Interface Ubíqua, Interoperativa e Escalável para uma Plataforma de Serviços PLN em *Big Data*

Proposta de Dissertação de Mestrado

Orientador: Sebastião Pais
Co-Orientador: João Paulo Cordeiro

Departamento de Informática @ UBI
2017/2018

1 Processamento da Linguagem Natural (PLN)

PLN é uma subárea da ciência da computação, inteligência artificial e da linguística que estuda os problemas da geração e compreensão automática de línguas humanas naturais. Sistemas de geração de linguagem natural convertem informação de base de dados de computadores em linguagem compreensível ao ser humano e sistemas de compreensão de linguagem natural convertem ocorrências de linguagem humana em representações mais formais, mais facilmente manipuláveis por programas de computador.

1.1 Aplicações

Nesta secção apresenta-se alguns dos trabalhos que se inserem no contexto de PLN. Note que alguns deles têm aplicações no mundo real, enquanto outros servem mais frequentemente como tarefas secundárias que são usadas para auxiliar na resolução de tarefas maiores. O que distingue essas tarefas de outras tarefas potenciais e reais de PLN não é apenas o volume de pesquisa dedicado a elas, mas o fato de que para cada uma há tipicamente uma definição de problema bem especificada, uma métrica padrão para avaliar a tarefa, corpora padrão em que a tarefa pode ser avaliada e as competições dedicadas à tarefa específica.

- Sumarização automática: Produz um resumo legível de uma parte do texto.
- Análise do Discurso: Esta rubrica inclui uma série de tarefas relacionadas. Uma tarefa é identificar a estrutura discursiva do texto conectado, isto é, a natureza das relações discursivas entre frases. Outra possível tarefa é reconhecer e classificar os atos de fala em um pedaço de texto.
- Tradução automática: Traduzir automaticamente texto de uma linguagem humana para outra. Este é uma das tarefas mais difíceis.
- Segmentação morfológica: Separa palavras em morfemas individuais e identifica classes de morfemas.
- Reconhecimento de entidade nomeada: Dado um fluxo de texto, determina quais são itens no mapa de texto para nomes próprios, como pessoas ou locais e qual é o tipo de cada nome (por exemplo, pessoa, local, organização).

- Geração de linguagem natural: Converte informações de base de dados de computador ou intenções semânticas em linguagem humana legível.
- Compreensão da linguagem natural: Converte pedaços de texto em representações mais formais, como estruturas de lógica de primeira ordem, que são mais fáceis de manipular pelos programas de computador.
- Marcação de classe gramatical: Dada uma frase, determina a classe gramatical de cada palavra.
- Análise sintática (Parsing): Determina a árvore de análise (análise gramatical) de uma frase.
- Respostas a perguntas: Dada uma questão de linguagem humana, determina sua resposta.
- Análise de subjetividade (sentiment analysis ou opinion mining): Extrai informações subjetivas geralmente de um conjunto de documentos, muitas vezes usando revisões online para determinar a "polaridade" sobre objetos específicos. É especialmente útil para identificar tendências da opinião pública nas mídias sociais, para fins de marketing.
- Reconhecimento de fala: Dado um som de uma pessoa ou pessoas a falar, determina a representação textual do discurso. É o oposto da síntese de fala e é uma das áreas mais difíceis.
- Desambiguação: Muitas palavras têm mais que um significado, assim temos que selecionar o significado que faz mais sentido no contexto. Para este problema, em geral é dada uma lista de palavras e sentidos de palavras associadas de um dicionário ou recurso online, como o WordNet.
- Recuperação de informação (IR): Trata-se de armazenar, pesquisar e recuperar informações.
- Extração de informação (IE): Trata-se, em geral, da extração de informação semântica a partir do texto.

2 Objetivos

Assim, esta proposta assenta numa investigação, conceptualização e respetivo desenvolvimento experimental de uma Interface Ubíqua, Interoperativa e Escalável para uma Plataforma de Serviços PLN em *Big Data*, com base no HULTIGCorpus.V1. A conceptualização e desenvolvimento desta interface será integrada no centro de investigação HULTIG - Human Language Technology Information Group, para disponibilização à comunidade cientifica.

3 Tarefas a Realizar

1. Investigação Preliminar e Especificação de Requisitos Iniciais
   - Contextualização da problemática apresentada nesta proposta
   - Especificar requisitos iniciais de infraestruturas cloud de suporte
   - Investigar e especificar requisitos iniciais da Interface
   - Especificar a arquitetura funcional dos componentes da Interface

2. Investigação, Conceptualização e Desenvolvimento Experimental da Infraestrutura Cloud de Suporte
   - Arquitetura, modelos e abordagem de armazenamento seguro de informação de suporte e de promoção da multi-linguagem
   - Interfaces aplicacionais de suporte e procedimentos associados à integração com plataformas externas e comunicação segura de informação
3. Investigação, Conceptualização e Desenvolvimento Experimental da Interface
   - Modelos de interfaces que otimizem a experiência de utilização
   - Mecanismos de gestão de serviços PLN e de interação com os mesmos
   - Dashboards de análise individualizada de histórico de uso.
   - Modelos de análise e gestão de serviços PLN
4. Integração e desenvolvimento experimental para conceito de prova.
5. Redação da dissertação de tese de mestrado

4 Contactos

Sebastião Pais (sebastiao@di.ubi.pt) - Gabinete 4.1
João Paulo Cordeiro (jpaulo@di.ubi.pt) - Gabinete 4.3

UBI, Departamento de Informática
Rua Marquês d'Ávila e Bolama
6201-001 Covilhã