Tópicos de Aprendizagem Automática aplicada a PLN

Aquisição Automática de Recursos Semânticos

Luís Sarmento

Simpósio Doutoral da Linguateca

3, 4 de Outubro 2006

1. Introdução e Motivação

AA?

- Este tutorial não vai valar sobre Aprendizagem Automática (AA) "em geral"
  - área tão vasta que é infrutífero falar dela numa hora
- Vamos focar em aplicações da AA que sejam relevantes em tarefas de PLN, nomeadamente aquelas que necessitam de **recursos semânticos** cuja construção manual seja **difícil/custosa** como:
  - Léxicos categorizados semanticamente
  - Bancos de regras de análise
- Nestes casos, a AA pode ser a única forma prática de construção de tais recursos em larga escala

Aquisição (Semi-)Automática de Recursos Semânticos

- Em muitos caso é simplesmente inviável construir de forma totalmente manual, recursos como:
  - bases de dados lexicais / léxicos-semânticos
  - dicionários e tesaurus
  - ontologias de domínio específico ou gerais
  - bancos de regras/ padrões
- A aprendizagem automática pode ajudar:
  - gerar recursos de raíz para posterior verificação manual (quando possível)
  - a expandir automaticamente recursos inicialmente criados à mão

Problemas com a construção manual de recursos

- um recurso produzido manualmente levanta sempre problemas de **consistência**
- a construção manual de recursos, devido ao enorme esforço envolvido, tende a limitar a sua **abrangência**
- dificuldades em manter a **representatividade** dos recursos:
  - efeitos perversos para o processamento automático
- **viés semântico** introduzido pelo esforço de conceptualização:
  - a visão do mundo do "criador" do recurso

O nosso enquadramento…

- métodos que permitam a construção de recursos semânticos cuja integração e utilização num sistema de análise seja quase imediata (ex: para REM)
- métodos não-supervisionados, ou apenas levemente supervisionados:
  - essencialmente *agrupamento* e *bootstrapping*
- Ou seja:
  - **não** vamos falar de como se infere um analisador sintáctico a partir de um corpus anotado

Mais especificamente…

- Métodos que são capazes de aprender sobre texto (anotado ou não) elementos como:
  - **grupos de palavras** semelhantes ou pertencentes à mesma classe semântica
  - **relações léxicais entre palavras** como a sinonímia / antonímia ou paráfrase léxical
  - **padrões de relacionamento ou de extracção** para classificação e para EI
  - **relações genéricas entre conceitos**: relações taxonómicas tradicionais (hiponímia, meronímia) e outro tipo de relações de *domínio aberto*.

Neste tutorial

- Iremos apenas abordar métodos que permitem "aprender" **grupos de palavras** semelhantes ou pertencentes à mesma classe semântica
  - As restantes 3 opções serão alvo de outros tutoriais

Métodos e Sistemas

- Iremos essencialmente rever sistemas que ilustrem os métodos
- Não faremos uma profunda abordagem teórica aos métodos mas focaremos em tarefas práticas
- Espera-se inspirar os presentes para a inclusão de métodos de AA nos seus próprios sistemas

2. A aquisição de grupos palavras

3 exemplos ilustrativos

A aquisição de grupos de palavras: o que é?

- **Objectivo:** conseguir agrupar em classes as palavras de um léxico (substantivos, verbos, nomes), normalmente também eles inferidos / aprendidos a partir de corpora.
- Os **agrupamentos de palavras** obtidos são em muitos casos suficientemente interessantes para serem considerados um "produto final".
- Noutros casos, podem ser um produto intermédio no processo de inferência de **etiquetas de classificação** e descoberta de **relações entre conceitos / classes**

A ideia base…

- A maior parte dos trabalhos nesta área baseia-se na "Hipótese Distribucional do Significado":
  - as palavras que são "semelhantes" semanticamente ocorrem em contextos "semelhantes" ou com distribuições "semelhantes"
- As grandes questões que se colocam são:
  - o que se entende por "contexto semelhante"?
  - o que se entende por "distribuição semelhante"?

3 Exemplos:

- Agrupamento por Representantes
  - (Lin & Pantel 2002)
- Aquisição usando Modelos de Grafos
  - (Widdows & Dorow, 2002)
- Etiquetando Classes Semânticas
  - (Pantel & Ravichandran, 2004)

2.1. Agrupamento por Representantes

- (Lin & Pantel 2002) propõem a aproximação de Agrupamento por Representantes (Clustering by Committee - CBC)
- os autores observam que, em inglês, contextos de duas palavras parecem ser suficientes para caracterizar semanticamente uma dada palavra
- os contextos podem ser usados para obter a representação vectorial de cada palavra
  - Modelo do Espaço Vectorial / Vector Space Model

