A Era de Sistemas Cognitivos: Um Olhar Interno sobre o IBM Watson e Como Ele Funciona

IBM WATSON

Redguides

para Lideres de Negócios

Rob High

- Aprenda como sistemas cognitivos, como o IBM Watson, podem transformar a forma de pensar, agir e operar das organizações
- Entenda as capacidades de processamento de idioma nativo e muito mais do IBM Watson
- Veja como respostas baseadas em evidências podem orientar melhores resultados

Redbooks

Visão geral executiva

IBM® Watson™ representa uma primeira etapa em sistemas cognitivos, uma nova era de computação. O Watson constrói a era atual de computação programática, mas é diferente de formas significativas. A combinação dos recursos a seguir torna o Watson exclusivo:

- *Processamento de idioma nativo* ajudando a entender as complexidades de dados não estruturados, que compõem mais de 80% dos dados no mundo hoje
- *Geração e avaliação de hipótese* aplicando análises avançadas para ponderar e avaliar um painel de respostas com base apenas em evidência relevante
- *Aprendizado dinâmico* ajudando a melhorar o aprendizado com base em resultados para ficar mais inteligente com cada iteração e interação

Apesar de nenhum destes recursos sozinho ser exclusivo para o Watson, a combinação entrega uma solução poderosa:

- Mover além das restrições de computação pragmática
- Mover de confiança em dados estruturados e locais para explorar o mundo de dados globais e não estruturados
- Mover de aplicativos deterministas orientados a árvore de decisão para sistemas probabilísticos que se desenvolvem junto com seus usuários
- Mover de busca baseada em palavra-chave que fornece uma lista de locais onde uma resposta possa (ou não) estar localizada, para um meio intuitivo e conversacional de descobrir um conjunto de respostas classificadas em confiança

Sistemas cognitivos, como o IBM Watson, podem transformar a forma de pensar, agir e operar das organizações no futuro. Esta publicação do IBM Redguide™ descreve como o Watson combina processamento de idioma nativo, aprendizado dinâmico e geração e avaliação de hipótese para dar respostas diretas e baseadas em confiança.

O que é linguagem e porque ela é difícil para os computadores entenderem

Linguagem é a expressão de ideias. É o meio pelo qual nos comunicamos e entendemos coisas entre pessoas. É como transmitimos medo, esperança, história e direções para o futuro. Alguns dizem que ela é o que nós usamos para pensar, especular e imaginar. Ela está na base de nossa cognição, de nossa capacidade de entender o mundo à nossa volta ou pelo menos na base de nossa capacidade de manipular e trocar este entendimento.

E ela é incrivelmente imprecisa.

Nossa linguagem é cheia de insinuações, idiossincrasias, expressões idiomáticas e ambiguidade. Temos narizes que correm e pés que cheiram. Como uma *boa chance* e uma *grande chance* serem a mesma coisa, mas *homem esperto* e *cara esperto* serem opostos? Como uma casa pode *queimar* se for *incendiada*? Por que *preenchemos* um formulário *completando-o*?

E ainda, ela pode ser incrivelmente precisa.

Nós transmitimos muito significado, e realizamos muita colaboração mesmo no meio de todas as dificuldades com idioma. De alguma forma, podemos ver as lacunas, as inconsistências e contradições, a irregularidade e a falta de clareza e ainda nos entender uns aos outros com uma grande precisão.

Esta diferença entre exatidão e precisão é importante. *Exatidão* é a exatidão mecânica ou científica pode podemos localizar em uma passagem de um texto. Podemos determinar se uma palavra específica existe em uma passagem com um alto grau de exatidão. *Precisão* é o grau em que uma passagem infere que outra pode ser considerada verdadeira por pessoas razoáveis.

E se, quando dizemos "2 + 2", queremos dizer a configuração de um carro, como em dois bancos dianteiros e dois bancos traseiros?

A resposta para "2 + 2" é exatamente 4. A matemática nos ensina este fato. Ela também nos ensina que, independente de quantos zeros você coloca depois do número decimal para representar uma exatidão maior, a resposta sempre deriva em 4. Mas e se, quando dizemos "2 + 2", não queríamos que isto fosse interpretado literalmente como uma fórmula matemática, mas como uma expressão idiomática para a configuração de um carro, como em *dois bancos dianteiros e dois bancos traseiros*? Ou e se um psicólogo estiver usando "2 + 2" para se referir a uma família com *dois pais e duas crianças*? Nestes outros contextos, a resposta *quatro* pode não ser uma interpretação precisa do que estamos tentando transmitir na linguagem.

Na verdade, para responder com precisão uma pergunta, você deve com frequência considerar o contexto disponível para ela. Sem informações evidenciais suficientes, é difícil responder com precisão uma pergunta, mesmo se for possível responder com exatidão elementos na pergunta literalmente.

Processamento de idioma nativo superficial

Muitos sistemas de idiomas nativos têm tentado enfatizar a exatidão nos confins de regras bem formuladas específicas. Por exemplo, a análise de sentimento com frequência procura um conjunto específico de palavras e seus sinônimos em um site de mídia social. Estes sistemas, então, sem mais avaliação do contexto em que estas palavras estão sendo usadas, calcula o número de vezes que estas palavras são localizadas juntamente com alguma marca na mesma frase. Por exemplo, ele pega a frase , "… parada pelo Loja de

Rosquinhas IBM para um café esta manhã, foi bom …" e, então afirma que a colocação do nome da marca e o termo "bom" são uma indicação de um sentimento positivo. Entretanto, considere se o restante da frase é , "…, foi bom ouvir que um novo Café Fictício será aberto em breve, então não serei tentado a comer rosquinhas toda manhã." Em seguida, o sistema pode perder o fato de que o sentimento não é sobre a Loja de Rosquinhas IBM. Chamamos este conceito de *processamento de idioma nativo superficial (NLP)* porque, apesar de ele poder ser bastante exato em seus focos mais restritos, ele não é muito preciso.

