Reconhecimento de entidades mencionadas em português

Documentação e actas do HAREM,
a primeira avaliação conjunta na área

Diana Santos e Nuno Cardoso
editores

Linguateca, 2007

Reconhecimento de entidades mencionadas em português

Documentação e actas do HAREM,
a primeira avaliação conjunta na área

Diana Santos e Nuno Cardoso
editores

Linguateca, 2007

1ªEdição, Novembro de 2007.
1st Edition, November 2007.

Publicação Digital. *Digital Print.*

ISBN 978-989-20-0731-1

O capítulo 12, "Functional Aspects of Portuguese NER", foi anteriormente publicado em Renata Vieira, Paulo Quaresma, Maria da Graça Volpes Nunes, Nuno J. Mamede, Cláudia Oliveira e Maria Carmelita Dias (eds.), *Computational Processing of the Portuguese Language: 7th International Workshop, PROPOR 2006, Itatiaia, Brazil, May 2006. Proceedings.* p. 80-89, na série LNAI, Vol. 3960 da editora Springer Verlag, ISBN-10: 3-540-34045-9.
The chapter 12, "Functional Aspects of Portuguese NER", was republished from Renata Vieira, Paulo Quaresma, Maria da Graça Volpes Nunes, Nuno J. Mamede, Cláudia Oliveira & Maria Carmelita Dias (eds.), Computational Processing of the Portuguese Language: 7th International Workshop, PROPOR 2006, Itatiaia, Brazil, May 2006. Proceedings. pp. 80-89, Springer Verlag, LNAI series, Vol. 3960, ISBN-10: 3-540-34045-9.

O capítulo 16, "Directivas para a identificação e classificação semântica na colecção dourada do HAREM", foi previamente publicado como Relatório Técnico DI/FCUL TR–06–18, Departamento de Informática, Faculdade de Ciências da Universidade de Lisboa.
The chapter 16, "Directivas para a identificação e classificação semântica na colecção dourada do HAREM*", was previously published as Technical Report DI/FCUL TR–06–18, Department of Informatics, Faculty of Sciences, University of Lisbon.*

O texto do capítulo 17, "Directivas para a identificação e classificação morfológica na colecção dourada do HAREM", foi previamente publicado como Relatório Técnico DI/FCUL TR–06–19, Departamento de Informática, Faculdade de Ciências da Universidade de Lisboa.
The chapter 17, "Directivas para a identificação e classificação morfológica na colecção dourada do HAREM*", was previously published as Technical Report DI/FCUL TR–06–19, Department of Informatics, Faculty of Sciences, University of Lisbon.*

O capítulo 18, "Avaliação no HAREM: Métodos e medidas", foi previamente publicado como Relatório Técnico DI/FCUL TR–06–17, Departamento de Informática, Faculdade de Ciências da Universidade de Lisboa.
The chapter 18, "Avaliação no HAREM*: Métodos e medidas", was previously published as Technical Report DI/FCUL TR–06–17, Department of Informatics, Faculty of Sciences, University of Lisbon.*

Prefácio

Não quisemos que a divulgação do trabalho feito no HAREM sofresse um atraso tão significativo como o que ocorreu por ocasião das Morfolimpíadas (cujo livro saíu à luz quatro anos depois). Por isso, decidimos publicar a presente obra de forma electrónica e gratuita, de forma a maximizar o seu alcance e minimizar o tempo de saída.

Isso não obstou, naturalmente, a que tivéssemos seguido um processo editorial rigoroso, com revisão cruzada entre os autores, além de amplos comentários e sugestões pelos dois editores, numa tentativa de tornar os capítulos mais homogéneos entre si, e ainda a leitura crítica da primeira versão completa do livro por vários especialistas em processamento computacional do português, que resultou em várias sugestões valiosas e observações pertinentes.