A representação escolhida…

- Se considerarmos p_i que co-ocorre com vários contextos *anteriores* e *posteriores* de duas palavras:
  - e_k e_j ... (à esquerda)
  - ...d_l d_m (à direita)
- a representação vectorial da palavra p_i, $[p_i]$, pode ser obtida calculando o valor da *Associação* A entre a p_i e cada um dos contextos c_i
  - $[p_i] = [A (p_i , c_1); A (p_i, c_2); A (p_i, c_3) ... A (p_i, c_n)]$
- A Medida de Associação usada foi a Informação Mútua (Church & Hanks, 1990) à qual se adiciona um factor de correcção aos "casos raros"

Medida de Semelhança

- A partir da representação vectorial, torna-se simples proceder ao calculo da "semelhança" de duas palavras fazendo uso de medidas tradicionais de semelhança vectorial
  - os autores usaram o Coeficiente do Coseno

O algoritmo de agrupamento CBC

- Os autores propõem o algoritmo de agrupamento CBC:
  - **Fase 1: Determinação dos elementos mais semelhantes.** Na primeira fase são calculados os k elementos mais semelhantes a cada palavra a agrupar p_i.
  - **Fase 2: Pesquisa de Representantes (Comittees).** Na segunda fase, o algoritmo tenta recursivamente encontrar agrupamentos muito coesos ao longo do espaço de semelhança.
  - **Fase 3: Atribuiçao dos Elementos aos Agrupamentos.** Nesta fase, depois de determinados os Agrupamentos Representantes são calculados os seus centroides. A formação dos Agrupamentos finais é feita associando cada elemento do conjunto ao centroide mais próximo

Procedimento Experimental

- O algoritmo foi experimentado e comparado com outros métodos de agrupamento, em várias configurações:
  - K-Means, Buckshot, Bissecting K-Means, Chameleon / Average-link, Complete-Link, Single-Link
- Um corpus de 1GB foi processado usando o parser Minipar que permite a selecção de certos contextos sintácticos em torno de substantivos
  - ou seja nem todos os bigramas foram utilizados como contextos para a representação vectorial!

Procedimento Experimental

- A qualidade dos agrupamentos resultantes foi avaliada relativamente ao conteúdo do próprio WordNet
  - foram construídos dois padrões de teste, um incluindo 13403 palavras e 202 classes (S_{13403}) e outro apenas com 3566 palavras e com 150 classes (S_{3566})
- Calculou-se o grau de sobreposição entre os agrupamentos obtidos por cada um dos algoritmos e os conjuntos de teste
  - concluiu-se que o algoritmos proposto, o Clustering By Committe, obteve desempenhos superiores
- O método permitiu encontrar elementos que são relevantes nos agrupamentos encontrados, e que não tinham sido incluídos no WordNet:
  - um exemplo de como se podem expandir recursos já existentes!

2.2. Aquisição usando Modelos de Grafos

- Em (Widdows & Dorow, 2002) é apresentado um modelo de grafos para lidar com os problemas que a polisemia traz ao processo de agrupamentos de palavras
- os autores assumem do princípio que duas palavras são semelhantes se possuirem contextos e distribuições de contextos semelhantes:
  - novamente a Hipótese Distribucional

Construção de um grafo semântico

- Partem do BNC (marcado gramaticalmente)
  - das várias relações possíveis focam na coordenação: {Nome} {e / ou} {Nome}
- Cada Nome (substantivo ou nome próprio) é considerado um nó num grafo
- Se dois nomes aparecem coordenados no BNC é marcada uma aresta entre os nós correspondentes
  - peso da aresta é proporcional ao número de vezes que a palavras co-ocorrem coordenadas
  - Para evitar arestas espúrias, foi considerado que as arestas têm de ter um peso mínimo

O Grafo

- Após vários testes foi considerado que se deveria apenas manter apenas as ***n*** arestas com maior peso para cada nó
  - exclui ligações para palavras mais raras, sacrificando abrangência por precisão
  - decisão de ***n*** puramente ad-hoc (!!)
- O grafo resultante produzido a partir do BCN:
  - 99545 nós e 587475 arestas,
  - 400000 nomes do BNC o grafo engloba ~25%

Obtendo "semelhantes"

- Sejam A = $\{a_1, a_2, ..., a_n\}$ um conjunto de n nós do grafo e seja $N(A)$ o conjunto de nós vizinhos aos elementos de A
- o nó b mais semelhante aos elementos de A é um elemento não pertencente a A que possua maior proporção de arestas para elementos de $N(A)$
- o nó b mais semelhante é aquele que maximiza a relação:

Algoritmos robusto!

- Partindo do princípio que se começa com dois nós semente "orange", "banana", e "apple" e independentemente de existir no grafo uma ligação (possivelmente forte) entre "apple" e "novell", "novell" nunca será adicionado ao agrupamento inicial porque não possui ligações com outros dos vizinhos desse agrupamento para além de "apple":
  - Isto permite identificar palavras polisémicas cujos agrupamentos a que pertencem são *disjuntos* isto é pertencem a partes do grafo fracamente conexas

A avaliação deste método

- comparar os grupos gerados para certas classes semânticas, com classes que se podem obter usando a informação hierarquica do WordNet:
  - crimes, locais, ferramentas, meios de transporte, instrumentos musicais, roupas, doenças, partes do corpo, disciplinas académicas e comida

Algumas conclusões

- Tendo em conta os resultados obtidos, os **autores** afirmam que:
  - o método permite uma melhor performance no agrupamento de palavras semelhantes do que propostas anteriores
  - este tipo de aproximação "data driven", que utiliza apenas a informação de PoS de um corpus base de grandes dimensões, atinge resultados muitos melhores do que aqueles obtidos usando um corpus de menores dimensões embora alvo de análise linguística mais profunda.

