Uma visão geral sobre o uso de sistemas de perguntas e respostas na ciência cognitiva

Sergio Varga
Universidade Estadual de Campinas, Faculdade de Engenharia Elétrica e Computação, Campinas, Brasil
sergiovarga@yahoo.com

***Resumo* — A partir da década de 50 uma nova ciência foi criada resultante da congruência de outras ciências e originou uma nova àrea de pesquisa, a ciência cognitiva. Uma das sub-àreas relacionadas a ela é a Inteligência Artificial e mais específicamente os Sistemas de Perguntas e Respostas. Esse trabalho visa apresentar uma visão geral desses sistemas e, mais especificamente, uma maior explicação do sistema DeepQA, desenvovldo para competir com seres humanos em um programa de perguntas e respostas, e como eles se relacionam com a ciência cognitiva.**

***Palavras Chaves*— ciência cognitiva, sistemas de perguntas e respostas, inteligência artificial**

I. INTRODUÇÃO

A mente humana, desde a antiguidade, sempre foi um objeto de estudo curioso e cativante. Seja por ser uma coisa extremamente complexa, se tratar de um fenômeno subjetivo, de difícil definição, e ter vários objetos de pesquisa relacionadas como percepção, linguagem, raciocínio, memória, atenção, inteligência, emoção e significado, ou seja pelas caracteristicas relacionados aos processos mentais como representação, consciência, intencionalidade e livre-arbítrio.

Essa fascinação veio a emergir uma nova ciência com o objetivo de estudar os processos cognitivos envolvidos na aquisição, representação e uso do conhecimento humano relacionados aos fenômenos da mente e da inteligência.

O surgimento da ciência cognitiva remonta do pós-guerra e pode-se dizer que foi um processo convergente englobando várias àreas relacionadas como apresentado na Figura 1 abaixo.

Alguns eminentes pesquisadores foram fundamentais

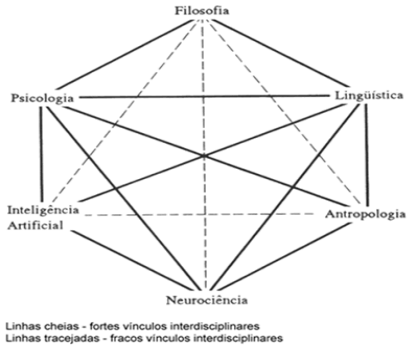

Fig. 1 - Disciplinas Relacionadas a Ciência Cognitiva

para o desenvolvimento dessa nova ciência [1]. Na matemática e computação Alan Turing, com sua máquina de Turing e John von Neumann com o programa de computador para instruir a máquina de Turing; Warren McCulloch e Walter Pitts com a identificação de que as operações das células nervosas podiam ser modeladas em termos de lógica; Norbert Wiener com a síntese cibernética e a teoria do controle e comunicação; Claude Shannon com a Teoria da Informação e os estudos relacionados a incapacidades cognitivas resultantes de danos cerebrais em especial na figura de Herbert Simon.

As àreas apontadas foram importantes contribuidores para a formação da ciência cognitiva e os seguintes tópicos foram incorporados a nova ciência como objeto de estudo:

- Da filosofia foram agregados os estudos da mente e a relação com o corpo, o conceito de intencionalidade e os fenômenos mentais.
- Da psicologia, tópicos relacionados a desenvolvimento da inteligência, representações mentais e o processo do pensamento.
- Da inteligência artificial, tópicos relacionados ao uso de computadores para resolver problemas específicos ou genéricos e o debate do papel das máquinas em substituição ao homem.
- Da linguística, tópicos relacionados ao desenvolvimento linguístíco, representação da linguagem e reconhecimento sintático e semântico.
- Da antropologia, tópicos relacionados à cultura, evolução da linguagem e sociedade.
- Da neurociência, tópicos relacionados a estudos sobre como o cérebro funciona, identificação de partes do cérebro e padrões.

