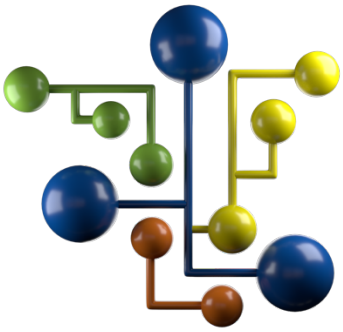


Data Science Academy

www.datascienceacademy.com.br

Introdução à Inteligência Artificial

Modelos Probabilísticos Para Extração de Informação

Quando se deve experimentar a extração de informações a partir de entrada ruidosa ou variada, a simples abordagem do estado finito é fraca. É muito difícil conseguir todas as regras e suas prioridades corretamente; é melhor utilizar um modelo probabilístico em vez de um modelo baseado em regras. O modelo probabilístico mais simples para sequências com o estado oculto é o modelo oculto de Markov, ou MOM. Um MOM modela uma progressão através de uma sequência de estados ocultos, xt, com uma observação et em cada etapa. Para aplicar MOMs para extração de informações, podemos construir um grande MOM para todos os atributos ou construir um MOM separado para cada atributo.

Faremos o segundo. As observações são as palavras do texto, e os estados ocultos indicam se estamos na parte do alvo, prefixo ou sufixo do modelo de atributo, ou em segundo plano (não faz parte do modelo). Por exemplo, aqui está um texto breve e o caminho mais provável para o texto de dois MOMs, um treinado para reconhecer o locutor de um comunicado e outro treinado para reconhecer datas. O "-" indica um estado em segundo plano:

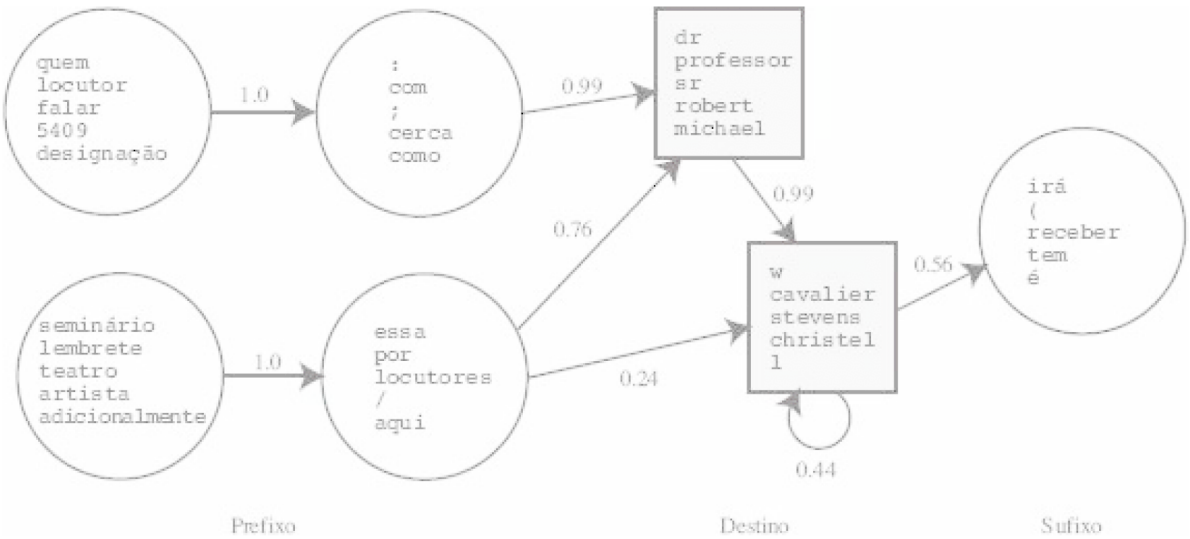
Texto:	Haverá	um	seminário	pelo	Dr.	Andrew	McCallum	na	sexta-feira
Locutor:	-	-	PRÉ	PRÉ	DESTINO	DESTINO	DESTINO	PÓS	-
Data:	-	-	-	-	-	-	-	PRÉ	DESTINO

Os MOMs são probabilísticos e, portanto, tolerantes ao ruído. Em uma expressão regular, se um único caractere esperado estiver ausente, o regex falha; com os MOMs há uma degradação elegante com relação a caracteres/palavras ausentes e temos uma probabilidade indicando o grau de correspondência, não apenas uma falha/correspondência booleana. Em segundo lugar, os MOMs podem ser treinados a partir dos dados, pois não necessitam de modelos de engenharia trabalhosos e, assim, podem ser mantidos atualizados mais facilmente à medida que o texto se altera ao longo do tempo.

Observe que assumimos certo nível de estrutura em nossos modelos MOM: todos eles consistem em um ou mais estados de destino e quaisquer estados de prefixo devem preceder os destinos, os estados de sufixo devem seguir os de destino, e outros estados devem ficar em segundo plano. Essa estrutura facilita a aprendizagem de MOMs a partir de exemplos. Com uma estrutura parcialmente especificada, o algoritmo para a frente e para trás pode ser utilizado para aprender tanto as probabilidades de transição P(Xt | Xt-1) entre estados como o modelo de observação, P(Et | Xt), que informa qual a probabilidade de cada palavra estar em cada estado. Por exemplo, a palavra "sextafeira" teria alta probabilidade de estar em um ou mais dos estados de destino do MOM e menor probabilidade de estar em outros lugares.

Com dados de treinamento suficientes, o MOM aprende automaticamente uma estrutura de datas que achamos intuitiva: a data do MOM pode ter um estado de destino em que a probabilidade alta de palavras são "segunda-feira", "terça-feira" etc., e que tem alta probabilidade de transição para um estado de destino com as palavras "Jan", "Janeiro", "Fev" etc. A figura abaixo mostra o MOM para o locutor de um comunicado, como aprendido dos dados. O prefixo abrange expressões como "locutor" e "seminário por", e o destino tem um

estado que abrange títulos e primeiros nomes e outro estado que abrange iniciais e sobrenomes.



Uma vez que os MOMs foram informados, podemos aplicá-los a um texto usando o algoritmo Viterbi para encontrar o caminho mais provável pelos estados do MOM. Uma abordagem é a aplicação de cada atributo MOM separadamente; nesse caso seria de esperar que os MOMs gastassem a maior parte do tempo nos estados em segundo plano. Isso é apropriado quando a extração é esparsa — quando o número de palavras extraído é pequeno em comparação com o comprimento do texto.

Outra abordagem é combinar todos os atributos individuais em um MOM grande que, então, encontraria um caminho que vagueia através de atributos de destino diferentes, primeiro encontrando um locutor destino, depois uma data destino etc. MOMs separados são melhores quando esperamos apenas um de cada atributo em um texto, e um MOM grande é melhor quando os textos têm o formato mais livre e denso com atributos. Com qualquer uma das abordagens, ao final temos uma coleção de observações de atributos de destino e precisamos decidir o que fazer com eles. Se todos os atributos esperados têm um preenchedor de destino, a decisão é fácil: temos um exemplo da relação desejada. Se houver múltiplos preenchedores, precisamos decidir qual escolher, como discutimos com o modelo baseado em sistemas. Os MOMs têm a vantagem de fornecer números de probabilidade que podem ajudar a fazer a escolha. Se alguns destinos estiverem faltando, precisamos decidir se essa é uma instância da relação desejada a todos ou se os destinos encontrados são falsos positivos. Um algoritmo de Machine Learning pode ser treinado para fazer essa escolha.

Referências:

Livro: Inteligência Artificial
Autor: Peter Norvig