

## SUPeRB

### Sistema Uniformizado de Pesquisa de Referências Bibliográficas

---


Luís Miguel Cabral

28 de Maio de 2007

Orientado Por:  
Professor Doutor Eugénio Oliveira (FEUP)  
Doutora Diana Santos (SINTEF)




## Índice da apresentação



- **Introdução**
- **Motivação**
- **Contribuições**
- **Especificação do SUPeRB**
  - Fontes
  - Arquitectura
  - Tarefas do SUPeRB
- **Métodos de avaliação**
- **Linhas de investigação futura**

2

## O que é o SUPeRB?




Sistema Uniformizado de Pesquisa de Referências Bibliográficas

- É um assistente automático na pesquisa de informação bibliográfica
  - Permite a pesquisa e processamento de informação bibliográfica
  - Permite uma melhor interacção por parte de vários tipos de utilizadores
    - Utilizadores que apenas pesquisam
    - Utilizadores que contribuem
    - Gestor de um recurso

3

## Informação bibliográfica (1)




- **Referência bibliográfica**  
 Texto em linguagem *quase* natural ou num formato específico que permite identificar um documento

[Marrafa & Ribeiro 2001]  
 Palmira Marrafa & António Ribeiro. "Quantitative Evaluation of Machine Translation Systems: Sentence Level". In *Proceedings of the MT Summit VIII: Fourth ISLE workshop* (Santiago de Compostela, 22 de Setembro de 2001), pp. 39-43. <http://www.eamt.org/summitviii/papers/marrafa.pdf>

4

## Informação bibliográfica (2)



- Elemento bibliográfico
 


São partes que compõem a referência bibliográfica e que identificam propriedades do documento tal como o autor, o título, a conferência, data, local, etc.

[Marrafa & Ribeiro 2001]
[Palmira Marrafa & António Ribeiro] [Quantitative Evaluation of Machine Translation Systems: Sentence Level] [in <i>Proceedings of the MT Summit VIII: Fourth ISLE workshop</i> ] [Santiago de Compostela] [22 de Setembro de 2001].
[pp. 39-43.] [ <a href="http://www.eamt.org/summitVIII/papers/marrafa.pdf">http://www.eamt.org/summitVIII/papers/marrafa.pdf</a> ]

□ Elementos bibliográficos

5

## Informação bibliográfica (3)



- Formato bibliográfico
 


Texto num formato estruturado, onde cada elemento bibliográfico está devidamente identificado e delimitado.

  - EndNote
  - RIS
  - BibTeX

```
@inproceedings{1141121239,
author =(Palmira Marrafa and António Ribeiro),
title =(Quantitative Evaluation of Machine Translation Systems: Sentence Level),
year =(2001),
booktitle =(Proceedings of the MT Summit VIII: Fourth ISLE workshop),
pages =(39-43),
location =(Santiago de Compostela),
url ={"url["http://www.eamt.org/summitVIII/papers/marrafa.pdf"]}
```

6

## Fontes da informação




- Fontes estruturadas
 

Acesso a repositórios que contêm informação bibliográfica devidamente processada e à qual é possível aceder aos dados devidamente identificados, através de protocolos próprios
- Fontes genéricas
 

Páginas html e documentos (doc, ps, pdf, txt, ppt)

7

## Índice da apresentação



- Introdução
- **Motivação**
- Contribuições
- Especificação do SUPeRB
  - Fontes
  - Arquitectura
  - Tarefas do SUPeRB
- Métodos de avaliação
- Linhas de investigação futura

8

## Motivação para o desenvolvimento do SUPeRB



- Facilitar o acesso a referências e a documentação científica
- A informação encontra-se dispersa na Web, em bases de dados bibliográficas e repositórios, páginas Web e documentos na Web.
- Motivar a participação de utilizadores
  - Permitindo a fácil introdução de informação bibliográfica