O marco da ciência cognitiva foi o Simpósio Hixon realizado em Setembro de 1948 com os principais pesquisadores da época nas mais diversas àreas. Em conjunto com esse simpósio alguns outros encontros posteriores, chamados de encontros catalíticos, serviram para o desenvolvimento dessa nova ciência.

Em Setembro de 1956 foi realizado o Simpósio sobre Teoria da Informação realizado no MIT onde importantes apresentações como “Máquina de Teoria Lógica” de Allen Newell e Herbert Simon; “Três Modelos de Linguagem” de Noam Chomsky e GeorgeMiller com seu artigo sobre a capacidade da memória humana de curto prazo, foram marcantes dentro da nova ciência que estava por vir.

Pode-se dizer que a partir dessa data a ciência cognitiva foi oficialmente reconhecida.

Dentro da ciência cognitiva existe a Inteligência Artificial (IA), mais ligada a engenharia, onde se pesquisa a cognição artificial, ou seja, onde se aplica o conhecimento teórico das ciências cognitivas para a criação de sistemas artificiais que emulem processos cognitivos com o objetivo de criar sistemas de cognição artificial que consigam representar o conhecimento, ter percepção, imaginação, saber categorizar, ter emoções, aprendizagem e gerenciar a memória, aprender comportamentos como seleção de ação, planejamento, tomar decisões, ter atenção e consciência e aprender a linguagem.

No início se questionava dentro da IA o que deveria ser tratado como escopo de estudo. Havia a idéia inicial de que ela desenvolveria sistemas que iriam substituir a mente humana. Essa idéia inicial não veio a se concretizar, pois os programas desenvolvidos na época não conseguiram atingir esse objetivo. Críticas foram apontadas e um importante teste veio a se tornar um questão crucial dentro da IA: o teste de Turing. O teste determinava que uma máquina era inteligente se um humano não distinguisse se estivesse falando com uma máquina. Outra crítica a IA foi a questão do quarto chinês proposto por John Searle, que questionava se os computadores tinham consciência do que estavam fazendo ou se estavam apenas emulando.

Outras àreas de interesse foram englobadas como pontos de estudo dentro da IA, sendo elas:

- Representação do Conhecimento e Raciocinio
- Aprendizado de Máquina
- Processamento de Linguagem Natural
- Sistemas Especialistas
- Robótica

O estudo relacionado a representação de conhecimento e processamento natural teve um grande avanço dentro da IA com alguns programas que conseguiram algum êxito e levaram a uma nova àrea de interesse denominada Sistemas de Perguntas e Respostas (SPR).

II. SISTEMAS DE PERGUNTAS E RESPOSTAS

Alguns programas desenvolvidos dentro dessa àrea de SPR tornaram-se bem conhecidos e foram marcantes para o desenvolvimento posterior da àrea dentro da IA:

STUDENT – programa em LISP desenvolvido em 1964 para resolver problemas de algebra que reconhecia linguagem natural.

ELIZA – desenvolvido em 1966 esse programa continha uma base de conhecimento interno e fazia o reconhecimento de linguagem natural e respondia as perguntas dos usuários com respostas pré-programadas passando a impressão de que se estava falando com uma pessoa. Utilizava baseamente correspondência de padrão de palavras [3].

SHRDLU – desenvolvido em 1970 e tinha objetivo de mover blocos. Se comunicava através de linguagem natural.

START – foi o primeiro programa de perguntas e respostas instalado na *web*, em 1993. Possuia uma base de conhecimento de várias fontes da *web* e tratava informação estruturada e não estruturada.

ALICE – outro programa de computador que simulava uma conversação com o usuário. Desenvolvido em 1995 e utiliza AIML (*Artificial Intelligence Markup Language*).

Os programas de SPR em geral eram definidos para reconhecer a linguagem natural e transformá-las para um formato estruturado que facilitasse a procura. Outra característica desses programas era a necessidade de ter um banco de dados pré-compilado. Essas duas caracteristicas eram fatores limitantes da tecnologia que os restringiam com relação a precisão da resposta e a engessavam pois necessitavam de uma estrutura de dados e esquema pré-definido.

