Linguateca e Processamento de Texto Livre na Área da Saúde: Alguns Comentários e Sugestões

Liliana Ferreira, António Teixeira
Instituto de Engenharia Electrónica e Telecomunicações de Aveiro/
Departamento de Electrónica, Telecomunicações e Informática
Universidade de Aveiro,
Campus Universitário de Santiago,
3810-193 Aveiro, Portugal
{lsferreira, ajst}@ua.pt

1 Introdução

A crescente utilização de sistemas de informação na área da saúde, levou a um aumento significativo da informação médica disponível electronicamente sob a forma de texto livre. A necessidade de gerir e processar grandes quantidades de dados motiva o recente interesse em aproximações semânticas, cujos principais objectivos se prendem com a redução de erros médicos, a melhoria da eficiência médica e uma maior satisfação e segurança dos pacientes. As tecnologias de Rede Semântica auxiliam na obtenção destes objectivos através de múltiplas ontologias populadas, anotação semântica automática de documentos e processamento de regras, entre outros.

A experiência do Instituto de Engenharia Electrónica e Telemática de Aveiro (IEETA) no desenvolvimento de sistemas de informação na área da saúde contribuiu para o recente interesse no desenvolvimento de sistemas de processamento de texto livre, capazes de extrair informação pertinente de um grande volume de textos médicos em linguagem natural. Um exemplo desta motivação é o projecto Rede Telemática da Saúde (RTS) [2] que pretende disponibilizar, de forma segura, o acesso a informação clínica e promover a comunicação entre profissionais de saúde credenciados, bem como envolver o cidadão na gestão da sua saúde, contribuindo para um melhor acesso aos cuidados de saúde. Esta Rede implementa um Processo Clínico Electrónico Regional resumido, que agrega informação clínica do utente, proveniente de várias fontes de informação clínica geograficamente distribuídas pelas várias instituições de saúde regionais. Entre outros, a RTS disponibiliza acesso a cartas de alta, boletins de análises clínicas, relatórios de exames de imagiologia, etc.

2 MedAlert

Motivado pela Rede Telemática de Saúde, está actualmente em desenvolvimento no IEETA o projecto MedAlert - Sistema de Processamento de Linguagem Médica, que tem por objectivo a utilização de técnicas de extracção automática de informação de textos médicos, de modo a inferir, de uma forma automática, irregularidades/dúvidas suscitadas pelas decisões tomadas pelos profissionais de saúde. Este sistema, que deverá tomar a forma dum módulo escalável e adaptável a diferentes configurações de sistemas de informação hospitalares, deverá usar técnicas de Processamento de Linguagem Natural (PLN) para extrair informação de um amplo conjunto de textos médicos disponibilizados pela RTS, particularmente cartas de alta e textos contendo directivas

médicas. Esta informação, bem como a proveniente recursos externos como ontologias e outras fontes de conhecimento médico, deverá ser utilizada no suporte e validação de decisões médicas.

Deste modo, tornou-se essencial o desenvolvimento de uma ferramenta capaz de extrair informação de uma forma automática a partir de texto livre. Na inventariação das ferramentas existentes e mais recentes na área, nomeadamente das ferramentas desenvolvidas para o português, a Linguateca teve um papel preponderante. Embora a delimitação de área, neste caso a medicina, imponha a necessidade de usar ferramentas direccionadas e o desenvolvimento de módulos específicos que identifiquem a informação relativa à aplicação em particular, o processamento inicial do texto pode ser feito com recurso a ferramentas de análise morfológica e sintáctica do Português, como as que já se encontram disponibilizadas e listadas pela Linguateca. Também os manuais e a diversa literatura apresentada pela Linguateca contribuíram para uma aprendizagem mais diversificada e célere.

No entanto, a escolha das ferramentas a usar no desenvolvimento de tal sistema não recaiu sobre os recursos disponibilizados pela Linguateca, mas sim na utilização e adaptação das componentes existentes em plataformas de processamento de informação não estruturada, como por exemplo o GATE [4] ou UIMA [5]. Estas plataformas permitem uma adaptação a diferentes sistemas operativos e disponibilizam as várias componentes de um sistema de processamento de linguagem natural em ambientes de desenvolvimento gráfico, facilitando a aprendizagem e a adaptação a diferentes línguas e domínios. Deste modo, considerou-se mais vantajoso e potencialmente mais rápida a utilização e adaptação de um ambiente deste tipo, do que a criação de algo de raiz.

No caso do MedAlert, começou por se desenvolver um sistema tendo por base a plataforma GATE. O GATE é uma infra-estrutura para o desenvolvimento de componentes de software, que processam linguagem natural, em desenvolvimento na Universidade de Sheffield desde 1995 e utilizado numa grande variedade de projectos. A arquitectura consiste em vários recursos de processamento independentes do domínio e aplicáveis a várias línguas, como o Tokenizador e a Separação em Frases. No entanto, o processamento principal, em particular o Reconhecimento de Entidades Mencionadas, foi efectuado com recurso a almanaques e um conjunto de regras gramaticais desenvolvidas em JAPE (Java Annotations Pattern Language) [3] que consideram conteúdos específicos da língua e do domínio, neste caso a medicina.