9

## Catálogo de publicações da Linguateca



- A **Linguateca** é um centro de recursos - distribuído - para o processamento computacional da língua portuguesa
- O catálogo de publicações da Linguateca, é um dos recursos que visa disponibilizar informação bibliográfica no âmbito do processamento computacional da língua portuguesa
  - Cerca de 1344 publicações (cerca de 280 publicações da Linguateca), 600 das quais em português.
  - Catálogo possui muitos processos manuais para confirmação e inserção de informação

10

## Objectivos da tese



- Criar um sistema de pesquisa e processamento de informação bibliográfica
- Capaz de executar determinadas tarefas vitais no processamento de referências bibliográficas
  - Colmatar as limitações do processamento automático com supervisão humana
  - Melhorar o catálogo de publicações da Linguateca

11


## Problemas tratados (1)



- Ao procurar referências bibliográficas por toda a Web coloca-se o problema de “**Como pesquisar na Web?**”
- Ao obter dados da Web, é necessário **filtrar** a informação relevante
- **Extrair** a informação dos **vários formatos** distintos e converter essa informação para um formato legível

12


### Problemas tratados (2)



- É necessário **processar** e extrair a informação relevante (bibliográfica)
- É necessário **correlacionar a informação** obtida
- É necessário **validar** os dados antes de os tornar públicos
- **Integrar** a informação recolhida num recurso publico

13


### Índice da apresentação



- Introdução
- Motivação
- **Contribuições**
- Especificação do SUPeRB
  - Fontes
  - Arquitectura
  - Tarefas do SUPeRB
- Métodos de avaliação
- Linhas de investigação futura

14


### Contribuições da tese



- Propôs-se uma arquitectura que permite ultrapassar as barreiras previstas, que integre métodos e tecnologias existentes
- Desenvolveu-se a arquitectura, aplicando-a ao catálogo da Liguateca
- Apresentou-se um método de avaliação para módulos específicos do sistema

15


### Índice da apresentação



- Introdução
- Motivação
- Contribuições
- **Especificação do SUPeRB**
  - **Fontes**
  - Arquitectura
  - Tarefas do SUPeRB
- Métodos de avaliação
- Linhas de investigação futura

16


## Fontes estruturadas



- Repositórios e bases de dados bibliográficas online
  - DBLP
  - Citeseer
  - Editoras (Springer, IEEE )
- Motores de pesquisa bibliográficos
  - Google scholar
  - Microsoft Live Academic
- Gestores bibliográficos online
  - Citeulike
  - Bibsonomy

17

## Fontes genéricas




Páginas html ou documentos que se encontrem na Web.

- Páginas de conferências
- Páginas de instituições
- Páginas pessoais (investigadores)
- Páginas Web de repositórios

18

## O que se obtém das fontes



- A informação que se encontra disponível pode não ser exacta
  - Pode ser inconsistente com os dados inseridos pelo utilizador
  - Pode conter informação irrelevante
  - Pode conter informação errada
  - Pode omitir determinada informação
  - Contém redundâncias

19

## Vantagens/Desvantagens




Repositórios e bases de dados bibliográficas online

- Vantagens
  - **Alta fiabilidade**
  - **Processo de pesquisa e processamento automáticos**
  - Facilitar meios de conversão para formatos exportáveis
  - Interação com outros sistemas
- Desvantagens
  - Limitados a áreas / eventos científicos específicos
  - Sistemas automáticos podem gerar erros ou produzir informação escassa
  - Não permitem introdução de dados

20

## Vantagens/Desvantagens (2)



Motores de pesquisa bibliográficos (Google Scholar, Microsoft Live Academic)

- Vantagens
  - Processo de pesquisa e processamento automáticos
  - Facilitar meios de conversão para formatos exportáveis
  - **Interação com outros sistemas e com a Web em geral**
- Desvantagens
  - Sistemas automáticos podem gerar erros ou produzir informação escassa
  - Não permitem introdução de dados
  - **Apenas interfaces para utilizadores**

21

## Vantagens/Desvantagens (3)




Gestores bibliográficos (CiteULike, Bibsonomy)

