Um sistema inteligente baseado em ontologia para apoio ao esclarecimento de dúvidas

Marta Talitha C. F. de Amorim[1], Davidson Cury[1], Crediné S. Menezes[1]

[1]Departamento de Informática – Universidade Federal do Espírito Santo (UFES)
Av. Fernando Ferrari, 514, Campus de Goiabeiras – 29.075-910 – Vitória – ES – Brasil

{martatcfa, dedecury, credine}@gmail.com

Abstract. *This paper presents an architecture of a system designed to allow questions and answers automatically. The architecture is supported by ontologies, AIML knowledge base, agents, and information retrieval techniques. The system is guided by the ontology and have ability to update the knowledge base in AIML, as well as to build semi-automatically the ontology of a given domain and its instances.*

Keywords: *Question-answering system, ontologies, concept maps, AIML, agent, information retrieval.*

Resumo. *Este artigo apresenta uma arquitetura de um sistema que visa receber perguntas e dar respostas de forma automática. A arquitetura é apoiada por ontologias, banco de conhecimento AIML, agentes e técnicas de recuperação de informação. O sistema se orienta pela ontologia e tem capacidade de atualizar a base de conhecimento AIML, bem como de construir semiautomaticamente a ontologia de um dado domínio e de suas instâncias.*

Palavras chaves: *Sistema pergunta-resposta, ontologias, AIML, agente e recuperação da informação.*

1. Introdução

Quando as pessoas querem aprender algum conceito, a forma mais comum é usar uma ferramenta de pesquisa, por exemplo: Google, Yahoo, Bing, dentre outros. Uma consulta é submetida a uma ferramenta e a pesquisa retorna uma grande quantidade de páginas relacionadas ao conceito pesquisado. Geralmente as páginas retornadas são listadas e organizadas principalmente baseando-se na combinação de palavras chaves ao invés de utilizar a relevância dos termos consultados. O usuário terá que ler uma grande quantidade de páginas e selecionar a mais apropriada à sua necessidade. Esse tipo de comportamento consome tempo e o usuário-aprendiz perde o foco do seu objetivo [Zhang et al. 2003]. Sistemas de pergunta-resposta e *help-desk* inteligentes pretendem resolver esse problema, apresentando respostas específicas para as perguntas.

Segundo [Maybury, 2004] um sistema de pergunta-resposta (*Question Answering*, QA) tem a seguinte definição: "É um processo interativo humano-computador que engloba uma compreensão precisa de informações dos usuários, tipicamente expressa em uma consulta em linguagem natural, recuperação de

documentos relevantes, dados ou conhecimento a partir de fontes selecionadas, extraindo, qualificando e priorizando as respostas disponíveis dessas fontes e apresentando e explicando respostas de uma maneira eficiente.". Uma outra definição de [Kupiec, 1993] diz que QA são sistemas que tentam responder a uma consulta do usuário que é formulada como uma pergunta, retornando uma frase substantiva apropriada tal como uma localização , uma pessoa , ou data.

A partir dessas definições pode-se conceber um sistema QA que tenha como base uma arquitetura que integra tecnologias de diferentes áreas de conhecimento. Dessa forma pode-se pensar em uma arquitetura geral [Vicedo et al. 2001], na qual cada componente utiliza-se de estratégias diferentes para que respostas possam ser obtidas. Os seguintes elementos de organização podem ser encontrados nas arquiteturas de um sistema QA:

1. base de dados: ontologias [Lopez et al. 2007], web [Zhang et al. 2003], híbrido [Chu-Carroll et al. 2003], entre outros.

2. sistema de análise da questão: Processamento de linguagem natural [Woods, 1978], algoritmos *string metrics* [Lopez et al. 2007], expressões regulares [Lopez et al. 2007], técnicas de Recuperação de Informação (RI) [Wilkens,2010], entre outros.

3. sistema para seleção e extração de resposta: Recuperação de informação baseado em vetor de pesos, ontologias [Lopez et al. 2007], aprendizagem de máquina [Prager et al. 2000], métodos para seleção das respostas [Prager et al. 2000], entre outros.

4. sistema para geração das respostas: Combinação de som, imagens e texto [van den Bosch, 2011], textual e web [Lopez et al. 2007], vídeos [Li et al. 2010], entre outros.

