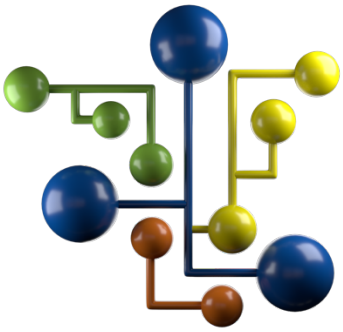

Data Science Academy

www.datascienceacademy.com.br

Introdução à Inteligência Artificial

Extração de Ontologias

Até agora pensamos em extração de informações, como encontrar um conjunto de relações específicas (por exemplo, locutor, hora, local) em um texto específico (por exemplo, um comunicado). Uma aplicação diferente da tecnologia de extração é a construção de uma base ampla de conhecimento ou ontologia de fatos de um corpus. Isso é diferente por três razões: primeiro, é em aberto — queremos adquirir fatos sobre todos os tipos de domínios, não apenas de um domínio específico. Em segundo lugar, com um corpus grande, essa tarefa é dominada pela precisão, não revocação, exatamente como perguntas e respostas na Web. Terceiro, os resultados podem ser agregados estatísticos recolhidos de várias fontes, em vez de ser extraídos de um texto específico. Por exemplo, Hearst (1992) examinou o problema de aprender uma ontologia de categorias e subcategorias de conceito de um corpus grande. (Em 1992, um corpus grande era uma enciclopédia com mil páginas; hoje seria um corpus de 100 milhões de páginas da Web.) O trabalho concentrou-se em modelos que são muito gerais (não vinculados a um domínio específico) e tinham alta precisão (eram quase sempre corretos quando correspondiam), mas com baixa cobertura (nem sempre coincidiam). Aqui está um dos modelos mais produtivos:

$$NP \textbf{ tal que } NP(, NP)^* \ (,)? \ ((\textbf{e} \mid \textbf{ou}) \ NP)?$$

Aqui as palavras em negrito e vírgulas devem aparecer literalmente no texto, mas os parênteses são do agrupamento, o asterisco significa repetição de zero ou mais e o ponto de interrogação significa opcional. NP é uma variável que suporta um sintagma nominal. Vamos assumir que sabemos que algumas palavras são substantivos e outras palavras (tais como verbos) podemos assumir confiantemente que não são parte de um sintagma nominal simples. Esse modelo corresponde ao texto “doenças como a raiva afetam o seu cão” e “suporta protocolos de rede tais como DNS”, concluindo que a raiva é uma doença e DNS é um protocolo de rede. Pode-se construir modelos semelhantes com as palavras-chave “inclusive”, “especialmente” e “ou outros”. Certamente esses modelos vão deixar de coincidir com muitas passagens relevantes, como “A raiva é uma doença”. Isso é intencional. O modelo “NP é um NP” de fato, por vezes, indica uma relação de subcategoria, mas muitas vezes significa outra coisa, como em “Há um Deus” ou “Ela está um pouco cansada”. Com um corpus grande podemos nos dar ao luxo de ser exigentes; usar apenas os modelos de alta precisão. Vamos perder muitas declarações de um relacionamento de subcategoria, mas muito provavelmente encontraremos uma paráfrase da declaração em outro lugar no corpus de uma forma que possamos usar.

A relação de subcategoria é tão fundamental que vale a pena fazer manualmente alguns padrões para ajudar a identificar os casos de ocorrências no texto em linguagem natural. Mas o que dizer dos milhares de outras relações no mundo? Criar e depurar padrões para todos eles, tomando como base alunos de pós-graduação de IA pelo mundo, não é suficiente. Felizmente, é possível aprender padrões a partir de alguns exemplos e então usá-los para aprender mais exemplos, de onde se pode aprender mais padrões, e assim por diante. Em uma das primeiras experiências desse tipo, Brin (1999) começou com um conjunto de dados de apenas cinco exemplos:

("Isaac Asimov", "The Robots of Dawn")
("David Brin", "Startide Rising")
("James Gleick", "Chaos-Making a New Science")
("Charles Dickens", "Great Expectations")
("William Shakespeare", "The Comedy of Errors")

Claramente, esses são exemplos da relação autor-título, mas o sistema de aprendizagem não tem conhecimento de autores ou títulos. As palavras nesses exemplos foram utilizadas em uma pesquisa em um corpus na Web, resultando em 199 combinações. Cada combinação é definida como tuplas de sete sequências,

(Autor, Título, Ordem, Prefixo, Meio, Sufixo, URL)

onde **Ordem** será verdadeiro se o autor vier primeiro e falso se o título vier primeiro, **Meio** são os caracteres entre o autor e título, **Prefixo** são os 10 caracteres antes da correspondência, **Sufixo** são os 10 caracteres após a correspondência e URL é o endereço da Web onde a correspondência foi feita. Dado um conjunto de correspondências, um esquema simples de geração de padrões pode encontrar padrões para explicar as correspondências. A linguagem dos padrões foi projetada para ter um mapeamento junto às correspondências, para ser passível de aprendizagem automatizada e para enfatizar a alta precisão (possivelmente com o risco de menor cobertura). Cada padrão tem os mesmos sete componentes como correspondência. O **Autor** e o **Título** são expressões regulares consistindo em quaisquer caracteres (mas começando e terminando em letras) e restritos a ter um comprimento de metade do comprimento mínimo dos exemplos até duas vezes o comprimento máximo. Prefixo, meio e sufixo são restritos a sequências literais, que não sejam expressões regulares. O meio é o mais fácil de aprender: cada sequência de meio distinta em um conjunto de correspondências é um modelo de candidato distinto. Para tal candidato, o padrão de prefixo é então definido como o sufixo comum mais longo de todos os prefixos nas correspondências, e o sufixo é definido como o prefixo comum mais longo de todos os sufixos nas correspondências. Se qualquer um deles for de comprimento zero, o padrão será rejeitado. A URL do padrão é definida como o prefixo mais longo das URLs nas correspondências. No experimento executado por Brin, as primeiros 199 combinações geraram três padrões. O padrão mais produtivo foi

<LI> <B> *Título* </ B> por *Autor* (
URL: www.sff.net/locus/c

Os três padrões foram então usados para recuperar mais 4.047 exemplos (autor, título). Os exemplos foram usados para gerar mais padrões, e assim por diante, acabando por produzir mais de 15.000 títulos. Dado um bom conjunto de padrões, o sistema pode coletar um bom conjunto de exemplos. Dado um bom conjunto de exemplos, o sistema pode construir um bom conjunto de padrões. A maior fraqueza dessa abordagem é a sensibilidade ao ruído. Se um dos

primeiros poucos padrões estiver incorreto, os erros podem se propagar rapidamente. Uma forma de limitar esse problema é não aceitar um novo exemplo a menos que seja verificado por vários padrões, e não aceitar um novo padrão, a menos que descubra vários exemplos que são também encontrados por outros padrões.

Referências:

Livro: Inteligência Artificial
Autor: Peter Norvig