Proposta de Projeto de Doutoramento a Desenvolver no Âmbito do 1º Concurso para Atribuição de Bolsas de Investigação na Área de Engenharia Informática

1. Título do projeto

Título: Tradução Automática da língua Bantu Moçambicana *Emakhuwa*

Palavras-chave: Processamento de Linguagem Natural; Tradução automática; Línguas com poucos recursos

Referência: CEE_EI_FEUP2

2. Instituições envolvidas

Instituição onde o doutoramento será realizado: FEUP

Outras instituições participantes no projeto de investigação: Universidade Lúrio - Moçambique

3. Equipa de Orientação

Orientador: Henrique Lopes Cardoso, Faculdade de Engenharia da Universidade do Porto (FEUP/LIACC)

Coorientador: Rui Sousa Silva, Faculdade de Letras da Universidade do Porto (FLUP/CLUP)

4. Descrição do Projeto

Tradução automática é o processo pelo qual um computador traduz automaticamente um documento de uma língua para outra. O Google Tradutor, por exemplo, permite atualmente traduzir textos em cerca de 109 línguas. Estas ferramentas apoiam-se em algoritmos sofisticados e grandes quantidades de dados paralelos (em pares de línguas) para alcançar uma tradução automática aceitável. Estes algoritmos não podem ser diretamente generalizados para todas as línguas, devido a especificidades de cada língua, nomeadamente ambiguidades sintáticas e semânticas – uma palavra ou expressão pode ter sentidos diferentes em contextos diferentes. Assim, estes algoritmos necessitam que haja uma quantidade considerável de recursos, quer linguísticos quer computacionais, para poder treinar modelos de modo a alcançar uma tradução automática aceitável.

Por este motivo, línguas com recursos linguísticos computacionais limitados, como o Emakhuwa, tendem a ser menos exploradas pela comunidade académica, pese embora haver benefícios claros no desenvolvimento de ferramentas computacionais, particularmente: na preservação da língua, na educação, na melhoria da compreensão e conhecimento da língua, na administração da justiça e na aplicação em situações de emergências ou desastres naturais.

A língua Emakhuwa é a língua materna falada em todas províncias do Norte de Moçambique, nomeadamente, Niassa, Cabo Delgado e Nampula. É também falada na Zambézia, que é a província que une o centro e o norte de Moçambique. O Emakhuwa é a língua mais falada em Moçambique, onde se estima que cerca de 5,8 milhões de moçambicanos a usem no seu dia-a-dia, em alternativa à língua oficial, o Português.

Este projeto de Doutoramento visa desenvolver um conjunto de recursos para facilitar o processamento computacional da língua Emakhuwa, potenciando a tradução automática para esta língua, nomeadamente a

partir do Português e do Inglês. Para tal, será necessário recolher, anotar e criar *corpora* – coleções de dados construídas a partir recursos como dicionários, bíblias, documentos legislativos, entre outros. A diversidade de dados permitirá almejar processos de tradução com especificidades do domínio, e ao mesmo tempo um processo de tradução o mais abrangente possível. Uma vez que os recursos existentes para a língua Emakhuwa são escassos, serão desenvolvidas e aplicadas técnicas e métodos utilizados para abordar línguas com poucos recursos, como por exemplo *transfer learning*, que consiste em emprestar recursos de outras línguas (como o Inglês ou o Português) para auxiliar no treino de modelos de tradução automática para a língua Emakhuwa. Com o desenvolvimento deste projeto, serão disponibilizados publicamente os recursos criados, o que alavancará o desenvolvimento de atividades de Processamento de Linguagem Natural (PLN) da língua Emakhuwa, permitindo assim que outros investigadores, desenvolvedores de *software* e empresas beneficiem do trabalho e das ferramentas disponibilizadas e adicionem novo conhecimento sobre o conhecimento existente.

Uma das linhas de investigação do Laboratório de Inteligência Artificial e Ciência de Computadores (LIACC/FEUP) é o PLN. Mais especificamente, no LIACC têm sido exploradas diversas técnicas de aprendizagem automática multi-lingue, baseadas em *transfer learning*, de modo a permitir tirar partido de grandes coleções de dados para Inglês no desenvolvimento de modelos computacionais para Português, em tarefas semanticamente exigentes. Uma das linhas de investigação do Centro de Linguística da Universidade do Porto (CLUP/FLUP) é a tradução, explorando a tradução técnica e científica em diversas línguas, e a sua interação com a linguística aplicada. A qualidade da formação nesta área na FLUP é certificada pela chancela da rede EMT (European Masters in Translation) da Comissão Europeia.

5. Referências Bibliográficas

Surafel M. Lakew, Matteo Negri, Marco Turchi: Low Resource Neural Machine Translation: A Benchmark for Five African Languages. In: AfricaNLP workshop at ICLR 2020.

Imankulova, A., Dabre, R., Fujita, A., Imamura, K.: Exploiting out-of-domain parallel data through multilingual transfer learning for low-resource neural machine translation. In: Machine Translation Summit XVII, pp. 128–139. Dublin, Ireland (Aug 2019)

Gil Rocha, Henrique Lopes Cardoso: A Comparative Analysis of Unsupervised Language Adaptation Methods. In: 2nd Workshop on Deep Learning for Low-Resource NLP (DeepLo 2019), EMNLP-IJCNLP 2019, Hong Kong, November 3, 2019. DOI: 10.18653/v1/D19-6102

Gil Rocha, Christian Stab, Henrique Lopes Cardoso and Iryna Gurevych: Cross-Lingual Argumentative Relation Identification: from English to Portuguese. In 5th Workshop on Argument Mining, EMNLP 2018, pp. 144-154. DOI: 10.18653/v1/W18-5217

Jiang, C., Yu, H.F., Hsieh, C.J., Chang, K.W.: Learning word embeddings for low-resource languages by PU learning. In: 2018 Conf. of the North American Chapter of the ACL. pp. 1024–1034. DOI: 10.18653/v1/N18-1093

Nguyen, T.Q., Chiang, D.: Transfer learning across low-resource, related languages for neural machine translation. In: Eighth Int. J. Conf. on NLP. pp. 296–301. Asian Federation of NLP, Taipei, Taiwan (Nov 2017)

Zoph, B., Yuret, D., May, J., Knight, K.: Transfer learning for low-resource neural machine translation. In: 2016 Conf. on Empirical Methods in Natural Language Processing. pp. 1568–1575. ACL, Austin, Texas (Nov 2016). DOI: 10.18653/v1/D16-1163