Entretanto, também é importante perceber que o NLP superficial realmente tem uma função importante em muitos sistemas. Se sua intenção é criar uma avaliação relevante estatisticamente de tendências de sentimentos sobre grandes quantidades de informações, a falta de precisão para cada exemplo individual provavelmente não é um problema. Assumindo que haja aproximadamente tantos falsos positivos quanto falsos negativos em um conjunto de amostra suficientemente grande, elas se cancelam. E se o conjunto de cálculos cancelados permanecer relativamente constante nos conjuntos de amostras no decorrer do tempo, os dados não cancelados remanescentes produzirão informações de tendência estatisticamente relevantes. Portanto, os custos de processamento adicional que são necessários para a precisão adicional para qualquer ocorrência pode não ser autorizados.

Processamento de idioma nativo superficial pode ser bastante exato em seu foco mais restrito, mas não é muito preciso.

Entretanto, quando as ocorrências individuais são importantes, os sistemas que são projetados para ser exatos sem se concentrar em altos níveis de precisão tendem a ser quebradiços. Isto é, eles executam bem nos parâmetros limitados de seu design intencionado, mas não executam bem quando estes paramentos mudam. Comparamos estes sistemas a usar técnicas de construção de alvenaria. Tijolos são fortes e razoavelmente fáceis de serem usados em construção. Por décadas e séculos, refinamos a técnica de construção de alvenaria para ser bastante exata. Somos capazes de construir estruturas relativamente grandes, ornamentado e duráveis. Entretanto, apesar de construções de tijolo terem grande força de carga, elas são pobres em força de tensão. Eles caem facilmente em terremotos e não duram períodos grandes. E depois de um certo ponto, sua força de carga falhará também.

É possível observar estas mesmas limitações em alguns produtos do consumidor hoje. Por exemplo, você pode usar seu assistente pessoal ativado por voz e dizer, "Localize pizza para mim." Em troca, você obtém uma lista local de pizzarias, que exatamente o que deseja. Agora você diz, "*Não* localize pizza para mim." Você ainda obtém uma lista local de pizzarias, que não é exatamente o que você pediu. Da mesma forma, se disser "Localize pizza para mim *próximo*" ou "Localize pizza para mim *longe*", as mesmas listas locais são retornadas. O ponto é que estes sistemas são projetados de acordo com um conjunto específico de regras e estão procurando por combinações específicas de palavras-chave para determinar a resposta a produzir. Estes sistemas não sabem como distinguir entre coisas para as quais não há regra. Elas podem ser exatas, mas não necessariamente muito precisas.

Processamento profundo de idioma nativo

Para superar as limitações da construção de tijolos, mudamos para usar aço e concreto reforçado para construções maiores. Da mesma forma, estamos vendo um mudança em técnicas de construção para processamento de idioma nativo quando precisão é necessária ao invés de exatidão limitada. Estas técnicas incorporam muito mais contexto na avaliação da pergunta. Nos referimos a este conceito como *processamento profundo de idioma nativo*, o que às vezes é chamado *Pergunta-Resposta Profunda (DeepQA)* quando o problema é sobre responder grandes perguntas naturais.

Estamos vendo um mudança em técnicas de construção para processamento de idioma nativo quando precisão é necessária.

IBM Watson é um sistema de NLP profundo. Ele alcança a precisão tentando analisar a maior quantia de contextos possíveis. Ele obtém este contexto tanto da passagem da pergunta quanto da base de conhecimento (chamada de *corpus*) que está disponível para ele para localizar respostas.

Ao ser preparado para o programa de perguntas, JEOPARDY!, foram feitas as perguntas (pistas) a seguir para o Watson a partir da categoria Lincoln Blogs:

"Sec. do Tesouro. Chase acabou de enviar isto pela terceira vez - adivinhe parceiro, desta vez eu estou aceitando."

Primeiro, observe a abreviação, "Sec.", que tinha que ser entendida como *Secretária*. Além disso, observe que *Secretário* não é mencionado aqui como sendo alguém que toma ditados e gerencia uma agenda de compromissos. Os termos combinados *Secretária do Tesouro* são importantes aqui como um nome e uma função. Portanto, para responder esta pergunta, o Watson teve que localizar uma passagem que envolvesse enviar e aceitar algo entre as Secretarias do Tesouro Chase e Lincoln (a categoria da pista). Entretanto, observe também que a categoria não diz "Presidente Lincoln" necessariamente. A resposta correta então seria "O que é uma renúncia?".

Ao descrever este exemplo em uma escola de ensino básico às vezes após uma transmissão do IBM Watson jogando JEOPARDY!, um estudante do quinto ano ofereceu "O que é uma solicitação de amigo?" como uma resposta possível.

Sem contexto, estaríamos perdidos.

A resposta deste estudante é interessante em parte porque ela diz muito sobre o grau em que a mídia social penetrou profundamente na malha da próxima geração de sociedade. Entretanto, também é instrutiva porque também pode ser tomada como uma resposta relativamente razoável para a pista. Mas sabemos que esta resposta é imprecisa porque temos contexto histórico. Sabemos que o Facebook não estava disponível no final do século dezenove. Observe que o contexto é o que nos permitiu aumentar a precisão do sistema em produzir esta resposta. Sem o contexto, estaríamos perdidos.

