UNIVERSIDADE DE COIMBRA
FACULDADE DE CIÊNCIAS E TECNOLOGIA
***Departamento de Engenharia Informática***
Pólo II da Universidade, Pinhal de Marrocos
3030 Coimbra, Portugal
*Tel.: 7000000 — Fax: 701266*

**PLANO DE ESTUDOS**

PARA INVESTIGAÇÃO CONDUCENTE À OBTENÇÃO DO GRAU DE DOUTOR
PELA FACULDADE DE CIÊNCIAS E TECNOLOGIA DA UNIVERSIDADE DE COIMBRA,
ESPECIALIDADE DE ENGENHARIA INFORMÁTICA

**Título**: Geração de uma ontologia lexical para o português: métodos e avaliação

# Introdução

A área do Processamento de Linguagem Natural (PLN) tem vindo a ganhar cada vez mais interesse na nossa sociedade de informação. O grande impulsionador nos meios empresariais é o facto de as empresas produziram enormes quantidades de informação em formato de texto. Informação essa que por não estar no formato facilmente interpretado por computadores acaba por ser subaproveitado pelos sistemas de informação hoje existentes. Daí surgir a necessidade de encontrar técnicas capazes interpretar esses documentos e colocar o conhecimento neles implícito à disposição dos utilizadores, por forma estes poderem tomar decisões informadas. Para além da perspectiva empresarial, hoje em dia encontramos vários exemplos de sistemas que lidam com a linguagem natural e que já fazem parte integrante do nosso quotidiano. Temos como exemplos os correctores ortográficos e verificadores gramaticais, assim como os motores de pesquisa que começam a procurar técnicas de PLN que os possam ajudar a compreender melhor as necessidades dos seus utilizadores.

No PLN é frequente a utilização de recursos que representam conhecimento sobre uma linguagem ou um determinado domínio de estudo. Alguns desses recursos têm a forma de uma rede semântica que relacionam os conceitos de determinada linguagem e cultura específica. Estes recursos são actualmente conhecidos como *Ontologias,* ou *Ontologias Lexicais* quando se pretende impor uma estrutura semântica ao nosso léxico. O aspecto linguístico que motiva a utilização deste tipo de recursos é a intuição de que os seres humanos estabelecem relações semânticas entre palavras e têm-nas presentes na produção e compreensão de informação expressa em linguagem natural. O presente trabalho pretende explorar esta hipótese e extrair de texto as relações semânticas que guiaram a sua geração. Assim sendo, o processo de extracção de relações semânticas pode ser encarado como um processo de engenharia inversa, ou reversa ("reverse engineering") sobre material escrito (ou falado) de forma a obter o conhecimento (uma ontologia) inicialmente utilizado na produção desse mesmo artefacto.

Baseado em investigação realizada pelo candidato no âmbito da sua tese de mestrado e também na grande tendência que se verifica na comunidade científica na utilização destes recursos, julga-se necessário a criação de uma ontologia lexical para o português. No entanto, a criação deste recurso, de uma forma (semi) automática, exigirá um estudo aprofundado da utilização da língua portuguesa e a concepção de estratégias computacionais e linguísticas que possam atingir os objectivos.

# Objectivos

Os objectivos deste trabalho podem ser divididos segundo dois eixos ortogonais mas complementares:

1. Definição e implementação de metodologias de extracção (semi) automática de Ontologias a partir de texto.
2. Definição e implementação de metodologias de avaliação dessas mesmas Ontologias.

O primeiro ponto refere-se ao estudo e implementação de técnicas computacionais capazes de extrair relações semânticas entre conceitos referenciados em texto. A satisfação deste objectivo produzirá um conjunto de metodologias computacionais considerado eficaz, tendo em conta o estado da arte na área, e que deverão servir de suporte à génese de uma ***Ontologia Lexical do Português***. Algumas metas intermediárias passam pelo estudo e construção de Ontologias de dimensão mais reduzida e especializada a um determinado domínio. A restrição do domínio deverá facilitar o estudo exploratório inicial e permitir estabelecer as bases necessárias sobre as quais o restante trabalho poderá assentar.