- Vantagens
  - Facilitar meios de conversão para formatos exportáveis
  - **Permite a introdução de informação bibliográfica**
  - **Maior supervisão humana**
  - **Partilha de informação entre utilizadores**
  - **Permite a classificação de utilizadores**
- Desvantagens
  - Não permite a pesquisa fora do sistema
  - A introdução de informação não possui meios de verificação
  - Não oferece garantias de fiabilidade

22

## Vantagens/Desvantagens (4)




Fontes genéricas (Através de motores de pesquisa genéricos)

- Vantagens
  - **Um conjunto mais extenso de informação**
  - Informação em primeira mão
- Desvantagens
  - Baixa fiabilidade
  - Apenas de leitura
  - Informação apresentada em linguagem quase natural o que leva à necessidade de métodos adicionais para extrair a informação relevante

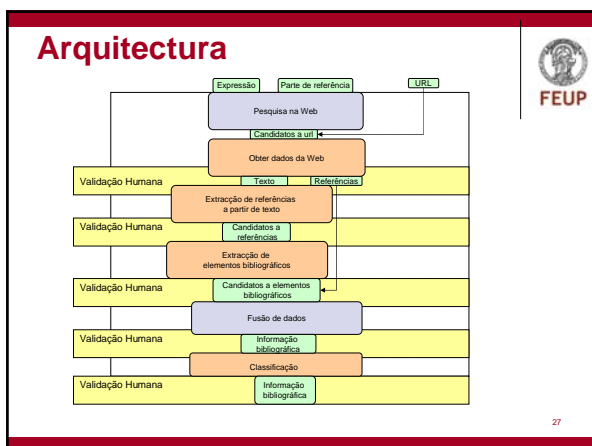
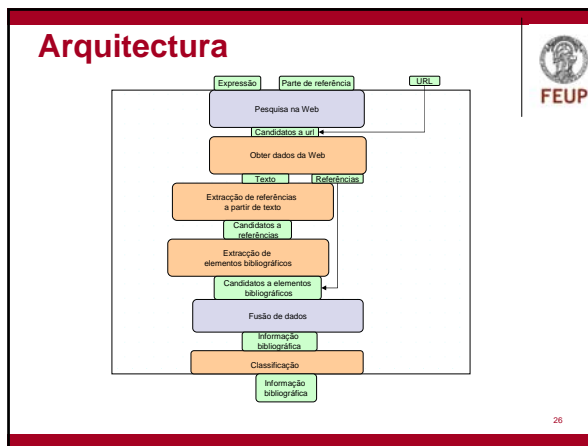
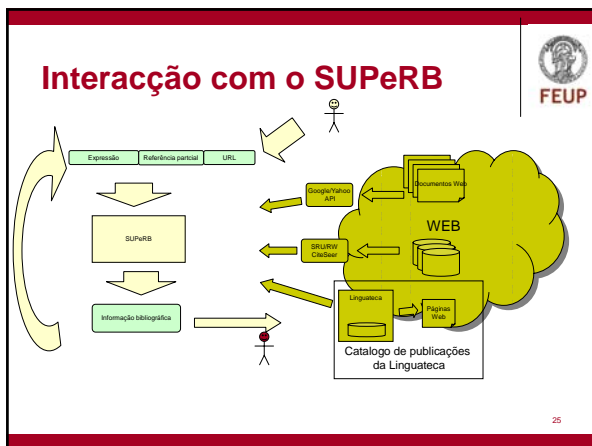
23

## Índice da apresentação



- Introdução
- Motivação
- Contribuições
- **Especificação do SUPeRB**
  - Fontes
  - **Arquitectura**
  - Tarefas do SUPeRB
- Métodos de avaliação
- Linhas de investigação futura


24



- ### Índice da apresentação
- Introdução
  - Motivação
  - Contribuições
  - **Especificação do SUPeRB**
    - Fontes
    - Arquitectura
    - **Tarefas do SUPeRB**
  - Métodos de avaliação
  - Linhas de investigação futura

Pesquisa na Web

## Pesquisa nas fontes genéricas




- **Dada:**
  - Uma expressão específica introduzida pelo utilizador
  - Ou parte de uma referência bibliográfica
- **Obter um conjunto de documentos da Web com informação relevante**
- **Método:**
  - Acede a motores de pesquisa, através de API próprias, para obter uma lista de candidatos
- **Refinamento da pesquisa**
  - Adicionando palavras de carácter específico aos padrões utilizados

29

Pesquisa na Web

## Pesquisa nas fontes genéricas



publicações  
publications  
...  
bibliografia  
bibliography

Informação do utilizador Palavras específicas

Marrafa quantitative evaluation...

