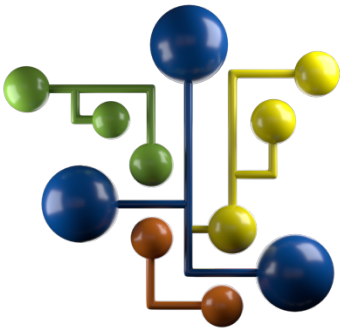

Data Science Academy

www.datascienceacademy.com.br

Introdução à Inteligência Artificial

Leitura de Máquina

A construção de padrões automatizados é um grande passo para a construção do padrão artesanal, mas ainda requer um punhado de exemplos rotulados de cada relação para começar.

Para construir uma grande ontologia com muitos milhares de relações, mesmo essa quantidade de trabalho seria onerosa; gostaríamos de ter um sistema de extração sem entrada humana de espécie nenhuma — um sistema que pudesse ler por conta própria e construísse sua própria base de dados. Tal sistema seria independente da relação; trabalharia para qualquer relação. Na prática, esses sistemas trabalham com todas as relações em paralelo, devido às exigências de E/S de corpora de grande dimensão. Eles se comportam menos como um sistema de extração de informações tradicional que se destina a poucas relações e mais como um leitor humano que aprende do próprio texto; por esse motivo, o campo tem sido chamado de leitura de máquina.

Um sistema de leitura de máquina representativo é o TEXTRUNNER (Banko e Etzioni, 2008). O TEXTRUNNER utiliza cotreinamento para aumentar seu desempenho, mas precisa de algo para conseguir desenvolver-se por si mesmo. No caso de Hearst (1992), padrões específicos forneceram o reinicialização e, para Brin (1998), foi um conjunto de cinco pares de autor-título. Para o TEXTRUNNER, a inspiração original foi uma taxonomia de oito modelos sintáticos muito gerais, como mostrado na figura abaixo. Sentiu-se que um pequeno número de padrões como esse poderia abranger a maior parte das formas como as relações são expressas em inglês. O processo é acelerado a partir de um conjunto de exemplos rotulados e extraídos do Penn Treebank, um corpus de sentenças analisadas. Por exemplo, da análise da sentença “Einstein recebeu o Prêmio Nobel em 1921”, o TEXTRUNNER é capaz de extrair a relação (“Einstein”, “recebeu”, “Prêmio Nobel”).

Dado um conjunto de exemplos rotulados desse tipo, o TEXTRUNNER treina uma cadeia linear CAC para extrair mais exemplos de textos sem rótulo. As características no CAC incluem palavras funcionais como “para” e “de” e “o”, mas não substantivos e verbos (e não frases nominais ou verbais). Como o TEXTRUNNER é independente de domínio, ele não pode confiar em listas predefinidas de substantivos e verbos.

Tipo	Modelo	Exemplo	Frequência
Verbo	NP_1 *Verbo* NP_2	X estabeleceu Y	38%
Substantivo-Prep	NP_1 *NP Prep* NP_2	X acordou com Y	23%
Verbo-Prep	NP_1 *Verbo Prep* NP_2	X se moveu para Y	16%
Infinitivo	NP_1 **para** *Verbo* NP_2	X planeja adquirir Y	9%
Modificador	NP_1 *Verbo* NP_2 *Subst*	X é vencedor de Y	5%
Subst-composto	NP_1 (, | **e** | - |:) NP_2 NP	X-Y tratar	2%
Verbo-composto	NP_1 (, | **e**) NP_2 *verbo*	X, Y fundir	1%
Aposto	NP_1 *NP* (: |,)? NP_2	X cidade natal : Y	1%

O TEXTRUNNER atinge precisão de 88% e cobertura de 45% F(1 de 60%) em um grande corpus da Web. O TEXTRUNNER extraiu centenas de milhões de fatos a partir de um corpus de meio bilhão de páginas da Web. Por exemplo, mesmo não tendo conhecimentos médicos predefinidos, extraiu cerca de 2.000 respostas para a consulta [o que mata as bactérias]; respostas corretas incluem antibióticos, ozônio, cloro, Cipro e brotos de brócolis. As respostas duvidosas incluem “água”, que veio da sentença “água fervendo por pelo menos 10 minutos vai matar a bactéria”. Seria melhor atribuir isso a “água fervendo” em vez de apenas a “água”. Com as técnicas delineadas neste capítulo e novas invenções contínuas, estamos começando a chegar mais perto do objetivo da leitura automática de máquina.

Referências:

Livro: Inteligência Artificial
Autor: Peter Norvig