Recentemente, foi realizada uma experiência na área da vacinação com o objectivo de extrair informação do Plano Nacional de Vacinação (PNV) [7]. A informação considerada relevante foi extraída e associada às entidades ACRONIMO, IDADE, PARTE_CORPO, DOENCA, DOSE, INTERACCAO, REACCAO e PESO de acordo com o conteúdo expresso. A figura 1 apresenta a interface gráfica do GATE com um excerto do PNV anotado com várias entidades e a tabela 1 os resultados obtidos, em termos de precisão e abrangência.

Tabela 1: Resultados da tarefa de Extracção de Informação para cada uma das entidades definidas.

	DOENCA	ACRONIMO	IDADE	PARTE_CORPO	DOSE	INTERACCAO	REACCAO	PESO
Saídas correctas	225	294	181	14	90	10	156	8
Parcialmente correctas	0	0	1	0	3	0	0	0
Em falta	0	0	6	0	0	0	0	0
Total	**225**	**294**	**188**	**14**	**93**	**10**	**156**	**8**
Abrangência	1,00	1,00	0,96	1,00	0,97	1,00	1,00	1,00
Precisão	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00
Medida F	**1,00**	**1,00**	**0,98**	**1,00**	**0,98**	**1,00**	**1,00**	**1,00**

O desenvolvimento de uma arquitectura semelhante às referidas, que integre alguns dos recursos já existentes para o português, ou a produção de recursos tendo por base ambientes já existentes e que permitam a adaptação a diferentes técnicas e áreas de uma forma rápida e facilmente adaptável, permitiria estruturar os recursos actuais de uma forma mais útil. Estas arquitecturas permitem a utilização da tecnologia não só por profissionais da área, mas também por grupos que procuram sistemas eficientes e prontos a usar, podendo contribuir para uma maior divulgação da área e da tecnologia. Um exemplo da utilidade destas arquitecturas constituídas por módulos é a sua aplicação no ensino, em particular no ensino de pós-graduação, pelo facto de

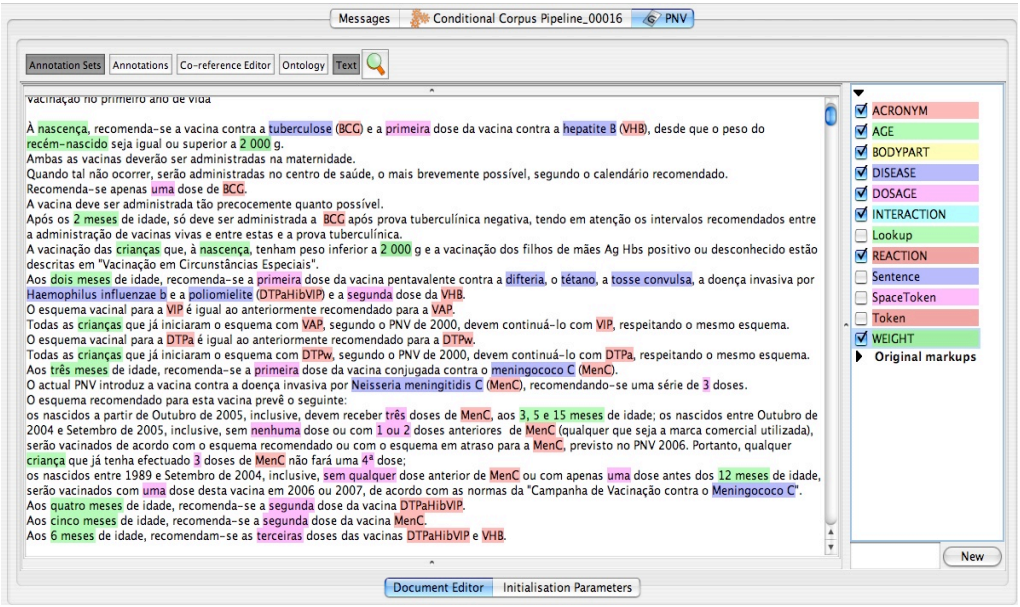

Figura 1: Interface gráfica do GATE com um excerto do Plano Nacional de Vacinação anotado.

permitir a produção de recursos independentes e facilmente integráveis e escaláveis. Embora a Linguateca disponibilize de uma forma pública os recursos criados, a sua utilização e/ou adaptação pressupõe muitas vezes a existência de informação e de conhecimento que os interessados podem não possuir, podendo tornar o processo moroso e complicado.