Para que conste, aqui fica a nossa profunda gratidão a essa comissão informal de redacção, que foi constituída (por ordem alfabética) por António Teixeira, Daniel Gomes, Graça Nunes, Jorge Baptista, Luís Costa e Paulo Gomes. Agradecemos também a leitura aturada do primeiro capitulo pelo Eugénio Oliveira com valiosos comentários, e queremos fazer uma menção especial à Cristina Mota pelo cuidado e pormenor com que reviu todos os outros capítulos do livro, fazendo sugestões valiosíssimas. Embora como trabalho de bastidores, foi também muito importante a contribuição do Luís Miguel Cabral para o processamento das referências bibliográficas.

A organização de uma avaliação conjunta de raiz é algo que exige um grande empenhamento e muito trabalho, por isso nos parece importante que aquilo que se aprendeu e que foi feito possa ser reaproveitado por outros – os leitores do presente livro. Ao contrário de fechar aqui o trabalho nesta área e partir para outra, pretendemos também com este livro potenciar e possibilitar a preparação de futuras avaliações conjuntas em REM,

e em particular o Segundo HAREM que, à data de escrita deste prefácio, acaba de ser iniciado. Assim, tivemos o cuidado de republicar as directivas no presente volume e criar uma documentação mais cuidada dos próprios programas de avaliação, para facilitar a sua utilização e mesmo reprogramação.

Como nunca é demais ser repetido, na organização do HAREM não estivemos sós: contámos com a preciosa colaboração (por ordem alfabética) de Anabela Barreiro, Luís Costa, Paulo Rocha, Nuno Seco, Rui Vilela e Susana Afonso. E gostávamos de agradecer também a todos os participantes no Primeiro HAREM e também aos participantes no Encontro do HAREM no Porto pela participação e valiosas sugestões, participação e ideias essas que tudo fizemos para se encontrarem fielmente reflectidas pelo presente volume.

Como todo o trabalho feito no âmbito da Linguateca, o que nos moveu foi o desejo de uma melhoria significativa das condições do processamento computacional da língua portuguesa e, na esteira do modelo IRA (informação, recursos e avaliação), além da avaliação conjunta propriamente dita criámos recursos importantes para o REM em português (a colecção dourada, e os sistemas de avaliação). Com este livro, estamos a pôr em prática a terceira vertente, de informação.

Resta-nos agradecer a todos quantos tornaram este projecto (HAREM, e a própria Linguateca) possível, e acusar com gratidão o financiamento recebido, através dos projectos POSI/PLP/43931/2001 (2001-2006) e POSC 339/1.3/C/NAC (2006-2008).

Oslo e Lisboa, 5 de Novembro de 2007

Os editores

Diana Santos e Nuno Cardoso

Preface

This is a book about the First HAREM, an evaluation contest in named entity recognition in Portuguese, organized in the scope of the Linguateca project to foster R&D in the computational processing of Portuguese.

Although inspired by MUC, the path followed in HAREM was based on a different semantic model, aiming at identifying and classifying all proper names in text with the help of a set of 10 categories and 41 subcategories (called types), and allowing vague categories in the sense of merging two or more interpretations (as the geopolitical class in ACE, which conflates place and organization, but not only in that case).

HAREM had 10 participants in its first edition, which in fact included two evaluation events, the first event and Mini-HAREM (only for those who had participated before), which allowed us to perform some statistical validation studies and increase the evaluation resources. Because we had participants from non-Portuguese speaking countries (Denmark, Spain and Mexico), we have four chapters in English in this book, and therefore a preface in English is due as well.

This book reflects the participation and the discussion in the final HAREM workshop that took place in July 2006 after Linguateca's first summer school in Porto. It is organized in three parts, after an encompassing introduction:

1. Fundamentals of HAREM: history, preliminary studies, comparison with MUC and ACE, discussion of the semantic choices, statistical validation, a proposal for future venues, and a chapter summing up what was achieved and which future prospects we envisage.