É válido enfatizar o ponto de que nós, como humanos, temos pouca dificuldade de processar nossa linguagem, mesmo se ficamos confusos em algumas ocasiões. Mas geralmente fazemos muito melhor resolvendo o significado de informações que gravamos do que os computadores.

Temos uma qualidade inata sobre como desambiguizar a linguagem que desejamos capturar e aproveitar em sistemas de cálculo. Este conceito tem sido um objetivo chave da comunidade de inteligência artificial pelas quatro últimas décadas. E em grande medida, temos sido capazes de aumentar a exatidão do processamento de linguagem. Mas é apenas com o Watson que podemos finalmente atravessar o nível de precisão que é necessário para sistemas de informações funcionarem bem no mundo real de amplo idioma natural.

Além disso, uma grande força de orientação busca resolver este problema. Estamos vivenciando uma explosão de produção de dados. Noventa porcento de todos os dados no mundo foram produzidos nos últimos dois anos. A expectativa é que esta tendência cresça conforme nos interconectamos e instrumentamos mais de nosso mundo. E 80% de todas as informações no mundo são informações desestruturadas, que incluem textos como literatura, relatórios, artigos, relatórios de pesquisa, teses, emails, blogs, tweets, fóruns, bate-papos e mensagens de texto. Precisamos que os computadores possam entender este grande fluxo de informações de forma que possamos utilizá-lo da melhor maneira.

O IBM Watson entende a linguagem

Navegação efetiva através de uma enchente atual de informações não estruturadas requer um uma nova era de computação que chamamos de *sistemas cognitivos*. O IBM Watson é um exemplo de sistema cognitivo. É possível separar parte da linguagem humana para identificar inferências entre passagens de texto com precisão humana, e em velocidades e escalas que são muito mais rápidas e maiores que qualquer pessoa possa executar sozinha. Ele pode gerenciar um alto nível de precisão quando vem a entender a resposta correta para uma pergunta.

Entretanto, o Watson não entende de verdade palavras individuais da linguagem. Ao invés disso, ele entende os recursos de linguagem que são usados pelas pessoas. A partir destes recursos, é possível determinar se uma passagem de texto (que chamamos de *pergunta*) infere outra (que chamamos de *resposta*), com um alto nível de precisão sob circunstâncias alternantes.

No programa de perguntas JEOPARDY!, o Watson tinha que determinar se a pergunta, “Jodie Foster levou isto para casa por seu papel em ‘Silêncio dos Inocentes’” inferia a resposta “Jodie Foster ganhou um Oscar por seu papel em ‘Silêncio dos Inocentes’”. Neste caso, *levar algo para casa* inferiu *ganhar um Oscar*, mas nem sempre. Às vezes *levar algo para casa* infere um resfriado, guloseimas ou várias outras coisas quaisquer. Por outro lado, nem sempre você leva para casa as coisas que ganha. Por exemplo, você pode ganhar um contrato para trabalhar, mas isso não é algo que você leva para casa.

O contexto é importante. Restrições temporais e espaciais são importantes. Todos estes conceitos agregam em permitir que um sistema cognitivo se comporte com características humanas. E, para voltar a um ponto anterior, uma abordagem baseada em regras precisa de um número quase infinito de regras para capturar cada caso que podemos encontrar em linguagem.

Navegação efetiva através de uma enchente atual de informações não estruturadas requer um uma nova era de computação chamada sistemas cognitivos.

O Watson separa de lado a pergunta e respostas potenciais no corpus e, então, examina ele e o contexto da afirmação de centenas de maneiras. O Watson, então, usa os resultados para obter um grau de confiança em sua interpretação da pergunta e de potenciais respostas.

Mas devemos voltar um pouco. Como o Watson orienta suas respostas para perguntas? Ele utiliza o seguinte processo:

1. Quando uma pergunta é apresentada primeiro ao Watson, ele a analisa para extrair os principais recursos dela.
2. Ele gera um conjunto de hipóteses procurando através do corpus por passagens que tenham algum potencial de conter uma resposta de valor.
3. Ele executa uma comparação profunda do idioma da pergunta e do idioma de cada resposta em potencial usando vários algoritmos de raciocínio.

   Esta etapa é desafiadora. Há centenas de algoritmos de raciocínio, cada um dos quais executa uma comparação diferente. Por exemplo, alguns olham a correspondência de termos e sinônimos, alguns olham os recursos temporais e espaciais, e alguns olham origens relevantes de informações contextuais.
4. Cada algoritmo de raciocínio produz uma ou mais pontuações, indicando a extensão em que a resposta potencial é inferida pela pergunta com base na área específica de foco deste algoritmo.
5. Cada pontuação resultante é, então, ponderada com relação a um modelo estatístico que captura como se saiu este algoritmo em estabelecer as inferências entre duas passagens semelhantes para aquele domínio durante o “período de treinamento” para o Watson.

Este modelo estatístico pode, então, ser usado para resumir um nível de confiança que o Watson tem sobre a evidência que a resposta candidata é inferida pela pergunta.

6. O Watson repete este processo para cada uma das respostas candidatas até que ele possa localizar respostas que aparecem como candidatas mais fortes que as outras.

A Figura 1 ilustra como o Watson orienta uma resposta para uma pergunta.