Ainda na área da vacinação, e no âmbito do MedAlert, foi recentemente desenvolvido trabalho no sentido de criar uma representação conceptual das directivas contidas no PNV. Para tal, o PNV foi analisado manualmente e modelado segundo os conceitos e relacionamentos que descreve. Foi assim criada uma ontologia contendo todas as classes de vacinas, interacções e alergias descritas no PNV. A tarefa de popular a ontologia foi dividida em dois passos. Primeiro, informação automaticamente extraída do PNV foi adicionada à ontologia através da associação entre a classe e a entidade identificada. Posteriormente, adicionaram-se os relacionamentos entre as instâncias, usando uma abordagem baseada em Procura de Padrões Frequentes. Para tal, identificou-se no texto anotado padrões frequentes de entidades mencionadas, usando Procura de Regras de Associação [1]. Procurou-se identificar regras como, por exemplo, *doença* $\Rightarrow$ *vacina (80%)*, indicando, neste caso, que 4 em cada 5 vezes que uma doença é mencionada (ex. Tuberculose) é seguida pela referência a uma vacina (ex. BCG). Seguindo o exemplo, poder-se-ia concluir que a vacina mencionada, BCG, combate a doença Tuberculose, e automaticamente inferir e adicionar o correspondente triplo RDF [6]. Esta foi, no entanto, apenas uma primeira experiência na tentativa de automatizar o processo de adição de relacionamentos entre instâncias em ontologias. No seguimento deste trabalho de criação de ontologias para o Português, o caminho usual da Linguateca de disponibilização e publicação do processo e ferramentas utilizadas, de que é exemplo o PAPEL, será certamente útil na continuação de criação das ontologias para a nossa área de aplicação.

3 Participação no II HAREM

Recentemente, participamos pela primeira vez na avaliação conjunta de reconhecedores de entidades mencionadas organizada pela Linguateca, o II HAREM. Este modelo de avaliação, em que vários grupos comparam o progresso dos seus sistemas usando uma métrica consensual, representou uma importante oportunidade para perceber quais os desafios inerentes ao reconhecimento

de nomes próprios em textos não especializados na área da saúde e deste modo desenvolver diferentes técnicas de delimitação e classificação destas entidades. No caso do sistema desenvolvido em Aveiro, o desafio representou o desenvolvimento de um sistema capaz de recorrer a fontes de conhecimento externas, como a Wikipedia, de modo a melhorar a classificação e a diminuir a utilização de listas e almanaques. Este sistema, denominado REMMA – Reconhecimento de Entidades Mencionadas do MedAlert, foi desenvolvido tendo por base o sistema de processamento de linguagem não estruturada Apache UIMA [5]. UIMA é uma plataforma para o desenvolvimento de sistemas de software capazes de analisar grandes volumes de informação não estruturada. O REMMA contem, entre outras, uma componente capaz de explorar a Wikipedia como fonte de conhecimento. A impossibilidade de construir ou aceder a um almanaque de grande dimensão e qualidade, motivou a decisão de extrair categorias e tipos através da análise da primeira frase do artigo Wikipedia. Esta experiência, embora direccionada a textos não especializados, permitiu perceber a utilidade de tais abordagens e procurar recursos e soluções semelhantes para a área em que nos concentramos.

4 Conclusões e sugestões finais

Não sendo certamente possível, ou mesmo desejável, que a Linguateca desenvolva corpos e ferramentas para domínios específicos como o nosso, atrevemo-nos a sugerir que seja efectuada uma inventariação o mais completa possível de módulos e sistemas existentes e disponíveis para utilização por todos; concentração do esforço da Linguateca na criação de recursos e ferramentas ainda não disponíveis; e que haja um esforço de criação de um sistema integrado. Para facilitar a avaliação, para além dos eventos como o HAREM, seria importante a criação de directivas genéricas que facilitariam a avaliação comparativa em domínios mais específicos como o que nos interessa. Particularmente, a criação de directivas sobre construção de "Colecções Douradas", de métricas de avaliação e possivelmente um sítio para disponibilização destes recursos possibilitaria o desenvolvimento por parte dos grupos interessados de acções de avaliação conjunta em áreas especificas, como a Medicina.

Referências

[1] Rakesh Agrawal e Ramakrishnan Srikant. Fast algorithms for mining association rules. Em *20th International Conference Very Large Data Bases*, 1994.

[2] João Paulo Silva Cunha, Isabel Cruz, Ilídio Oliveira, António Sousa Pereira, César Telmo Costa, Ana Margarida Oliveira e Amândio Pereira. The RTS project: Promoting secure and effective clinical telematic communication within the Aveiro region. Em *Em eHealth 2006 High Level Conference*, Málaga, Espanha, Maio de 2006. p. 1–10.

[3] Hamish Cunningham, Diana Maynard, Kalina Bontcheva, Valentin Tablan e Cristian Ursu. The GATE User Guide. http://gate.ac.uk/, 2000.

[4] Hamish Cunningham, Diana Maynard, Kalina Bontcheva e Valentin Tablan. GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications. Em *40th Anniversary Meeting of the Association for Computational Linguistics (ACL'02)*, Lisboa, Julho de 2002.

[5] David Ferrucci e Adam Lally. UIMA an architectural approach to unstructured information processing in the corporate research environment. *Natural Language Engineering*, 10(3–4): 327–348, 2004.

[6] Ora Lassila e Ralph Swick. Resource Description Framework (RDF) Model and Syntax. W3C, World Wide Web Consortium. http://www.w3.org/TR/WD-rdf-syntax/, 1998.

[7] Direcção Geral da Saúde. Programa Nacional de Vacinação. Orientações Técnicas, 2006.