Uma das principais caracteristicas dos programas de SPR é o tipo de aplicação para que ele é definido. Existem dois tipos de domínios. No domínio aberto os programas são voltados a responder perguntas genéricas. No domínio fechado são voltados a um conjunto específico pré-definido e com escopo fechado.

A estrutura genérica de um SPR está descrito na

Figura 2 abaixo e compõem basicamente de quatro componentes [2]:

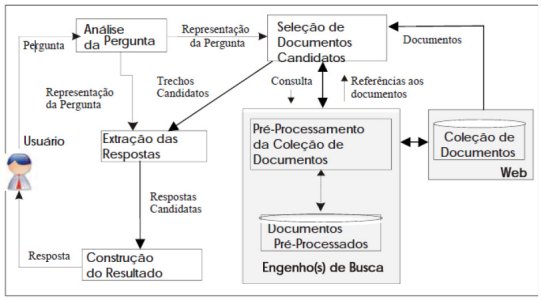

Fig. 2 - Arquitetura de um SPR

Análise de Pergunta – ocorre o tratamento da pergunta identificando o que está sendo perguntado. A pergunta é representada para um formato que possa ser utilizado para pesquisa. Existem três tipos principais de análise efetuadas por esse componente: análise sintática, semântica e pragmática (contextualização). Algumas resoluções gramaticais são efetuadas por esse componente como anáforas e elipses.

Seleção de Documentos – ocorre a seleção de documentos relacionados com o que foi perguntado e envolve a consulta a base de dados do sistema ou fontes externas. Nesse processo são identificados tipos de entidades e relacionamentos dentro do texto.

Extração das Respostas – ocorre a seleção de trechos coletados com o tipo da pergunta com o objetivo de gerar um universo de repostas para a pergunta.

Construção do Resultado – ocorre a seleção da resposta candidata baseada na relação de possíveis respostas encontradas no processo anterior.

Um grande avanço nesses tipos de SPR ocorreu com a introdução da *Text Retrieval Conference* (TREC) em 1991. Voltados para sistemas de aplicações de domínio fechado onde um conjunto de documentos e questões são fornecidos, os SPR são executados contra essa massa e retornam uma lista do documentos melhores ranqueados baseados na exploração de um único fato da pergunta. Os resultados são apresentados em uma conferência anual onde os participantes compartilham experiência e pesquisas.

III. DEEPQA

Em Janeiro de 2011 a IBM participou de um programa de perguntas e respostas chamado *Jeopardy*!, com um sistema chamado Watson, contra os dois maiores vencedores da história do programa.

A característica das perguntas era baseada em domínio aberto, e com vários tipos de perguntas relacionados a conhecimento geral. O sistema do Watson se baseou fortemente em uma arquitetura chamada DeepQA que era baseada em tratamento de informação não estruturada - *Unstructured Information Management Architecture* (UIMA) e várias outras técnicas para tratamento de linguagem natural, recuperação de informação e representação e entendimento do conhecimento. Foi um projeto que durou três anos e consumuiu em torno de vinte pesquisadores.

A Figura 3 apresenta a arquitetura do DeepQA [4].

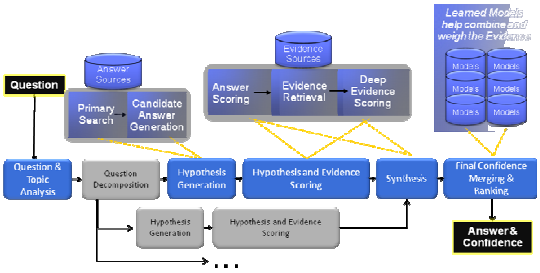

Fig. 3 - Arquitetura DeepQA

A arquitetura compreendia vários componentes e a base de dados usada pelo sistema foi extensivamente manipulada à medida que o sistema foi evoluindo. Os dados iniciais foram coletados da *Wikipedia* e dados adicionais foram sendo adicionados à medida que o sistema não respondia corretamente por falta de informação.