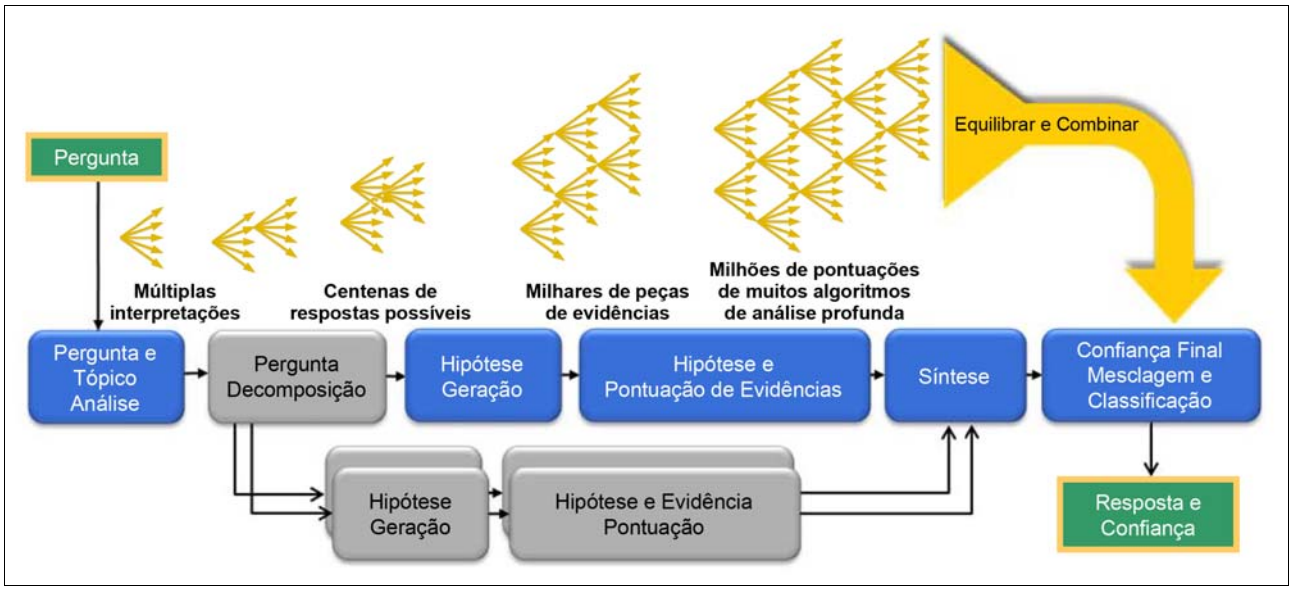

Figura 1 Como o Watson orienta uma resposta a uma pergunta

Um *corpus de conhecimento* é de importância primordial para a operação do Watson. Este corpus consiste em todos os tipos de conhecimento não estruturado, como livros de texto, diretrizes, manuais de com fazer, Perguntas Mais Frequentes, planos de benefícios e notícias. O Watson se alimenta o corpus, indo através de todo o corpo do conteúdo para colocá-lo num formato com o qual seja mais fácil de trabalhar. O processo de alimentação também organiza o conteúdo. Isto é, ele se concentra em se o corpus contém conteúdo apropriado, desprezando os artigos ou páginas que estão desatualizados, que são irrelevantes ou que vêm de fontes potencialmente não confiáveis.

Alguns algoritmos de raciocínio se concentram em recursos espaciais e temporais da passagem, que são críticos para desambiguizar uma enorme quantia do que humanos dizem e escrevem. Quando dizemos, “Localize pizza para mim,” é dado como certo que queremos dizer algo próximo. Mas o que está próximo é sempre relativo. Em outros casos, relacionamentos espaciais parecem relativos para mercados geográficos, por exemplo, uma vizinhança em uma cidade ou um estado em um país. Da mesma forma, recursos temporais também estão presentes no contexto de muito do que escrevemos. Quando dizemos, “Leve queijo da loja em seu caminho para casa,” um prazo é inferido. Provavelmente o escritor e o destinatário têm um entendimento contextual compartilhado de quando estarão em seu caminho para casa.

Avaliação espacial e temporal tem que ser executada tanto na pergunta quanto na resposta candidata.

A afirmação, “Em maio de 1898, Portugal celebrou o 400º aniversário da chegada deste explorador’ à Índia,” demonstra tanto dimensões espaciais quanto temporais. A celebração ocorreu em Portugal, mas o evento que eles estavam celebrando era a chegada do explorador’ à Índia. A afirmação sugere que o explorador foi de Portugal à Índia? Ele já esteve em Portugal? Observe que a celebração ocorreu em 1898, mas o evento ocorreu 400 antes. Portanto, o evento ocorreu realmente em 1498. A passagem que fornecia a

resposta para a pergunta dizia, “Em 27 de maio de 1498, Vasco da Gama desembarcou em Kappad Beach.” A avaliação espacial e temporal tinha que ser executada tanto na passagem da pergunta quanto na da resposta candidata.

O contexto é derivado tanto das informações imediatas quanto do conhecimento que está disponível mais amplamente. O Watson pode derivar informações imediatas do título de um documento, de outras passagens em um documento ou do banco de dados de origem de onde elas são originadas. O contexto também pode vir mais amplamente de um histórico compartilhado. Lembre-se de que sabíamos que “O que é uma Solicitação de Amigo?” era provavelmente uma resposta incorreta para a pista no Lincoln Blogs. O motivo é porque compartilhamos um contexto histórico comum, que nos diz quando certas coisas aconteceram em relação umas às outras. Sabemos que o Facebook foi criado muito recentemente, mas sabemos que Abraham Lincoln vivem há 150 anos atrás, bem antes de o Facebook se tornar popular. O contexto e o raciocínio nos ajudam a criar uma base cognitiva para processar linguagem.