O segundo ponto permite validar o trabalho feito no primeiro. É evidente que a avaliação do trabalho feito é essencial para que este possa ser considerado credível. No entanto, ainda não existe uma metodologia de avaliação consensual no seio da comunidade científica referente à extracção automática de ontologias a partir de texto. Por este facto julgamos que este objectivo deve ser explicitado de forma a salientar a sua verdadeira importância.

# Programa de Trabalhos

Ponto essencial no qual a nossa investigação pretende incidir é na definição das técnicas computacionais de extracção do conhecimento para a construção da ontologia lexical. Pretendemos utilizar várias fontes de informação (textuais) tais como:

1. Dicionários – contendo informação semântica explícita.
2. Diários ("logs") de motores de pesquisa – permitem estudar as relações que os utilizadores estabelecem entre palavras quando procuraram informação.
3. Outros textos que sejam propícios à extracção automática de relações semânticas.

O recurso resultante desta investigação consiste num conjunto abrangente de relações semânticas entre itens do vocabulário da língua portuguesa, tais como hiperonímia, hiponímia, meronímia assim como outras que se considerem ser relevante extrair e disseminar.

O trabalho a realizar pode ser dividido em três fases:

### *1ª Fase*

Esta fase consistirá principalmente na pesquisa, recolha e análise crítica da bibliografia existente relevante à questão da extracção de ontologias a partir de textos. No término desta fase deverá ser produzido um documento crítico sintetizando os trabalhos e recursos estudados.

### *2ª Fase*

Esta fase deverá dar início à implementação do sistema de extracção. A metodologia escolhida para a implementação baseia-se em várias iterações das seguintes etapas:

- Análise
- Implementação
- Avaliação
- Documentação

Em que a etapa de análise tentará encontrar soluções para os problemas identificados durante a etapa de avaliação da iteração anterior, propondo assim novos algoritmos e/ou novos recursos capazes de anular as deficiências identificadas.

O primeiro problema a abordar, que dará início à primeira iteração, será a identificação das relações a extrair.

Após a apreciação das relações e para que este modelo iterativo seja exequível é imprescindível a definição de um conjunto de métricas e critérios que possam ser utilizadas durante a avaliação de modo a medir-se o impacto de cada alteração introduzida. Assim, o início desta fase deverá servir para estabelecer as métricas e critérios a utilizarem.

Prevê-se que algumas das iterações poderão incidir sobre um domínio mais restrito, em vez de considerar todo o vocabulário português, o que permitirá um maior controlo sobre as avaliações efectuadas.

Durante a realização desta fase e caso se venham a descobrir novas técnicas e/ou obterem-se resultados favoráveis, julga-se importante a escrita de artigos científicos.

### *3ª Fase*

A última fase destina-se à produção da documentação, nomeadamente a dissertação de doutoramento a ser submetida.

## Calendarização

O presente plano de estudos assume a duração total de três anos. Assim sendo julga-se que a primeira fase não deverá ocupar mais do que os primeiros 4 meses. No entanto será sempre necessário estar atento aos desenvolvimentos no estado da arte e actualizar adequadamente a documentação produzida nesta fase.

A segunda fase ocupará os restantes 8 meses do primeiro ano, assim como todo o segundo ano. É nesta fase que quase todo o trabalho de investigação e desenvolvimento será realizado.

Finalmente o terceiro ano corresponderá à terceira fase constituída essencialmente pela escrita da dissertação.

A escrita de relatórios de progresso e artigos científicos será feita em paralelo e sempre que se justificar.

## Trabalhos Relacionados

Talvez o recurso lexical mais conhecido e que se enquadra no contexto do trabalho que se pretende realizar é o WordNet [1]. O WordNet é uma ontologia lexical construída manualmente por lexicógrafos e é baseado em fundamentos psico-linguísticos. Um aspecto bastante desenvolvido deste recurso é a sua taxonomia, isto é, a sua hierarquia de categorização semântica baseada na relação de hiperonímia/hiponímia. Apesar de muitos autores só reconhecerem a componente taxonómica do WordNet, este recurso fornece bastante mais conhecimento, e tem sido amplamente utilizado em PLN como sugere o número de contribuições existentes no sítio do recurso assim como os artigos publicados[1] que o referenciam.