Além disto, existem diversas abordagens na literatura para construção de sistemas de pergunta-resposta: Helpdesk inteligente [Greer, 2000]; *Frequently Ask Questions* (FAQ) [Berger, 2000], teste da compreensão da leitura [Chen, 2010], domínios restritos [Cao et al. 2011], domínios irrestritos [Nyberg, 2005] e para fins educacionais [Grignetti, 1975], para apoiar a BI (*Business Inteligent*) [Vila, 2011] dentre outros.

Este artigo apresenta a arquitetura de um sistema de pergunta-resposta independente de domínio e está organizado da seguinte forma: Na Seção 2 apresenta-se o estado da arte dos sistemas de pergunta-resposta em geral. A Seção 3 descreve a arquitetura proposta. A Seção 4 esboça uma instanciação da arquitetura para o domínio do sistema Linux. Por fim, na Seção 5 citam-se os trabalhos futuros e as considerações finais do artigo.

2. Revisão da Literatura dos Sistemas de Pergunta-Resposta

Utilizaram-se aqui apenas trabalhos disponíveis para *download*. Apresenta-se, a seguir, uma explanação breve dos sistemas de pergunta-resposta mais atuais. A comparação entre sistemas se baseou na métrica "percentual de questões respondidas corretamente" (*recall),* conforme relatado nos artigos descritores desses sistemas.

O FREyA [Damljanovic et al. 2010] é um sistema QA que combina análise sintática e ontologias a fim de se adequar a um novo domínio com o mínimo de customizações. A dinâmica do sistema pode ser resumida nos seguintes passos: identifica e verifica os conceitos da ontologia, gera a consulta, identifica o tipo da resposta e apresenta o resultado para o usuário. Os testes obtiveram um recall de 92.4% sobre um total de 250 questões.

O trabalho de [Oh et al. 2011] apresenta uma arquitetura na qual o componente da análise da questão emprega várias técnicas de análise linguística (*tagging*, *chuncking*, entre outras). Um algoritmo de aprendizagem é utilizado para auxiliar a seleção da resposta. Das cinco estratégias de teste, o melhor resultado obtido (*Automatic strategy-drive*) foi realizado com 500 questões e apresentou 84% de respostas corretas.

Em [Liu et al. 2010] é proposto um método de recuperação de QA baseado em *frequently ask questions* (FAQ). Geralmente sistemas QA baseados em FAQ, combinam a pergunta do usuário com as do banco de dados de pergunta-resposta, e retornam as respostas ao usuário. O sistema é estruturado em três componentes: interpretação da questão, recuperação da informação e gerenciamento do banco de dados de FAQ (atualizar e estabelecer). O teste apresentou um percentual médio das respostas corretas de 72.1%.

O PowerAqua [Lopez et al. 2011] é uma evolução de outro sistema chamado Aqualog, um sistema QA baseado em ontologia. Na arquitetura do PowerAqua, o componente análise da questão utiliza um componente linguístico para processar a consulta. A saída desse componente é um conjunto de triplas linguísticas (< sujeito, predicado, objeto>) que é mapeado para a consulta do usuário. Assim é possível realizar buscas das respostas em bases OWL[1] e RDF[2]. Os resultados obtidos nos testes apresentaram 48 (69,5%) questões respondidas das 69 questões totais.

O sistema QA apresentado por [Konopík et al. 2010] é especialmente efetivo em respostas para questões "*Wh*" (O que, Quem , Quando, Onde, Por que, De quem, Qual e Como) sobre pessoas, datas, nomes e localizações. A resposta é construída a partir de dados colhidos na internet, em ontologias públicas, no conhecimento da linguagem Tcheca. Na apresentação do resultado, o usuário avalia se a resposta é correta ou incorreta. Essa avaliação é armazenada e utilizada para otimizar o sistema. Um teste realizado com um conjunto de 100 questões apresentou 64% de respostas corretas.

Uma síntese dos sistemas apresentados pode ser encontrada na Tabela 1, onde as escolhas com respeito aos elementos de organização das arquiteturas são explicitados.

Os vários sistemas apresentados trazem, cada um, inovações importantes para o tratamento do problema. Neste trabalho busca-se incorporar contribuições importantes de cada um, integrando-as por intermédio de uma arquitetura multiagente conforme se descreve na seção a seguir.