Entender a linguagem é apenas o começo

Definimos sistemas cognitivos confirme aplicamos características humanas para transmitir e manipular ideias. Quando combinadas com as forças inerentes da computação digital, elas podem ajudar a resolver problemas com maior precisão, mais resiliência e em uma escala massiva sobre grandes corpos de informações.

Podemos decompor um sistema cognitivo como tendo vários elementos chave (Figura 2). A caixa sombreada média indica os recursos atuais de sistemas cognitivos. As caixas sombreadas mais claras indicam os recursos futuros dos sistemas cognitivos.

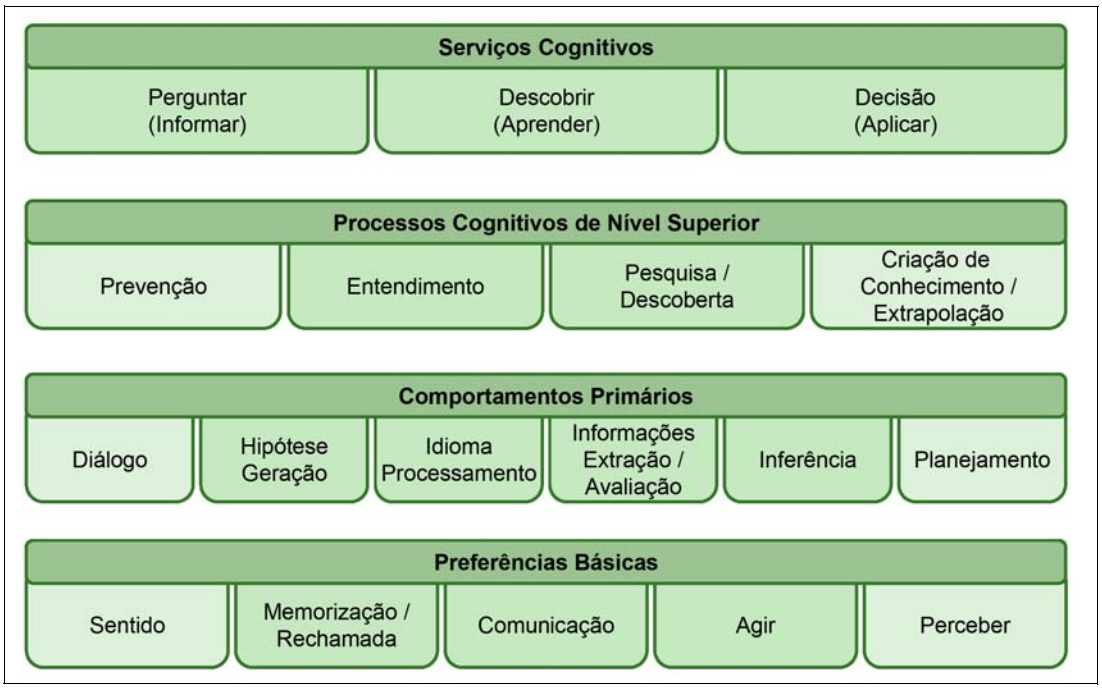

Figura 2 Elementos de um sistema cognitivo

Semelhante aos humanos, os sistemas cognitivos têm um meio de reunir, memorizar e rechamar informações, que é o equivalente às memórias humanas. Sistemas cognitivos

também têm uma capacidade básica de se comunicar e agir. Estas capacidades são organizadas por certas construções comportamentais, como os exemplos a seguir:

- A capacidade de criar e testar hipóteses
- Capacidade de separar de lado e criar inferências sobre linguagem
- A capacidade de extrair e avaliar informações úteis (como datas, locais e valores)

Estas qualificações são fundamentais, sem as quais os computadores não humanos podem determinar a correlação correta entre questões e respostas.

Processos cognitivos mais antigos e maiores podem alavancar comportamentos fundamentais para atingir um nível de entendimento. Entender algo requer que sejamos capazes de quebrá-lo em pedaços, até elementos mais finos que se comportem de formas bem-orientadas em uma dada escala. Como as coisas funcionam em física em escalas humanas não é como as coisas funcionam em escalas cósmicas ou subatômicas. Da mesma forma, sistemas cognitivos são projetados para funcionar como escalas humanas, embora sobre enormes coleções de humanos. Como tal, entender linguagem começa com entender as regras mais finas de linguagem, não apenas gramática formal, mas as convenções gramaticais informais do dia a dia.

Assim como os humanos, sistemas cognitivos são orientados a entender coisas decompondo expressões de uma ideia e, então, combinando isto com contexto.

Entretanto, como humanos, os sistemas cognitivos são levados a entender conceitos decompondo expressões de uma ideia e, então, combinando os resultados com o contexto e a probabilidade que certos termos na passagem estão sendo usados de certa forma. E, como com humanos, nossa confiança é proporcional à evidência que suporta estas probabilidades e o número de algoritmos de raciocínio que temos disponível para testar nossas hipóteses.

Depois de estabelecer um certo nível de entendimento, decompondo o problema com relação a sua provável intenção, sistemas cognitivos podem recompor os elementos de várias maneiras, cada uma das quais pode ser testada para imaginar novos conceitos. Estas combinações podem, então, ser usadas para orientar nova descoberta e insight, ajudando-nos a localizar respostas para perguntas e a perceber as perguntas que nunca pensamos em responder.