Outro recurso frequentemente citado é o Cyc [2]. Embora não tenha sido destinado a ser um recurso linguístico é frequentemente citado e utilizado na comunidade de PLN. Ao contrário do WordNet, o Cyc é altamente formalizado e todo o conhecimento é descrito utilizando uma linguagem baseada na lógica de predicados de primeira ordem. O Cyc é provavelmente um dos recursos mais ricos em conhecimento, sendo também um dos primeiros projectos na área de representação de conhecimento. Infelizmente, o Cyc é um recurso baseado no Inglês o que impede a sua utilização pelos investigadores interessados na língua portuguesa, para além de ser um recurso

[1]Uma lista de publicações até ao final de 2004 está acessível em http://engr.smu.edu/~rada/wnb/.

em que a utilização só é possível mediante a compra do mesmo.

Outro recurso semelhante ao WordNet mas tendo como língua alvo o Chinês é o HowNet [3]. O HowNet para além de servir de ontologia lexical permite ainda estabelecer mapeamentos entre conceitos existentes no inglês e no chinês.

No que diz respeito ao português, têm existido algumas iniciativas com o mesmo objectivo, das quais destacamos:

- WordNet.PT (http://www.instituto-camoes.pt/bases/lingua/wordnet.htm)
- WordNet.BR (http://www.nilc.icmc.usp.br/nilc/projects/wordnetbr.htm)

No entanto todas as tentativas de obtenção destes recursos têm sido em vão, indicando que o seu desenvolvimento está parado ou que os recursos não são de domínio público nem provavelmente virão a ser.

Todos os recursos anteriormente referidos partilham da mesma característica: são construídos manualmente. Com o avanço da tecnologia e consequente aumento do poder computacional, hoje em dia é possível a utilização de técnicas de processamento de texto que facilitam a extracção automática de relações e conceitos.

Dos trabalhos que seguiram uma abordagem automática ou semi-automática destacamos o trabalho desenvolvido pela Marti Hearst no início de década de 90 (para mais detalhes ver [4]). Ela definiu um conjunto de padrões léxico-sintácticos que utilizou para procurar e extrair relações de hiperonímia/hiponímia de texto. O trabalho realizado, apesar de muito pouco abrangente pois só estudou a relação de hiperonímia/hiponímia, provou ser motivador e digno de mais empenho. Desde do trabalho inicial de Hearst têm surgido vários projectos de investigação com o objectivo de extraírem relações automaticamente de texto e que de alguma forma estendem o trabalho de Hearst.

Outro trabalho relevante é o realizado por Stephen Richardson no âmbito da sua dissertação de doutoramento [5]. Neste trabalho o autor extrai relações semânticas de um dicionário (em formato digital) utilizando técnicas de PLN. Estas relações são posteriormente aglomeradas de forma a gerar uma ontologia lexical[2] que mais tarde veio ser conhecida como MindNet [6].

## Importância dos Resultados e Aspectos Inovadores

Todos os trabalhos disponíveis e enumerados na secção anterior estão relacionados com a língua inglesa, provocando uma lacuna evidente no que diz respeito ao português. O presente trabalho visa preencher essa falta, fornecendo um recurso utilizável por outros investigadores, assim como metodologias de extracção de conhecimento a partir de texto que poderão ser reutilizadas. As técnicas de avaliação também serão um contributo válido e original, visto que esta questão ainda está pouco elaborada no seio da comunidade científica interessada na extracção automática de conhecimento a partir de texto.

## Experiência Anterior

O candidato já detém experiência na área do PLN e na utilização de ontologias semelhantes à que se pretende construir. Ainda no âmbito da licenciatura, colaborou no projecto ReBuilder [7], onde adquiriu conhecimentos em PLN, Raciocínio Baseado em Casos e na utilização/exploração do WordNet. A componente mais relevante do seu trabalho, para o presente plano, foi a concepção de um módulo de interpretação de especificações de software descrito em linguagem natural [8].