Tabela 1. Elementos de Organização das Arquiteturas de QA Encontradas

Sistema	Base de dados	Análise da Questão	Seleção e Extração das Respostas	Geração das Respostas
FREyA [Damljanovic et al. 2010]	*Dataset* (http://www.mooney.net/geo)	*Parser*, raciocínio na ontologia, algoritmo que mede similaridade entre string, identifica o tipo da questão.	Gera e realiza consulta SPARQL, aprendizagem de máquina semi-supervisionada.	Web/textual
(sem nome) [Oh et al. 2011]	Web	Técnicas de análise linguística (*POS tagging, chunking, named entity tagging*),análise semântica, entre outros.	Aprendizagem de máquina, cálculo baseado em peso, entre outros.	Textual
(em nome) [Liu et al. 2010]	*Dataset* de FAQ relacionado ao domínio de faculdade e universidade.	Remoção de *stopwords*, determinação de palavras chaves, ontologias, entre outros.	Gera e realiza consulta SPARQL, cálculo de similaridade baseado em métodos estatísticos e semânticos.	Textual
PowerAqua [Lopez et al. 2011]	DBpedia (http://wiki.dbpedia.org) , entre outros.	Análise sintática, expressões regulares, Wordnet, algoritmo que mede similaridade entre string, entre outros.	Consulta SPARQL, algoritmo selecionar resposta, entre outros.	Web/textual
(sem nome) [Konopík et al. 2010]	Web (ex: Google)	Lematização, Wordnet, classificação e extração das entidades (*named entity*) em categorias pré-definidas, *POS Tagging*, entre outros.	Processamento estatístico	Textual

3. A Arquitetura Proposta

A arquitetura proposta difere-se das analisadas pela sua generalidade e também por combinar um banco de ontologias com uma família de agentes de software. Com o uso desses agentes pretende-se flexibilizar a arquitetura, favorecendo assim o uso de mecanismos inteligentes específicos, sintonizados com as necessidades de cada ação do sistema. Espera-se com isto obter um *framework* capaz de viabilizar a construção de sistemas pergunta-resposta mais abrangente e com respostas mais precisas.

A Figura 1 apresenta a arquitetura do sistema. Nela podem-se destacar dois grandes blocos. O sistema QA propriamente dito e o sistema de aquisição de conhecimento (inicial e continuado). Nas Subseções 3.1 e 3.2 a seguir é descrito cada um desses subsistemas.