Podemos, então, usar estes recursos para resolver problemas que se ajustam a certos padrões comuns. Podemos fazer perguntas que produzem respostas. Podemos usar o sistema para descobrir novos insights e perceber conceitos que não reconhecemos anteriormente. E podemos usar estes sistemas para apoiar boas decisões ou pelo menos ajudar as pessoas nas decisões que precisam tomar.

Conforme os sistemas cognitivos ficam enriquecem, é esperado que eles obtenham a capacidade de entender.

No futuro, conforme os sistemas cognitivos ficarem mais ricos, esperamos que eles ganhem a capacidade de entender. Esperamos que eles façam mais que apenas ler texto, mas que vejam, escutem e sintam de forma que possam ter um reconhecimento básico de seu ambiente. E esperamos que estes sistemas sejam capazes de perceber informações, como reconhecer formas e alterar condições que darão mais informações a seu contexto e capacidade de inferir e raciocinar. Também esperamos que eles adotem comportamentos e processos cognitivos de maior ordem, como ter um diálogo, planejar diferentes estratégias para resolver problemas e ganhar conhecimento antecipado e extrapolá-lo em novo conhecimento.

Em essência, sistemas cognitivos internalizarão muitos dos comportamentos que humanos acham “naturais,” e os aplicarão em escala massiva para ajudar as pessoas a resolverem os problemas que hoje frequentemente não se encaixam em suas lacunas. Estamos iniciando uma nova era. Nesta era, os computadores vão além de apenas executar tarefas processuais de rotina mais eficientemente, para empregar cognição humana para tornar as pessoas mais inteligentes sobre o que elas fazem.

Os problemas vêm em diferentes formas

Conforme seguimos em frente com o IBM Watson, estamos descobrindo outros usos para ele. No clássico “Pergunte ao Watson,” um usuário faz ao Watson uma pergunta (ou fornece uma pista, um registro paciente, e assim por diante), da qual o Watson deriva uma resposta. O Watson possui a confiança de que a pergunta infere a resposta e a evidência que suporta a resposta. O Watson localizou uma casa nos campos de Diagnóstico de Oncologia, Gerenciamento de Utilização (isto é, pré-aprovação de cobertura de seguro para procedimentos médicos planejados), Análise de Crédito, e pesquisa básica.

O Watson ajuda se um profissional precisar de assistência em obter as informações mais relevantes para seu espaço de problema.

Ao fazer perguntas mais importantes, é possível começar a pensar sobre seus problemas de negócios de uma forma totalmente diferente.

Uma das maiores revelações sobre o Watson é que, usando ele para ajudar a responder perguntas, você pode perceber que está fundamentalmente fazendo as perguntas erradas. Quando o Watson responde suas perguntas, mesmo respondendo corretamente, você pode perceber que precisa fazer outras perguntas, melhores e mais importantes, para ajudar a considerar seu problemas de negócios de uma forma totalmente nova. Você começa a pensar de uma maneira que o ajuda a entender as ameaças de concorrência e as oportunidades em seu mercado de trabalho que nunca lhe ocorreu antes.

Estas alocações de tipo de descoberta estão sendo melhoradas com trabalho que a IBM está fazendo agora nos laboratórios IBM Research e Software Development. Avanços recentes em encadeamento de inferência (determinando se *isto* infere *naquilo*, que infere em outra coisa, e assim por diante) estão criando insight mais profundo. Saber que diabetes causa alto nível de açúcar no sangue é importante. Entretanto, tomar a próxima etapa para desenhar inferência entre alto nível de açúcar no sangue que causa cegueira é mais crítico para se preocupar para todo o paciente. Estes tipos de inferências de diversos níveis podem ser capturados como um gráfico de inferência no qual podemos observar um amplo espectro de considerações de recebimento de dados. Mais importante, a convergência no gráfico é uma forma poderosa de derivar mais inferências significativas, como respostas que podem revelar insights mais profundos e consequências ocultas. Ao unir valores de confiança precedentes, podemos agregar e estabelecer maior confiança em uma resposta como sendo a resposta preferencial a uma pergunta.

Podemos produzir inferências reversas, em efeito, descobrindo perguntas para respostas que nunca foram pedidas.

Além disso, podemos produzir inferências reversas, que na verdade significam que descobrimos as perguntas para as respostas que nunca perguntamos. Determinar que um paciente que tem um histórico de tremores em repouso e uma “face inexpressiva” pode inferir que ele tenha mal’ de Parkinson. Entretanto, determinar que o paciente também tem dificuldade de andar pode também revelar dano no sistema nervoso na Substância Negra, que pode ter sido perdido sem solicitar perguntas não feitas anteriormente.

A IBM está investindo nestes tipos de melhorias maiores no Watson que acreditamos que levarão a mais avanços em assistência médica, finanças, centrais de contato, governo, indústrias químicas e um planeta mais inteligente. Estes tipos de avanços podem ajudar a nos impulsionar para uma era de sistemas cognitivos.

Em muitas soluções, o Watson está sendo alavancado com outras formas mais tradicionais de cálculo, como análises estatísticas, processamento de regras e negócios, colaboração e relatório, para resolver problemas de negócios. `Por exemplo, considere a ideia de a IBM combinando outras análises estatísticas com a capacidade do Watson de responder perguntas sobre eventos potenciais que podem sinalizar um risco para um investimento. A IBM pode ajudar nossos clientes a melhorar seus processos de risco e avaliação para instituições financeiras. Da mesma forma, o insight que ganhamos sobre as respostas do cliente através de NLP profundo pode sugerir mudanças para o comportamento de compra e uso que de outra forma pode não ser evidente nos dados estruturados. No segmento de mercado de assistência médica, o Watson está sendo usado para ajudar empresas de seguro em seus processos para pré-aprovar tratamentos como parte de seus processos de Gerenciamento de Utilização.