Expressão de pesquisa  
Marrafa quantitative evaluation article


API 1  
API 2

URL candidato

30

Obter dados da Web

## Obter dados da Web: Análise de resultados




- **Filtragem e tratamento de URL**
  - Ignorar resultados
  - Obtenção da informação de repositórios bibliográficos
    - API Citeseer
    - Protocolo SRU/SRW
  - Extração do texto de páginas e de documentos (HTML, DOC, PDF, PS, PPT, TXT) => Texto

31

Obter dados da Web

## Análise de resultados



URL candidato

Documento

HTML  
Word  
PDF  
PS  
PPT

Documento


Texto

Informação bibliográfica

32

Extracção de referências a partir de texto

## Extracção de candidatos a referências bibliográficas a partir de texto




- Análise da estrutura e formato do documento  
Tentar identificar o tipo de documento (Lista de publicações? Académico? Apresentação?)
- Identificar as zonas do texto com informação bibliográfica e extrair potencial informação relevante recorrendo a heurísticas e à análise obtida
  - Auto-referência (referência do próprio documento)
  - Blocos de referências (Listas)

33

Extracção de referências a partir de texto

## Extracção de auto-referência: Exemplo



MALINCHE: A NER system for Portuguese that reuses knowledge from Spanish

Thamar Solorio<sup>\*</sup>  
University of Texas at El Paso  
Department of Computer Science  
April 19, 2007

**Abstract**  
Named Entity extraction is a problem that has been tackled in the NLP research community for some time now. First the hand-coded approaches, then the machine learning ones. Currently what we are looking for are methods that can be easily adapted to different domains and/or languages. In a previous work we showed how a hand-coded tagger can still be useful for NE extraction in different domains by using the output of the tagger as attributes in a machine-learning task. In this work we show how the same representation of the learning task, using the same hand-coded tagger for Spanish, can be also exploited to perform NE extraction in Portuguese.


### 1 Introduction

Due to the many potential uses of Named Entities (NEs) in higher-level NLP tasks, a lot of work has been devoted to developing accurate NE recognizers. Earlier approaches were primarily based on hand-coded knowledge, lists of gazetteers, and trigger words (e.g. [1, 18, 5, 29]). More recently, as machine learning has increased its

34

Extracção de referências a partir de texto

## Extracção de auto-referência: Exemplo



MALINCHE: A NER system for Portuguese that reuses knowledge from Spanish  
Thamar Solorio  
University of Texas at El Paso  
Department of Computer Science  
April 19, 2007


**Abstract**  
Named Entity extraction is a problem that has been tackled in the NLP research community for some time now. First the hand-coded approaches, then the machine learning ones. Currently what we are looking for are methods that

**Título:** MALINCHE: A NER system for Portuguese that reuses knowledge from Spanish  
**Autor:** Thamar Solorio  
**Instituição:** University of Texas at El Paso Department of Computer Science  
**Data:** April 19, 2007  
**Resumo:** Named Entity extraction is a problem that has been tackled in the NLP(...)

35

Extracção de referências a partir de texto

## Extracção de blocos de referências: Exemplo



regardless of the ones determined by the hand-coded tagger.


