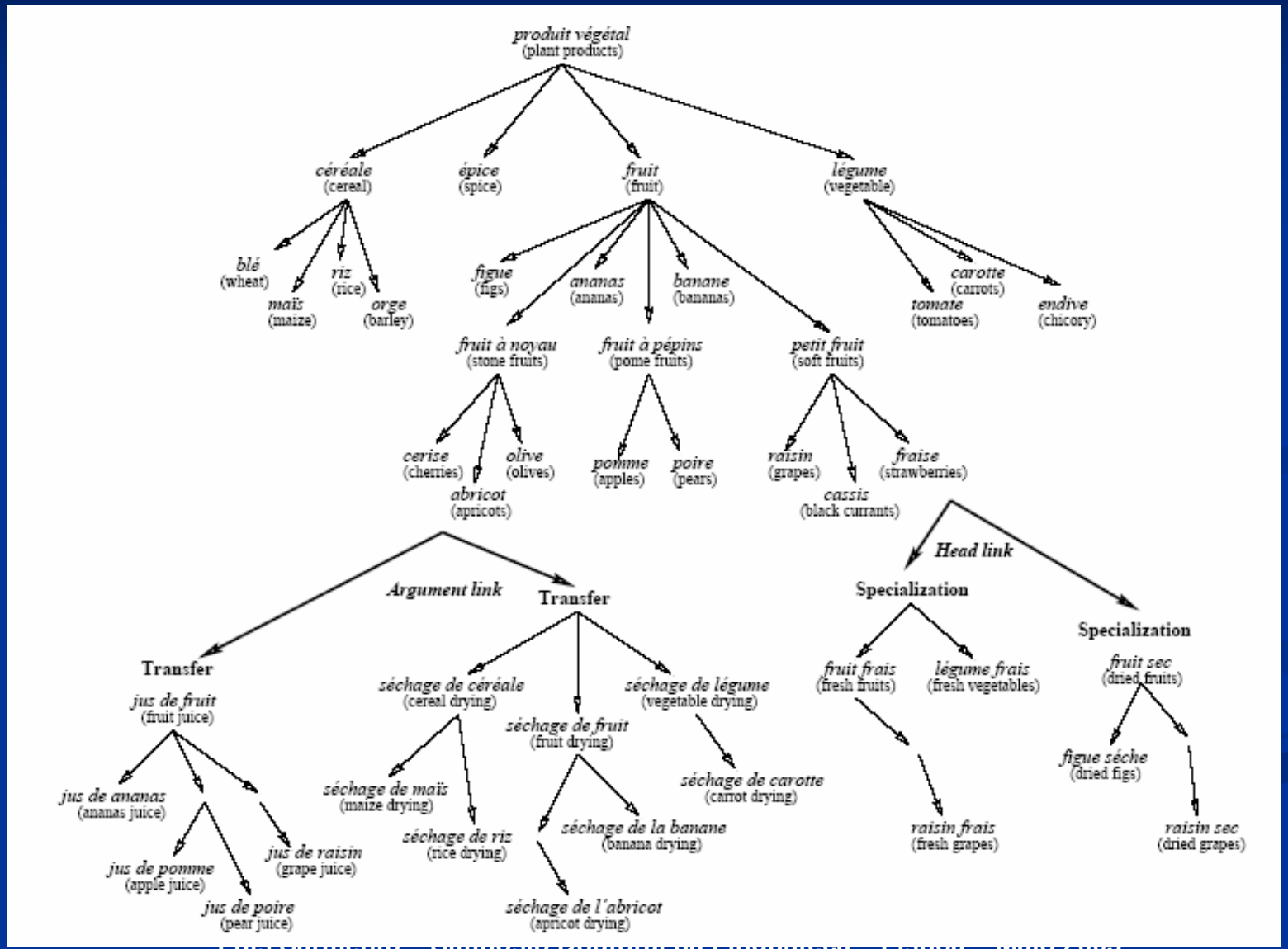
Classes	Hypernyms and co-hyponyms
arbres (trees)	*arbre, bouleau, chêne, érable, hêtre, orme, peuplier, pin, poirier, pommier, sapin, épicéa*
éléments chimiques (chemical elements)	*élément, calcium, potassium, magnésium, manganése, sodium, arsenic, chrome, mercure, sélénium, étain, aluminium, fer, cadium, cuivre*
céréales (cereals)	*céréale, maïs, mil, sorgho, blé, orge, riz, avoine*
enzymes (enzymes)	*enzyme, aspartate, lipase, protéase*
fruits (fruits)	*fruit, banane, cerise, citron, figue, fraise, kiwi, noix, olive, orange, poire, pomme, pêche, raisin*
olives (olives)	*fruit, olive, Amellau, Chemlali, Chétoui, Lucques, Picholine, Sevillana, Sigoise*
pommes (apples)	*fruit, pomme, Cartland, Délicious, Empire, McIntoch, Spartan*
légumes (vegetables)	*légume, asperge, carotte, concombre, haricot, pois, tomate*
polyols (polyols)	*polyol, glycérol, sorbitol*
polysaccharides (polysaccharides)	*polysaccharide, amidon, cellulose, styrène, éthylbenzène*
protéines (proteins)	*protéine, chitinase, glucanase, thaumatin-like, fibronectine, glucanase*
sucres (sugars)	*sucre, lactose, maltose, raffinose, glucose, saccharose*