A aquisição do conteúdo para armazenar as respostas candidatas e as fontes de evidências, que fiaram em torno de 400 terabytes, requereu um processo de dois estágios:

Análise da sentença onde o conteúdo coletado foi analisado sintaticamente e estruturas sintáticas sujeito-verbo-objeto foram identificadas e extraídas do conteúdo.

Análise semântica e estatística para então gerar a base de dados, baseada em quadros, para auxiliar na seleção de respostas.

Para exemplificar vejamos a seguinte frase:

"*Einstein, who has published more than 300 scientific papers, won the Nobel Prize for Physics in 1921*",

Aplicando análise sintática os seguintes quadros são gerados: "*Einstein wins Nobel Prize*", "*Einstein publishes papers*".

Aplicando o segundo estágio, análise semântica, os seguintes quadros são gerados: "*the best known thing Einstein wins is a Nobel Prize*", "*scientists publish papers*", "*scientists win Nobel prizes*".

A Figura 4 descreve um exemplo de como uma sentença é analisada sintáticamente [5].

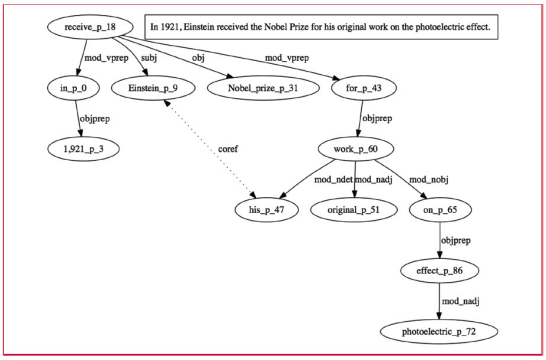

Fig. 4 – Análise Sintática

As informações identificadas, transformadas e expandidas são armazenadas em base de dados em formato de quadros que armazenam as relações de dependência entre as fontes de informações analizadas [6].

A Figura 5 descreve um exemplo dessa estrutura [7].

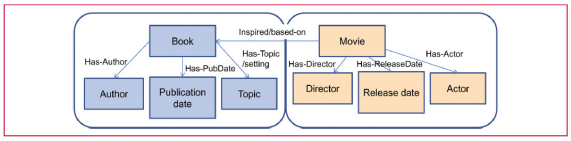

Figure 5 - Quadros

De posse de uma base de dados de conhecimento o primeiro passo para responder a pergunta é a análise da questão e em qual tópico está relacionado. A identificação do tópico é importante pois é fator para auxiliar na identificação da resposta [8]. Nessa etapa alguns elementos críticos da questão são identificados:

- A parte da questão que é a referência da questão, ou seja, o foco.
- Os termos da questão que indicam que tipo de resposta é solicitada.
- A classificação da questão relacionado ao tema perguntado, no caso específico do evento *Jeopardy*!
- Alguns elementos adicionais que possam representar papéis especificos e que requeiram tratamento específico.

O passo seguinte é a decomposição da questão. A questão é decomposta em fatos independentes que vão auxiliar na identificação da resposta. Em casos que há a necessidade identificar fatos independentes da questão

relacionado a uma entidade que irá auxiliar na identificação da resposta são gerados decomposições em sequência onde a resposta de uma decomposição é conectada a outra questão [9].

Para exemplificar vejamos a questão a seguir:

HISTORIC PEOPLE: The life story of this man who died in 1801 was chronicled in an A&E Biography DVD titled "Triumph and Treason".

A decomposição da questão vai gerar as seguintes questões:

Q1: HISTORY PEOPLE (A&E Biography DVD "Triumph and Treason"): This man who died in 1801.

Q2: HISTORY PEOPLE (1801): The life story of this man was chronicled in an A&E Biography DVD titled "Triumph and Treason".

O próximo passo é a geração de hipótese onde se pretende chegar com conteúdo para as respostas da questão. A base de dados gerada, em formato de quadros, é consultada utilizando várias técnicas de procura [10].