No âmbito do seu trabalho de mestrado continuou a trabalhar com o WordNet e adquiriu alguma experiência na área de recolha de informação, mais especificamente na componente da expansão dos termos de pesquisa (“query expansion”). O trabalho realizado passou ainda pela concepção e implementação de métricas computacionais de cálculo de semelhanças semânticas entre palavras [9], trabalho este que já se encontra a ser aplicado (em protótipos de investigação) em motores de

[2]O autor prefere a utilização do termo "Base de Conhecimento Lexical" traduzido do inglês *Lexical Knowledge Base*.

pesquisa (ver [10] para mais detalhes). Como os modelos computacionais de cálculo de semelhanças assumem a existência de uma ontologia lexical (mais especificamente o WordNet), o candidato também analisou métodos que permitissem a inclusão de novos conceitos na ontologia permitindo preencher lacunas semânticas e aumentando assim a sua utilidade e abrangência.

Actualmente encontra-se como bolseiro de investigação a trabalhar na Linguateca [11], um centro de recursos distribuído para o processamento computacional da língua portuguesa. As tarefas do candidato passaram pelo processamento de um corpus de texto (WPT03) contendo os documentos pertencentes à web portuguesa, resultando numa lista de frequências de todos os termos existentes no corpus, além de ter estudado e implementado estratégias de detecção de documentos duplicados. (A existência de documentos duplicados em colecções obtidas através de "web crawling" é um problema recorrente que os motores de pesquisa têm de considerar.) Ainda no âmbito desta bolsa participou na implementação e concepção do sistema de avaliação automático utilizado na avaliação conjunta de reconhecimento de entidades mencionadas em português promovido pela Linguateca [12].

## Referências


[1] Miller, G., Beckwith, R., Fellbaum, C., Gross, D., & Miller, K. J. (1990). Introduction to WordNet: an on-line lexical database. In International Journal of Lexicography, 235 – 244.

[2] Lenat, D. B. & Guha, R. V. (1990). Building Large Knowledge-Based Systems: Representation and Inference in the CYC Project. Addison-Wesley, Reading, Massachusetts.

[3] Dong, Z. & Dong, Q. (2004). HowNet. http: //www.keenage.com/.

[4] Marti Hearst. (1992) Automatic Acquisition of Hyponyms from Large Text Corpora. In Proceedings of the Fourteenth International Conference on Computational Linguistics.

[5] Stephen Richardson. (1997). Determining Similarity and Inferring Relations in a Lexical Knowledge Base. PhD thesis, City University of New York, Department of Computer Science.

[6] Stephen Richardson et al. (1998). MindNet: acquiring and structuring semantic information from text. In Proceedings of the 17th International Conference on Computational Linguistics.

[7] Paulo Gomes, Francisco C. Pereira, Paulo Paiva, Nuno Seco, Paulo Carreiro, José L. Ferreira, Carlos Bento. (2004). REBUILDER: A Case-Based Reasoning Design Reuse Tool. In journal of Engineering Intelligent Systems.

[8] Nuno Seco, Paulo Gomes, Francisco C. Pereira. (2004). Using CBR for Semantic Analysis of Software Specifications. In Proceedings of the 7th European Conference on Case-Based Reasoning, 778-792.

[9] Nuno Seco. (2005). Computational Models of Semantic Similarity in Lexical Ontologies. MSc. thesis, University College Dublin, Department of Computer Science.

[10] Giannis Varelas et al. (2005). Semantic Similarity Methods in WordNet and their Application to Information Retrieval on the Web. To appear in Proceedings of the 7th ACM International Workshop on Web Information and Data Management.

[11] Diana Santos et al. (2004). Linguateca: um centro de recursos distribuído para o processamento computacional da língua portuguesa. In Proceedings of the International Workshop "Taller de Herramientas y Recursos Linguísticos para el Espanõl y el Portugués", IX Iberoamerican Conference on Artificial Intelligence, 147-154.

[12] Nuno Seco, Diana Santos, Nuno Cardoso & Rui Vilela. (Submitted). HAREM: An Advanced NER Evaluation Contest for Portuguese.

Coimbra, 30 de Outubro de 2005.

| O candidato, | O orientador, | A orientadora, |
| --- | --- | --- |
| _____ | _____ | _____ |
| (Nuno Seco) | (Paulo Gomes) | (Diana Santos) |