E.g.: fruta – maçã – Cartland

Propagação de Relações: 2 tipos

1. Projecção por Transferência
   - as ligações entre 2 conceitos representados por UT de 1 palavra são transferidas para UT multi-palavra localizadas noutro ponto da hierarquia.
   - hiper(fruta, maçã) → hiper(sumo de fruta, sumo de maçã)
2. Projecção por Especialização
   - as ligações entre 2 conceitos representados por UT de 1 palavra são transferidas em paralelo para UT's multi-palavra que representam conceitos especializados de cada um dos conceitos base
   - hiper(fruta, figo) → hiper(frutos secos, figo seco)
     - (e "amêndoa"? -> não há problema porque "amêndoa seca" não é UT)

Propagação de Relações



Propagação de Relações: observações

- Estamos a trabalhar em domínios específicos o que reduz ambiguidades / polissemia / homografia
  - Por isso, estas propagações não devem gerar demasiados pares espúrios
- UT multi-palavra são normalmente específicas e menos frequentes em corpora
  - Por isso, nem todas as propagações possíveis podem depois ser encontradas / verificadas em corpora
- Esta técnica deve ser entendida como uma forma semi-automática de extensão de glossários

Avaliação das Projecções (1)

- Propagação das relações obtidas pelo Método 1: Prométhée + corpora
- Dos 1216 pares entre UT's de 1 palavra, seleccionaram-se 89
- Propagadas 584 novas relações entre UT multi-palavra

Classes	Specialization			Transfer		
	# Occ.	Correct occ.	P.	# Occ.	Correct occ.	P.
trees	1	1	100.0%	3	3	100.0%
chemical elements	8	4	50.0%	101	99	98.0%
cereals	6	1	16.7%	76	65	85.5%
enzymes	3	3	100.0%	29	20	69.0%
fruits	32	20	62.5%	214	172	80.4%
olives	4	1	25.0%	10	8	80.0%
apples	4	1	25.0%	16	12	75.0%
vegetables	3	2	66.7%	3	3	100.0%
polyols	0	-	-	0	-	-
polysaccharides	3	1	33.3%	13	11	84.6%
proteins	0	-	-	8	6	75.0%
sugars	13	11	84.6%	34	26	76.5%
Total	77	45	58.4%	507	425	83.8%

Avaliação das Projecções (1)