**References**

- [1] Douglas E. Appelt, Jerry B. Hahn, John Bear, David Leavel, Miguel Kuznetsov, Andy Kehler, David Martin, Kevin Moran, and Harby Tsou. SHE International BAVES system: MUC-6 test results and analysis. In *Proceedings of the Fifth Message Understanding Conference (MUC-6)*, pages 237-246. Columbia, MD, November 1995. NIST, Morgan Kaufmann Publishers.
- [2] Beatriz Arribas, Xavier Carreras, Lluís Màrquez, Toni Martí, Lluís Padró, and Maria José Simón. A proposal for wide-coverage Spanish named entity recognition. *Second Spanish symposium on Computational Linguistics* (Semin-02-02), Mar 2002.
- [3] Donald M. Bikel, Scott Miller, Richard Schwartz, and Ralph Weischedel. Nymble: a high performance learning named-finder. In *Proceedings of the Fifth Conference on Applied Natural Language Processing*, pages 191-201. Washington, D. C., 01 March 1 April 1997. Association for Computational Linguistics.
- [4] Donald M. Bikel, Richard Schwartz, and Ralph Weischedel. An algorithm that knows what's in a name. *Machine Learning, Special Issue on Natural Language Learning*, 3(1):211-221, February 1999.
- [5] William J. Blake, Fabio Bazzoli, and David Moscard. FUDGE: Description of the NE system used for MUC-7. In *Proceedings of the 5th Message Understanding Conference, MUC-7*, Fairfax, VA, April-Mar 1998.
- [6] Andrew Borthwick. *A Maximum Entropy Approach to Named Entity Recognition*. PhD thesis, New York University, New York, September 1999.
- [7] Xavier Carreras, Lluís Màrquez, and Lluís Padró. Named entity extraction using adaboost. In *In: Balth and April and the Book editors, Proceedings of the Sixth Conference on Computational Natural Language Learning, CoNLL-06*, pages 167-176. Taipei, Taiwan, August 11 September 1 2002.
- [8] Xavier Carreras, Lluís Màrquez, and Lluís Padró. Named entity recognition for Catalan using only Spanish re-

36

Extracção de referências a partir de texto

## Extracção de blocos de referências: Exemplo (2)




FEUP

37

Extracção de referências a partir de texto

## Extracção de blocos de referências: Exemplo (2)



FEUP

38

Extracção de elementos bibliográficos

## Extracção de elementos bibliográficos

- Extrair títulos, autores, datas, locais a partir de candidatos a referências relevantes
  - Atomização
  - Classificação dos átomos através do Repositório de Elementos Bibliográficos (REB: Pessoa, local, conferências, editoras.)
  - Identificação de padrões
    - ParaTools (Mike Jewell)

FEUP

39

## bibliográficos: Exemplo

Steve Lawrence, C. Lee Giles and Kurt D. Bollacker, "Autonomous Citation Matching." Proceedings of the Third International Conference on Autonomous Agents, Seattle, Washington, May 1-5, ACM Press, New York, NY, 1999.


Steve Lawrence, C. Lee Giles and Kurt D. Bollacker, "Autonomous Citation Matching." Proceedings of the Third International Conference on Autonomous Agents, Seattle, Washington, May 1-5, ACM Press, New York NY, 1999.

NOME, T, NOME, NOME, NOME, "TITULO", CONFERENCIA, LOCAL, LOCAL, DATA, EDITORA, LOCAL, LOCAL, DATA, NOME, NOME and NOME, "TITULO", CONFERENCIA, LOCAL, DATA, EDITORA, LOCAL, DATA.

**Autor:** Steve Lawrence  
**Autor:** C. Lee Giles  
**Autor:** Kurt D. Bollacker  
**Título:** Autonomous Citation Matching  
**Conferência:** Proceedings of the Third International Conference on Autonomous Agents  
**Local:** Seattle, Washington  
**Data:** May 1-5  
**Editora:** ACM Press  
**Address:** New York, NY  
**Ano:** 1999

FEUP


40

Fusão de dados	
----------------	---

## Fusão dos dados

- A fusão de dados pode ser necessária
  - Por terem sido obtidos mais do que um resultados de uma pesquisa na Web
  - Por ter sido identificado informação bibliográfica semelhante num repositório local
    - Uma referência incompleta
    - Título de uma conferência ou livro
  - É necessário então desambiguar elementos bibliográficos