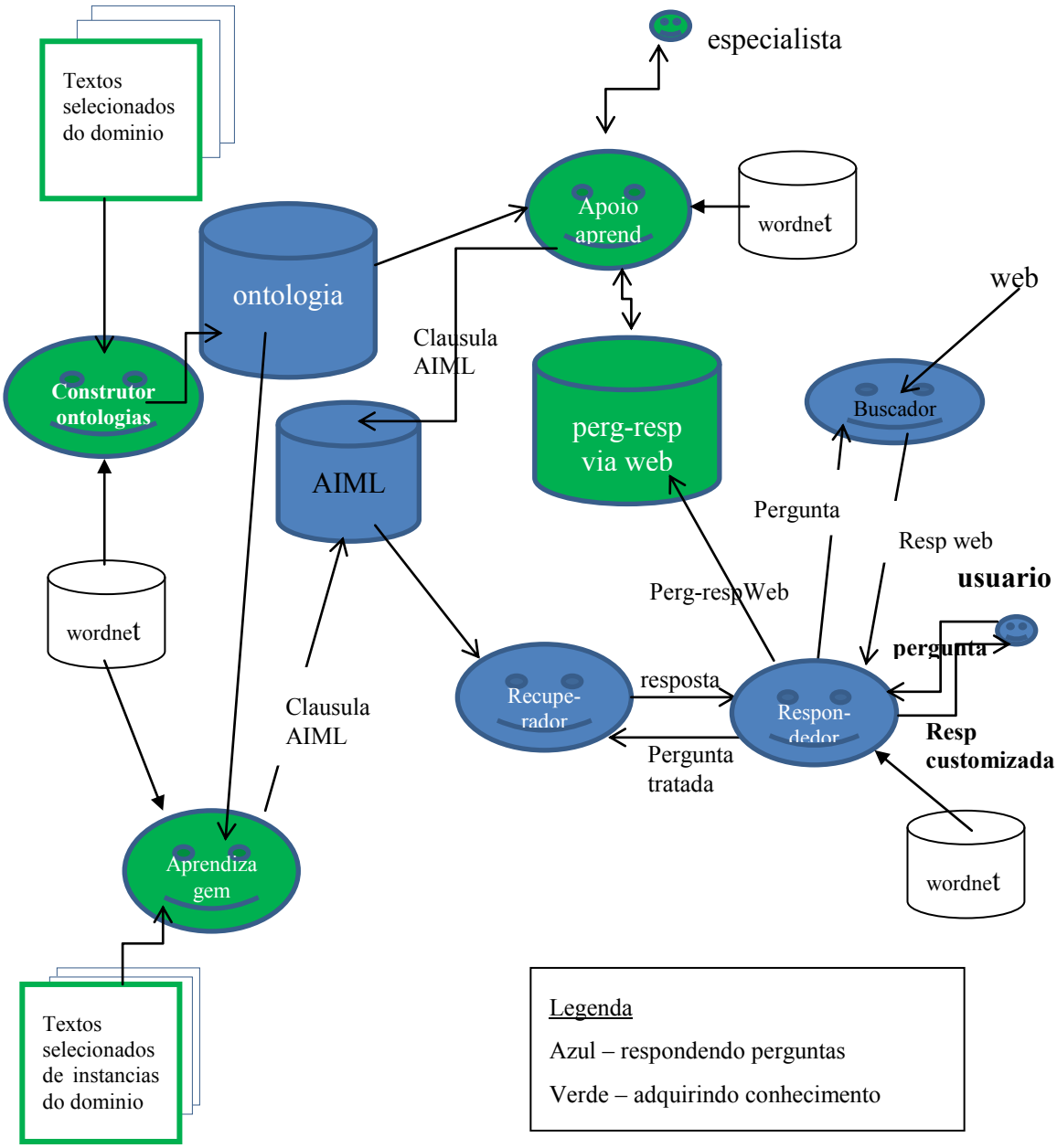

Figure 1. A arquitetura do sistema de pergunta-resposta

3.1 Respondendo perguntas

O trabalho começa a partir de uma consulta do usuário. De posse dessa consulta o agente respondedor acessa a *wordnet* para resolver questões de sinonímia. A pergunta, agora tratada, é mandada para o agente recuperador que consulta o banco AIML (*Artificial Intelligence Markup Language*). Se uma resposta adequada é encontrada, ela é

devolvida ao respondedor que a encaminha ao usuário, encerrando com isso o trabalho do sistema.

Quando não é encontrada nenhuma resposta, o agente recuperador devolve ao agente respondedor uma resposta evasiva, significando que não encontrou uma resposta adequada à pergunta. O respondedor percebe a natureza semântica da resposta e a envia ao agente buscador da web. Ao obter uma resposta da web, o buscador a devolve ao respondedor que, por sua vez, realiza duas ações: 1) encaminha-a ao usuário, sinalizando-lhe tratar-se de uma resposta obtida na web; 2) envia-a, juntamente com a pergunta original, ao banco de perguntas-respostas da web, sinalizando com isso as situações onde o sistema não conseguir responder precisamente a consulta de um usuário. Em se tratando de um usuário não anônimo seu endereço de *email* é também registrado para futuros contatos.

3.2 Adquirindo Conhecimento

A aquisição de conhecimento é multinivelada. Em um dos níveis constrói-se uma ontologia de domínio. No outro nível constroem-se cláusulas em AIML, com base em assuntos específicos. Em um terceiro nível ainda, é feita a aquisição continuada de conhecimento.

O agente construtor de ontologias recebe um conjunto de textos sobre o domínio e gera uma ontologia. A geração da ontologia do domínio em questão pode ser realizada de maneira semiautomática. Pretende-se usar para este fim a ferramenta ONTOCMAPS [Zouaq, 2011]. A transformação realizada pelo componente ONTOCMAPS inicia pela conversão do texto em mapas conceituais [Kowata et al. 2009] que são, por sua vez, convertidos em uma ontologia *lightweight*, em OWL. Os mapas conceituais são utilizados como uma linguagem intermediária para representação do conhecimento e nesta fase a validação humana é requerida.

O agente de aprendizagem recebe um conjunto de textos selecionados sobre instâncias do domínio. Apoiado na ontologia do domínio, ele constrói cláusulas AIML e as armazena no banco AIML. O trabalho de ambos é amparado pelo uso da *wordnet* para dirimir questões de sinonímias.

O agente de aquisição continuada de conhecimento entra em cena toda vez que um novo par pergunta-resposta é depositado no banco correspondente. Com este par, consulta a ontologia de domínio que lhe dá subsídios para a construção de uma nova cláusula AIML que será armazenada no banco AIML. Caso o agente não consiga gerar a cláusula, ou não consiga se decidir pela cláusula mais apropriada, ele ativa o agente especialista humano (especialista) para obter a resposta adequada. Quando o ambiente consegue adquirir um novo conhecimento por este mecanismo, os usuários envolvidos são contatados, oferecendo-lhes a nova resposta.