A precisão é melhorada através da generalização

Conforme a IBM continua a evoluir e desenvolver estes tipos de sistemas cognitivos, devemos exercer cuidado. Estamos em uma conjuntura clássica que nós humanos enfrentamos o tempo todo, que é especializar ou generalizar. Podemos especializar a tecnologia NLP para um domínio específico, concentrando-nos, por exemplo, apenas nos recursos linguísticos deste domínio. Esta abordagem é tentadora e pode até mesmo ser necessária nas fases iniciais de evolução para assegurar a viabilidade da tecnologia. Entretanto, esta abordagem provavelmente nos levará de volta à era de construir com tijolos. Se a capacidade de adaptar-se com destreza humana torna os sistemas cognitivos especiais, devemos generalizar. Precisamos reconhecer e desenhar inferências de um conjunto mais amplo de variação linguística, sob um conjunto mais amplo de circunstâncias, como nossas mudanças de conhecimento, como as mudanças de contexto e como mudança linguística contemporânea.

Usando esta abordagem, podemos nos adaptar mais prontamente a problemas novos e maiores. Já estamos aplicando o Watson nos segmentos de serviços de assistência médica e financeiro, o que tem os benefícios a seguir:

- Ele traz as vantagens do Watson para diferentes domínios com problemas de alto valor.
- Ele desenvolve os algoritmos de processamento de idioma do Watson para manipular um conjunto mais amplo de variação linguística.

Esta abordagem permite a adaptação mais fácil aos domínios e melhora o utilitário do Watson para nossos aplicativos de domínio existentes.

Estamos em uma conjuntura clássica, seja para especializar ou generalizar.

Aplicações de assistência médica são interessantes porque frequentemente precisam de exatidão e precisão. A precisão é necessária para interpretar adequadamente o texto na descrição de saúde' do paciente para inferir a condição' do paciente. Entretanto, a diretriz National Comprehensive Cancer Network (NCCN) para câncer de mama deve ser justificada pela presença de termos exatos no registro de saúde' do paciente. Em seguida, mais precisão é necessária para localizar evidência que apoie este tratamento.

Estamos no início de uma nova era de computação, que é menos exata, mas muito mais precisa.

Sempre que encontrarmos uma anomalia linguística (algo na linguagem que nunca encontramos antes), tomamos uma decisão sobre se o problema é exclusivo para o domínio ou se é característica de um conjunto mais amplo de problemas linguísticos. Sempre que possível, voltamos aos nossos algoritmos básicos para determinar se podemos generalizar o algoritmo para reconhecer e avaliar a nova situação. Assim como com humanos, usando esta abordagem, podemos mapear nosso entendimento de novas experiências e, portanto, aumentar a base contextual para o sistema.

O resultado esperado é que aumentemos a precisão, o escopo e a escala:

- Precisão de inferência linguística (obtendo a resposta correta, para o motivo correto, em tempo hábil)
- Escopo do espaço do problema
- Ajustando a escala para quantias massivas de dados e perguntas, em muito mais domínios

Esperamos ver valores ainda maiores nas soluções de "próxima melhor ação", análise de sentimento social, refinamento de petróleo o produtos químicos e muito mais aplicações. Estamos apenas no começo de uma nova e maior era de computação que está menos concentrada em exatidão, mas muito mais precisa. É uma era de aplicar comportamento humano a problemas de computação de larga escala. É a era de sistemas cognitivos.

Outros recursos para obter mais informações

Para obter mais informações sobre o Watson, consulte o website do IBM Watson em:

ibm.com/innovation/us/watson/index.shtml

A equipe do autor que escreveu este guia

Este guia foi escrito por Rob High em colaboração com o International Technical Support Organization (ITSO).

Rob High é um IBM Fellow, Vice President e Chief Technology Officer, IBM Watson Solutions, IBM Software Group. Ele tem a responsabilidade geral de orientar a estratégia técnica do IBM Watson Solutions e de liderança de pensamento. Como um membro chave da equipe do IBM Watson Solutions Leadership, Rob trabalha de forma colaborativa com as equipes de engenharia, pesquisa e desenvolvimento do Watson. Rob tem 37 anos de experiência de programação e trabalhou com monitores de transações distribuídas, orientadas a objeto, e baseadas no componente pelos últimos 21 anos, incluindo SOMObject Server, Component Broker e, mais recentemente, o IBM WebSphere® Application Server. Rob anteriormente atuou como Chief Architect para a fundação WebSphere com responsabilidade de arquitetura para o WebSphere Application Server e produtos relacionados que são integrados neste tempo de execução principal.

Agora é possível se tornar um autor publicado também!

Aqui' está uma oportunidade para destacar suas qualificações, crescer em sua carreira e se tornar um autor publicado — tudo ao mesmo tempo! Junte-se ao projeto de residência da ITSO e ajude a escrever um livro em sua área de experiência, enquanto afia sua experiência usando tecnologias de ponta. Seus esforços ajudarão a aumentar a aceitação do produto e a satisfação do cliente, conforme você expande sua rede de contatos técnicos e relacionamentos. As residências têm de duas a seis semanas de duração, e é possível participar pessoalmente ou como um residente remoto trabalhando de sua base de casa.