- Transferências são mais habituais: 507 vs. 77
- Certas classes são muito produtivas:
  - Elementos químicos, cereais e frutos
  - Termos muito genéricos
- Outras classes são muito pouco produtivas:
  - Polióleos, proteínas: UT muito específicos, ou melhor têm regras próprias de obtenção de variantes e de projecção
  - Árvores e vegetais: pouco frequentes nos corpora

Especialização vs. Transferência (1)

- Projecção por Especialização:
  - Menos frequentes: 77 tuplos
  - Precisão média relativamente baixa: 58.4 %
  - Desvio padrão elevado
- Projecção por Transferência:
  - Bastante frequentes: 507
  - Precisão média relativamente elevada: 83.8 %
  - Desvio padrão reduzido
- Precisão global do processo de Projecção: 80.5%

Avaliação das Projecções (2)

- Propagação das relações identificadas por consulta a um tesauro [AGROVAC: 15,800 UT's]
- Teste à robustez do Prométhée
- Seleccionaram-se 168 UT's de 4 tópicos: cultivo / anatomia vegetal / produtos vegetais / sabores
- Projectadas 371 novas relações entre UT multi-palavra

Classes	Specialization			Transfer		
	# Occ.	Correct occ.	P.	# Occ.	Correct occ.	P.
cultivation	0	-	-	0	-	-
harvesting	0	-	-	3	1	33.3%
pruning	0	-	-	0	-	-
plant anatomy	0	-	-	0	-	-
plant reproductive organs	5	1	20%	26	20	76.9%
inflorescences	0	-	-	0	-	-
flowers	0	-	-	0	-	-
leaves	0	-	-	3	3	100.0%
stems	2	-	-	0	-	-
tissues	1	1	100.0%	2	2	100.0%
plant products	8	6	75.0%	26	14	53.8%
cereals	5	1	20.0%	78	65	83.3%
spices (plant products)	0	-	-	1	1	100.0%
fruits (plant products)	3	3	100.0%	53	41	77.4%
stone fruits	2	1	50.0%	19	14	73.7%
pome fruits	7	3	42.9%	32	17	53.1%
soft fruits	8	7	87.5%	45	30	66.6%
oilseeds	0	-	-	31	25	80.6%
vegetables	4	3	75.0%	3	2	66.7%
flavourings	0	-	-	0	-	-
condiments	0	-	-	0	-	-
spices (flavourings)	0	-	-	1	1	100.0%
Total	45	26	57.8%	326	236	72.4%

Avaliação das Projecções (2)

- Resultados semelhantes ao caso anterior
- Projecção por Especialização:
  - Menos frequentes: 45 tuplos
  - Precisão média relativamente baixa: 57.8 %
- Projecção por Transferência:
  - Bastante frequentes: 326
  - Precisão média relativamente elevada: 72.4 %
- Precisão global do processo de Projecção: 70.6%

Corpora vs. Tesauros

- Os resultados da utilização da informação de tesauros são significativamente inferiores
- Os tesauros utilizam hierarquias mais profundas o que aumenta a distância média entre co-hipónimos
- Aparentemente, é mais difícil verificar a relação entre UT simples usando tesauros (não há transitividade?) e por isso são validadas menos relações entre TU variantes

Conclusões

- Um método que permite propagar algumas relações semânticas fáceis de estabelecer automaticamente (i.e. entre UT 1 palavra) a um conjunto de UT compostas cuja detecção de relações é mais complexa
- Boa precisão
- Passível de ser portado para PT
- Bom método para criar glossários
- Exemplo da riqueza da "extracção" de terminologia

Conclusões

- Um método automático interessante para obtenção de PLS
- Precisão razoável e talvez tenha poucas vantagens relativamente à aproximação manual
- Seria interessante pensar:
  1. como generalizar o método de propagação para quaisquer UT para partir de qualquer tamanho
  2. como enriquecer o PLS partindo das novas relações