Após definido a relação de respostas identificadas o próximo passo é o escore da evidência através da recuperação de provas adicionais para cada resposta identificada. Várias técnicas são utilizadas como racíocinio geoespacial, temporal, popularidade etc [11].

Devido a necessidade, por conta da disputa do *Jeopardy*!, de responder em torno de 3 segundos, essa etapa de geração de hipótese era efetuada em paralelo. Por conta disso ao término dela existe uma etapa de síntese onde as respostas resultantes são classificadas com base em um grau de confiança identificado para cada resposta. Quanto maior o grau de confiança, melhor a resposta.

Um último passo é o ranqueamento das 100 melhores respostas e refinamento até chegar às 5 melhores respostas. Nessa fase modelos pré-definidos são utilizados para auxiliar nos pesos e combinação das respostas [12].

A Figura 6 apresenta o processo final de avaliação de respostas baseado em modelos pré-definidos gerados durante o desenvolvimento da arquitetura.

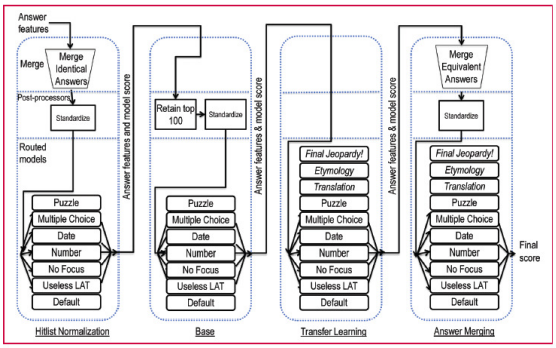

Fig. 6 - Avaliação Final da Resposta

Hitlist Normalization - nessa etapa as respostas são ranqueadas e as 100 melhores são mantidas.

Base - nessa etapa as as respostas sao particionadas em classes de questões.

Transfer Learning – nessa estapa ocorre a transferência de aprendizado para as classes de perguntas incomuns ou raras.

Answer Merging – nessa fase ocorre o grupamento de provas entre respostas equivalentes e seleciona a forma canônica, por exemplo: "John F. Kennedy","J.F.K.", and "Kennedy", corespondem a mesma resposta.

A última etapa efetuada pelo Watson era a resposta da pergunta. Um módulo de estratégia desenvolvido especificamente para o programa avaliava o grau de certeza da resposta e outras variáveis de acordo com o momento do jogo e tomava a ação de responder ou não a pergunta.

IV. RESULTADOS

O projeto do sistema Watson que utilizou a arquitetura DeepQA para participar do jogo *Jeopardy*! ocorreu em Janeiro de 2011 e conseguiu vencer os dois maiores competidores desse jogo.

Além da arquitetura DeepQA apresentada aqui outro componente (não representado neste artigo) foi desenvolvido para gerenciar a estratégia do jogo e decidir quando era oportuno responder a questão.

Como apresentado nesse artigo, várias técnicas foram adicionadas ao DeepQA em comparação a SPR tradicionais, que permitiram que o Watson chegasse num nível de confiança suficiente para poder disputar o jogo.

V. CONCLUSÃO

Foi verificado pouca evolução em SPR desde os primeiros sistemas elaborados por volta da década de 60 até o surgimento das conferências TREC em 1992. A partir dessa data os SPR começaram a adotar técnicas adicionais ao já utilizado mapeamento de padrão.

A arquitetura DeepQA trouxe aos SPR, utilizados em domínio aberto, um novo nível de excelência com a utilização de técnicas adicionais relacionadas a recuperação de informação, reconhecimento de linguagem natural e representação e entendimento do conhecimento.

Pelas referências bibliográficas analisadas foi identificado que o DeepQA consegue ter entendimento do que é perguntado (através de análise sintática e semântica) e, acessando sua base de dados, identificar várias possíveis respostas. Não é totalmente claro o quanto de entendimento o sistema possui, ou seja, se a identificação da resposta está relacionado somente com técnicas ou realmente possui entendimento, como o ser humano.