Encontre mais sobre o programa de residência, navegue pelo índice de residência e se cadastre online em:

`ibm.com/redbooks/residencies.html`

Fique conectado com o IBM Redbooks

- Encontre-nos no Facebook:

  `http://www.facebook.com/IBMRedbooks`
- Siga-nos no Twitter:

  `http://twitter.com/ibmredbooks`
- Procure por nós no LinkedIn:

  `http://www.linkedin.com/groups?home=&gid=2130806`
- Explore novas publicações do IBM Redbooks®, residências e workshops com a newsletter semanal do IBM Redbooks:

  `https://www.redbooks.ibm.com/Redbooks.nsf/subscribe?OpenForm`
- Fique atualizado sobre publicações recentes do Redbooks com Feeds RSS:

  `http://www.redbooks.ibm.com/rss.html`

Avisos

Estas informações foram desenvolvidas para produtos e serviços oferecidos nos Estados Unidos.

É possível que a IBM não ofereça os produtos, serviços ou recursos discutidos nesta publicação em outros países. Consulte um representante IBM local para obter informações sobre produtos e serviços disponíveis atualmente em sua área. Qualquer referência a produtos, programas ou serviços IBM não significa que apenas produtos, programas ou serviços IBM possam ser utilizados. Qualquer produto, programa ou serviço funcionalmente equivalente, que não infrinja nenhum direito de propriedade intelectual da IBM, poderá ser utilizado em substituição a este produto, programa ou serviço. a avaliação e verificação da operação de qualquer produto, programa ou serviço não IBM são de responsabilidade' do Cliente.

A IBM pode ter patentes ou solicitações de patentes pendentes relativas a assuntos tratados nesta publicação. O fornecimento desta publicação não concede ao Cliente direito algum sobre tais patentes. Pedidos de licença devem ser enviados, por escrito, para:
Gerência de Relações Comerciais e Industriais da IBM Brasil Av. Pasteur, 138-146 Botafogo Rio de Janeiro, RJ CEP 22290-240

O parágrafo a seguir não se aplica a nenhum país em que tais disposições não estejam de acordo com a legislação local: A INTERNATIONAL BUSINESS MACHINES CORPORATION FORNECE ESTA PUBLICAÇÃO "NO ESTADO EM QUE SE ENCONTRA", SEM GARANTIA DE NENHUM TIPO, SEJA EXPRESSA OU IMPLÍCITA, INCLUINDO, MAS NÃO SE LIMITANDO, ÀS GARANTIAS IMPLÍCITAS DE NÃO INFRAÇÃO, COMERCIALIZAÇÃO OU ADEQUAÇÃO A UM DETERMINADO PROPÓSITO. Alguns países não permitem a exclusão de garantias expressas ou implícitas em certas transações; portanto, essa disposição pode não se aplicar ao Cliente.

Essas informações podem conter imprecisões técnicas ou erros tipográficos. São feitas alterações periódicas nas informações aqui contidas; tais alterações serão incorporadas em futuras edições desta publicação. A IBM pode, a qualquer momento, aperfeiçoar e/ou alterar os produtos e/ou programas descritos nesta publicação, sem aviso prévio.

Referências nestas informações a websites não IBM são fornecidas apenas por conveniência e não representam de forma alguma um endosso a esses websites. Os materiais contidos nesses websites não fazem parte dos materiais desse produto IBM e o uso desses websites é de inteira responsabilidade do Cliente.

A IBM pode usar ou distribuir as informações fornecidas da forma que julgar apropriada sem incorrer em qualquer obrigação para com o Cliente.

Todos os dados de desempenho aqui contidos foram determinados em um ambiente controlado. Portanto, os resultados obtidos em outros ambientes operacionais podem variar significativamente. Algumas medidas podem ter sido tomadas em sistemas em nível de desenvolvimento e não há garantia de que estas medidas serão iguais em sistemas geralmente disponíveis. Além disso, algumas medidas podem ter sido estimadas por extrapolação. Os resultados reais podem variar. Os usuários deste documento devem verificar os dados aplicáveis para seu ambiente específico.

As informações relativas a produtos não IBM foram obtidas junto aos fornecedores dos respectivos produtos, de seus anúncios publicados ou de outras fontes disponíveis publicamente. A IBM não testou estes produtos e não pode confirmar a precisão de seu desempenho, compatibilidade nem qualquer outra reivindicação relacionada a produtos não IBM. Dúvidas sobre os recursos de produtos não IBM devem ser encaminhadas diretamente a seus fornecedores.

Estas informações contêm exemplos de dados e relatórios utilizados nas operações diárias de negócios. Para ilustrá-los da forma mais completa possível, os exemplos podem incluir nomes de indivíduos, empresas, marcas e produtos. Todos estes nomes são fictícios e qualquer semelhança com nomes e endereços utilizados por uma empresa real é mera coincidência.

LICENÇA DE COPYRIGHT:

Estas informações contêm programas de aplicativos de amostra na linguagem fonte, ilustrando as técnicas de programação em diversas plataformas operacionais. O Cliente pode copiar, modificar e distribuir estes programas de amostra sem a necessidade de pagar à IBM, com objetivos de desenvolvimento, utilização, marketing ou distribuição de programas aplicativos em conformidade com a interface de programação de aplicativo para a plataforma operacional para a qual os programas de amostra são criados. Esses exemplos não foram testados completamente em todas as condições. Portanto, a IBM não pode garantir ou implicar a confiabilidade, manutenção ou função destes programas.

Este documento, REDP-4955-00, foi criado ou atualizado em April 7, 2013.

Marcas Registradas