2.3. Etiquetando Classes Semânticas

- os trabalhos anteriores produziam classes (ou simples agrupamentos) de palavras sem no entanto serem capazes de etiquetá-los, com uma ou mais etiquetas.
- Em (Pantel & Ravichandran, 2004) é apresentado um método relativamente simples destinado a obter ***classes etiquetadas*** de palavras

Estratégia

- Dado um agrupamento de elementos semelhantes, o sistema tenta encontrar:
  - hiperónimo
  - nome da classe
- O método tem 2 etapas
  - CBC parea obter os agrupamentos a etiquetar
  - Utilização de padrões superficiais para etiquetar os agrupamentos

Padrões

- Aposição.
  - "... **Oracle**, a **company** known for its progressive emplyoment policies,..."
- Sujeito Nominal.
  - "... **Apple** was a hot young **company**, with Steve Jobs in charge..."
- Estruturas Exemplificativas.
  - "... **companies** such as IBM must be weary...";
  - "... **companies** like **Sun Microsystems** do not shy away from such challenges,..."

Cenário Experimental

- Corpus de 3GB de texto jornalístico,
- foram gerados 1432 classes constituídas por substantivos (nomes comuns ou nomes próprios
- Para cerca de 1.5 % não foi possível obter uma etiqueta, tendo executada avaliação manual sobre as restantes 1411 classes

Avaliação

- consistiu em escolher aleatoriamente 125 classes e as 5 etiquetas mais prováveis de acordo com o algoritmo de etiquetagem
- foi adicionada a cada uma das 125 categorias uma etiqueta fornecida por um anotador humano.
- recorreu-se também à informação presente no WordNet:
  - para cada uma dos agrupamentos obtidos com pelo menos 5 instâncias na hierarquia do WordNet, obteve-se o nome da primeira classe hierarquicamente superior a essas instâncias
  - Devido a baixa cobertura do o WordNet no que diz respeito a nomes próprios, só foi possível etiquetar 33 das 125 agrupamentos em avalição.

Juizes…

- Foi gerada uma tabela de etiquetas para cada grupo: (automática, manual, WordNet)
- Esta tabela foi depois submetida a avaliação por juizes humanos que desconheciam a origem de cada uma das etiquetas
- Foi pedido aos juízes para avaliarem cada etiqueta atribuída às classes como correcta, parcialmente correcta ou incorrecta
- Foi computado o Mean Reciprocal Rank (MRR) em que cada classe era pontuada com um valor de 1 / M sendo M a posição da primeira etiqueta que os juízes consideraram correcta

Resultados

- O sistema automático de etiquetagem:
  - MRR 72.2%
- Wordnet
  - MRR = 19.9%
- Considerando apenas as 33 classes das quais foi posível obter etiquetas do WordNet:
  - sistema automático de etiquetagem MRR= 75.3%
  - WordNet MRR = 82.7%
- Em 72.0% das classes, a primeira etiqueta atribuída pelo sistema foi considerada correcta

Conclusão

- os autores referem que a atribuição de uma etiqueta a uma classe é um problema complicado
  - especialmente porque não parece consensual que uma classe deva ter apenas uma etiqueta.
- O método proposto atribui várias etiquetas e um valor que tenta quantificar a ligação entre a classe e a etiqueta,
- autores argumentam que esta deverá ser a aproximação mais correcta afastando-se da vista mais tradicional de que as representações semânticas devem ser "rígidas"

3. Breves conclusões

Aquisição de grupos de palavras

- Principio "linguístico" base:
  - Hipótese Distribucional de Significado
- Tecnologia base:
  - agrupamento / clustering
- Recursos base:
  - Corpora anotado, texto crú e um parser (ou só texto crú embora não apresentado)
- Produção de "Etiquetas" para a Classe:
  - posterior ao agrupamento, usando padrões superficiais
- Avaliação:
  - Manual e normalmente recorre a um padrão semântico: WordNet

Bibliografia

- Church, K. & Hanks, P. Word association norms, mutual information, and lexicography Computational Linguistics, 1990, 16(1), 22–29
- Lin, D. & Pantel, P. Concept Discovery from Text COOLING 2002, 2002
- Pantel, P. & Ravichandran, D. Automatically Labeling Semantic Classes. HLT-NAACL, 2004, 321-328
- Widdows, D. & Dorow, B. A Graph Model for Unsupervised Lexical Acquisition Procedings of 19th International Conference on Computational Linguistics (COOLING 19), 2002, 1093-1099