3.3. Banco AIML

Baseado no trabalho desenvolvido por [Teixeira, 2005], que apresenta uma proposta de construção de bases de conhecimento AIML para um *chatterbot*. Na nova arquitetura esta base será expandida. O objetivo dessa base é apoiar a resolução das perguntas fornecendo um cache de respostas.

3.4 Wordnet

A Wordnet é uma grande base de dados léxica que contém substantivos, verbos, adjetivos e advérbios. Essa grande base tenta organizar a informação léxica em termos dos significados das palavras ao invés das formas das palavras, embora a morfologia seja considerada. Nesta arquitetura a Wordnet funciona como um dicionário dos sinônimos e hipônimos. Se um termo da pergunta não for encontrado na base, então o termo sinônimo é retornado a fim de dar continuidade na resolução da pergunta. Se o termo sinônimo não for encontrado o agente deve aprender o termo e incluí-lo na base. Dado que a wordnet para a língua portuguesa ainda se encontra em construção, aqui será usada a de língua inglesa.

4. Um estudo de caso

Em um trabalho anterior [TEIXEIRA, 2005] construiu-se um sistema QA para responder perguntas sobre o sistema Linux, adotando uma abordagem específica baseada em AIML, usando como fonte apenas um conjunto de perguntas e respostas disponível na web. O sistema não era dotado de mecanismo de aquisição continuada de conhecimento.

Com a intenção de ter um elemento a mais de comparação para nossos estudos, o primeiro caso que será realizado será sobre o sistema operacional Linux. Desta vez, entretanto, serão usados textos diversos, ao invés de um conjunto de perguntas já compiladas.

Igualmente, a titulo de experimentação de ferramentas novas, optou-se pelo uso da ONTOCMAPS [ZOUAK, 2010], uma ferramenta geradora da ontologia de domínio. Esta ferramenta promove a geração semiautomática de ontologias, usando mapas conceituais como representação intermediaria da ontologia. Uma vez gerado o mapa, é necessária a intervenção do especialista humano para a validação do mapa. ONTOCMAPS encontra-se em fase experimental.

5. Conclusão e Trabalhos Futuros

A web é aparentemente uma fonte ideal de respostas para uma grande variedade de questões. Geralmente, o que ocorre é que, para uma dada pergunta, o usuário recebe uma imensa quantidade de sites como resposta, o que vai lhe exigir uma verdadeira atividade de garimpo da resposta mais adequada à sua pergunta.

Alguns sistemas QA mais sofisticados já estão no mercado. Aqueles disponíveis para *download* foram analisados neste trabalho. Essa análise deu subsídios para a modelagem da arquitetura aqui proposta.

A arquitetura proposta trabalha com um conjunto de bancos de conhecimento baseados em ontologias. Um desses bancos contém conhecimento de um dado domínio enquanto que os demais contêm conhecimentos sobre instâncias do domínio. Uma família de agentes se encarrega da busca e da construção de respostas mais precisas aos questionamentos do usuário, bem como das atividades de aprendizagem, ou seja, da expansão da base de conhecimento AIML e da construção de novas ontologias.

Um estudo de caso será feito, como próxima tarefa, para o domínio de sistemas operacionais, usando uma de suas instâncias, o Linux. As ontologias serão construídas de modo semiautomático, pelo uso da ferramenta ONTOCMAPS. O protótipo será implementado em PHP.

Os trabalhos futuros convergem para a demonstração dos resultados de *recall* da arquitetura proposta, equiparando-os com o estado da arte.

6. Referências

Baeza-Yates, R., Ribeiro-Neto, B. (1999). Modern Information Retrieval, Addison Wesley, USA.

Berger, A., Mittal, O. V. (2000). Query-Relevant summarization using FAQs (2000). In Proceedings of the 38th Annual Meeting on Association for Computational Linguistics, Hong Kong, China, pp. 294-301.

Chen, Y. (2010). Improving reading comprehension using knowledge model. In Proceedings of the 5th International Symposium Advances in Computation and Intelligence, Wuhan, China, pp. 370-379.

Chu-Carroll, J., Ferrucci, D., Prager, J., Welty, C. (2003). Hybridization in Question Answering Systems. AAAI Press.

Damljanovic, D., Agatonovic, M., Cunningham, H. (2010). Natural Language Interfaces to Ontologies: Combining Syntactic Analysis and Ontology-based Lookup through the User Interaction. In Proceedings of the 7th Extended Semantic Web Conference (ESWC 2010), Creete, Greece, pp. 106-120.