A utilização do DeepQA para aplicações de domínio fechado pode ser uma ótima oportunidade para trabalhos futuros, aproximando-se dos sistemas especialistas. Outra oportunidade de utilização é na àrea de robótica, como base de dados para tomada de decisão em robôs que necessitem um grande conhecimento prévio.

REFERÊNCIAS BIBLIOGRÁFICAS

[1] Gardner, H., *A Nova Ciência da Mente*. São Paulo: EDUSP, 1995.

[2] RABELO, J. C. B. ; BARROS, F. A., *Pergunte!: Uma Interface em Português para Pergunta-Resposta na Web*. In: V Encontro Nacional de Inteligência Artificial/SBC, São Leopoldo. Anais do XXV Congresso da Sociedade Brasileira de Computação (SBC 2005). v. 1. p. 1114-1117.2005.

[3] J. Weizenbaum, "*ELIZA—A Computer Program for the Study of Natural Language Communication between Man and Machine*", Communications fit the. ACM, 2, pp. 36-45, 1966.

[4] D. A. Ferrucci, *Introduction to 'This is Watson'*, IBM J. Res. & Dev., vol. 56, no. 3/4, Paper 1, pp. 1:1–1:15, May/Jul. 2012.

[5] J. Fan, A. Kalyanpur, D. C. Gondek, and D. A. Ferrucci, *Automatic knowledge extraction from documents*, IBM J. Res. & Dev., vol. 56, no. 3/4, Paper 5, pp. 5:1–5:10, May/Jul. 2012..

[6] J. Chu-Carroll, J. Fan, N. Schlaefer, and W. Zadrozny, *Textual resource acquisition and engineering*, IBM J. Res. & Dev., vol. 56, no. 3/4, Paper 4, pp. 4:1–4:11, May/Jul. 2012

[7] A. Kalyanpur, B. K. Boguraev, S. Patwardhan, J. W. Murdock, A. Lally, C. Welty, J. M. Prager, B. Coppola, A. Fokoue-Nkoutche, L. Zhang, Y. Pan, and Z. M. Qiu, *Structured data and inference in DeepQA*, IBM J. Res. & Dev., vol. 56, no. 3/4, Paper 10, pp. 10:1–10:14, May/Jul. 2012.

[8] A. Lally, J. M. Prager, M. C. McCord, B. K. Boguraev, S. Patwardhan, J. Fan, P. Fodor, and J. Chu-Carroll, *Question analysis: How Watson reads a clue*, IBM J. Res. & Dev., vol. 56, no. 3/4, Paper 2, pp. 2:1–2:14, May/Jul. 2012.

[9] A. Kalyanpur, S. Patwardhan, B. K. Boguraev, A. Lally, and J. Chu-Carroll, *Fact based question decomposition in DeepQA*, IBM J. Res. & Dev., vol. 56, no. 3/4, Paper 13, pp. 13:1–13:11, May/Jul. 2012.

[10] M. C. McCord, J. W. Murdock, and B. K. Boguraev, *Deep parsing in Watson*, IBM J. Res. & Dev., vol. 56, no. 3/4, Paper 3, pp. 3:1–3:15, May/Jul. 2012.

[11] C. Wang, A. Kalyanpur, J. Fan, B. K. Boguraev, and D. C. Gondek, *Relation extraction and scoring in DeepQA*, IBM J. Res. & Dev., vol. 56, no. 3/4, Paper 9, pp. 9:1–9:12, May/Jul. 2012.

[12] D. C. Gondek, A. Lally, A. Kalyanpur, J. W. Murdock, P. Duboue, L. Zhang, Y. Pan, Z. M. Qiu, and C. Welty, *A framework for merging and ranking of answers in DeepQA*, IBM J. Res. & Dev., vol. 56, no. 3/4, Paper 14, pp. 14:1–14:12, May/Jul. 2012.