2. Participation in HAREM: most participants wrote a chapter describing their systems,

approaches and results in HAREM evaluations, often also suggesting improvements or changes for the future.

3. HAREM documentation: the material produced by the organization, such as the guidelines for the annotation of the golden collection, the evaluation metrics, the evaluation software architecture, and the distribution of the golden collection as a regular corpus as well.

Following the usual procedure in Linguateca, abiding by the IRE model (information - resources - evaluation), we organized the evaluation contest, we made the resources therein available to the community, and we now gather and produce information about the whole endeavour, in the form of the present book.

We thank all participants in HAREM, our fellow organizers (Susana Afonso, Anabela Barreiro, Paulo Rocha, Nuno Seco and Rui Vilela), Luís Miguel Cabral who processed the book's references, and all those who participated as book reviewers (Luís Costa, Daniel Gomes, Paulo Gomes, Cristina Mota, Graça Nunes and António Teixeira) and whose help led to a considerable increase in quality.

All work in HAREM was done in the scope of the Linguateca project, jointly funded by the Portuguese Government and the European Union (FEDER and FSE) under contract references POSI/PLP/43931/2001 and POSC/339/1.3/C/NAC.

Oslo and Lisbon, 5th November, 2007

The editors,

Diana Santos and Nuno Cardoso

Autores

Antonio Toral Departamento de Lenguajes y Sistemas Informáticos, Universidad de Alicante, Espanha.

Andrés Montoyo Departamento de Lenguajes y Sistemas Informáticos, Universidad de Alicante, Espanha.

Bruno Martins Universidade de Lisboa, Faculdade de Ciências, LaSIGE, Portugal, *agora* Instituto Superior Técnico, Universidade Técnica de Lisboa, Portugal.

Christian Nunes Aranha Cortex Intelligence, Brasil.

Cristina Mota Instituto Superior Técnico, Universidade Técnica de Lisboa, Portugal, *agora* Instituto Superior Técnico, Universidade Técnica de Lisboa, Portugal / L2F, INESC-ID, Portugal / New York University, EUA.

Diana Santos Linguateca, SINTEF ICT, Noruega.

Eckhard Bick VISL, Institute of Language and Communication, University of Southern Denmark, Dinamarca.

José João Dias de Almeida Departamento de Informática, Universidade do Minho, Portugal.

Luís Sarmento Linguateca, CLUP, Faculdade de Letras da Universidade do Porto, Portugal, *agora* Faculdade de Engenharia da Universidade do Porto, Portugal.

Marcirio Chaves Universidade de Lisboa, Faculdade de Ciências, LaSIGE, Portugal.

Marília Antunes Universidade de Lisboa, Faculdade de Ciências, Portugal.

Mário J. Silva Universidade de Lisboa, Faculdade de Ciências, LaSIGE, Portugal.

Max Silberztein LASELDI, Université de Franche-Comté, França.

Nuno Cardoso FCCN, Linguateca, Portugal, *agora* Universidade de Lisboa, Faculdade de Ciências, LaSIGE, Portugal.

Nuno Seco Linguateca, Grupo KIS, Centro de Informática e Sistemas da Universidade de Coimbra, Faculdade de Ciências e Tecnologia da Universidade de Coimbra, Portugal.

Óscar Ferrández Departamento de Lenguajes y Sistemas Informáticos, Universidad de Alicante, Espanha.

Paulo Rocha Linguateca, Grupo KIS, Centro de Informática e Sistemas da Universidade de Coimbra, Faculdade de Ciências e Tecnologia da Universidade de Coimbra, Portugal.

Rafael Muñoz Departamento de Lenguajes y Sistemas Informáticos, Universidad de Alicante, Espanha.

Rui Vilela Departamento de Informática, Universidade do Minho, Portugal.

Thamar Solorio Human Language Research Institute, Universidade do Texas, Dallas, EUA.

Zornitsa Kozareva Departamento de Lenguajes y Sistemas Informáticos, Universidad de Alicante, Espanha.