FocaLinux, “Guia Foca GNU/Linux”. Disponível em: <http://www.guiafoca.org>. Acesso em 24 de junho de 2011.

Gómez-Pérez, A., Fernández-López, M., Corcho, O. (2004). Ontological Engineering, Springer Verlag, London, UK.

Greer, J.,McCalla, G., Cooke, J., Collins, J., Kummar, V., Bishop, A., Vassileva, J. (1998). The Intelligent Helpdesk: Supporting Peer-Help in a University Course. Department of Computer Science, University of Saskatchewan, Canada.

Grignetti, M. C., Hausmann, C., Gould, L. (1975). An "intelligent" on-line assistant and tutor: NLS-scholar. ACM Digital Library Publishers.

Konopík, M., Rohlík, O. (2010). Question Answering for Not Yet Semantic Web. In Proceedings of the 13th International Conference Text, Speech and Dialogue, Brno, Czech Republic, pp. 125-132.

Kowata, J. H., Cury D., Boeres M. C. S. (2009). Caracterização das Abordagens para Construção (Semi) Automática de Mapas Conceituais. XX Simpósio Brasileiro de Informática na Educação, Florianopolis, Brasil.

Kupiec, J. (1993). MURAX: A robust linguistic approach For question answering using online encyclopedia. In Proceedings of the 16th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval, Pennsylvania, USA, pp. 181-190.

Liu, H., Lin, X., Liu C. (2010). Research and Implementation of Ontological QA System based on FAQ. In Journal of convergence information technology, Vol. 5, No. 3, pp. 79-85.

Lopez, V., Fernández, M., Stieler, N., Motta, E. (2011). PowerAqua: supporting users in querying and exploring the Semantic Web content. KMi The Open University, United Kingdom.

Maybury, M. T. (2004). New Directions in Question Answering. MIT Press, California, USA.

Moise, M., Gheorghe, C., and Zingale, M. (2010). Developing question answering (QA) systems using the patterns. In WSEAS International Conference on Computers, Vol. 9, No. 7, pp. 726-737.

Novak, J. D., Cañas, A. J. (2006). The Theory Underlying Concept Maps and How to Construct and Use Them. Acesso em 6 de junho de 2009, disponível em <http://cmap.ihmc.us/Publications/ResearchPapers/TheoryUnderlyingConceptMaps.pdf>

Nyberg, E., Mitamura, T., Frederking, R., Bilotti, M., Hannan, K., Hiyakumoto, L., Ko, J., Lin, F., Pedro, V., Schlaikjer, A. (2005). JAVELIN I and II Systems at TREC 2005. In Proceedings of Text Retrieval Conference.

Oh, H.-J., Jang, M.-G., and Myaeng S. H. (2011). Effects of answer weight boosting in strategy-driven question answering. Information Processing and Management, Elsevier.

Prager, J., Brown, E., Coden, and Raved, D. (2000). Question-Answering by predictive annotation. In Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Athens, Greece, pp. 184-191.

Teixeira, S. (2005). Chatterbots – Uma proposta para a construção de bases de conhecimento. Dissertação de Mestrado apresentado a Programa de Pós-Graduação em Informática do Centro Tecnológico, Universidade Federal do Espírito Santo.

van den Bosch, A., Bouma, G. (2011). Interactive Multi-modal Question Answering. Spring Verlag, London, UK.

Vicedo, J. L., Mollá, D. (2001). Open-Domain Question-Answering Technology: State of the Art and Future Trends. University of Alicante, Spain.

Vila, K., Ferrández, A. (2011). Model-driven restricted-domain adaptation of question answering systems for business intelligence. In Proceedings of the 2nd International Workshop on Business intelligencE and the WEB, Uppsala, Sweden, pp. 36-43.

Zhang, D., Lee, S. W. (2003). A web-based question answering system. In Proceedings of the SMA Annual Symposium 2003, Singapore.

Zouaq, A., Gasevic, D., Hatala, M. (2011). Ontologizing Concept Maps using Graph Theory. In Proceedings of the 2011 ACM Symposium on Applied Computing, New York, USA, pp. 1687-1692.

Wilkens, R., Villavicencio, A. (2010). Question answering for Portuguese: How much is needed. In Proceedings of the 20th Brazilian Symposium on Artificial Intelligence, São Bernando do Campo, Brazil, pp. 173-182.

Woods, A. W. (1978). Semantics and Quantification in Natural Language Question Answering. Morgan Kaufmann Publishers, San Francisco, CA, USA.