41

Fusão de dados	
----------------	---

## Fusão dos dados: Exemplo

Luis Sarmiento "Hunting Answers with RAPOSA (FOX)", in Alessandro Nardi, Carol Peters & José Luis Vicedo (eds.), Cross Language Evaluation Forum: Working Notes for the CLEF 2006 Workshop (CLEF 2006) (Alcántara, Espanha, [http://www.linguistica.pt/Repository/Sarmiento\\_RAPOSA.pdf](http://www.linguistica.pt/Repository/Sarmiento_RAPOSA.pdf))


+

Hunting Answers with RAPOSA (FOX), Luis Sarmiento Working Notes of the Cross-Language Evaluation Forum Workshop (CLEF 2006) Alcántara, Spain, 20-22 de September 2006. [http://paginas.fe.up.pt/~lisa/conteudo/pub/pw/CLEF\\_2006/Sarmiento\\_RAPOSA.pdf](http://paginas.fe.up.pt/~lisa/conteudo/pub/pw/CLEF_2006/Sarmiento_RAPOSA.pdf)

=

(Title Hunting Answers with RAPOSA (FOX))  
 (Autor Luis Sarmiento)  
 (Conference Working Notes of the Cross-Language Evaluation Forum Workshop (CLEF 2006))  
 (Editores Editores: Alessandro Nardi, C.P. & Vicedo, J.L.)  
 (Local Alcántara, Spain)  
 (Date 20-22 de September 2006)  
 (URL [http://paginas.fe.up.pt/~lisa/conteudo/pub/pw/CLEF\\_2006/Sarmiento\\_RAPOSA.pdf](http://paginas.fe.up.pt/~lisa/conteudo/pub/pw/CLEF_2006/Sarmiento_RAPOSA.pdf))  
 (URL [http://www.linguistica.pt/Repository/Sarmiento\\_RAPOSA.pdf](http://www.linguistica.pt/Repository/Sarmiento_RAPOSA.pdf))


42

Classificação	
---------------	---


## Classificação

- Marcação manual livre pelo utilizador (*tagging*)
- Classificação interna, com base em atributos específicos do documento (conferências específicas ou tipos de documentos tal como apresentações, ou língua)

43

Classificação	
---------------	---

## Interface de classificação



44

## Índice da apresentação



- Introdução
- Motivação
- Contribuições
- Especificação do SUPeRB
  - Arquitectura
  - Tarefas do SUPeRB
- **Métodos de avaliação**
- Linhas de investigação futura

45

## Avaliação de extracção de referências




- Textos analisados manualmente
- Listas de referências candidatas
- Categorização nominal de cada candidato e da info. conhecida
  - certo
  - com informação excedentária
  - informação incompleta
  - falta
  - erro
- Medidas de RI: precisão, abrangência, sobregeração etc.

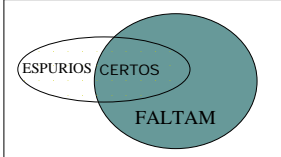


46

## Avaliação de extracção de elementos




- Referências analisadas manualmente
- Lista de elementos candidatos para cada referência: de notar que em alguns campos há mais do que um elemento
- Categorização nominal por campo ou por elemento
  - correcto
  - excedentário
  - incompleto
  - em falta
  - espúrio
- Precisão por referência
- Precisão por autor



47


## Índice da apresentação



- Introdução
- Motivação
- Contribuições
- Especificação do SUPeRB
  - Arquitectura
  - Tarefas do SUPeRB
- Métodos de avaliação
- **Linhas de investigação futura**

48

## Linhas de investigação futura



- Melhorar o conteúdo
  - Classificação automática
  - Análise de co-citações
- Aumentar as funcionalidades
  - Automatização de tarefas para permitir a actualização periódica autónomas (Sempre validadas pelo gestor)
  - Facilitar o conteúdo do recurso via serviços Web
- Usabilidade
  - Usabilidade da interface
  - Perguntas em linguagem natural

49

## SUPeRB

### Sistema Uniformizado de Pesquisa de Referências Bibliográficas

---

Luís Miguel Cabral

28 de Maio de 2007



Orientado Por:  
Professor Doutor Eugénio Oliveira (FEUP)  
Doutora Diana Santos (SINTEF)