

LÍNGUAS EM RISCO: UM DESAFIO PARA A TECNOLOGIA DA LINGUAGEM

Somos testemunhas de uma revolução digital que está a ter um impacto radical na forma de comunicarmos e na sociedade em que vivemos. Os recentes desenvolvimentos nas áreas das Tecnologias da Informação e da Comunicação são por vezes comparados com a invenção da imprensa por Gutenberg.

O que pode esta analogia dizer-nos sobre o futuro da sociedade de informação europeia e sobre as nossas línguas em particular?

Na sequência da invenção da imprensa por Gutenberg, os avanços na comunicação e na partilha de conhecimentos foram concretizados através de inúmeras realizações, das quais a tradução da Bíblia do Latim para as línguas vernáculas da Europa é apenas um dos aspetos mais reconhecidos. Nos séculos seguintes, foram desenvolvidas novas técnicas para melhor lidar com o processamento da linguagem e a partilha de conhecimento:

- a padronização ortográfica e gramatical das principais línguas permitiu a rápida divulgação de novas perspectivas científicas e intelectuais;
- o desenvolvimento das línguas oficiais tornou possível aos cidadãos comunicarem dentro de certas fronteiras (muitas vezes políticas);
- o ensino e a tradução de línguas permitiram uma partilha de conhecimento entre línguas;
- a criação de diretrizes editoriais e bibliográficas garantiu a qualidade e a disponibilidade do material impresso;

- o surgimento de diferentes meios de comunicação, como jornais, rádio, televisão, livros e outros suportes e formatos, veio dar resposta às diferentes necessidades de comunicação.

Estamos a testemunhar uma revolução digital com um impacto que tem sido comparado com a invenção da imprensa por Gutenberg.

De forma análoga, nos últimos vinte anos, as Tecnologias da Informação e da Comunicação vieram ajudar ainda mais a automatizar e a facilitar o processamento da linguagem e a comunicação:

- as aplicações para edição de texto (*desktop publishing software*) substituem a datilografia e a composição tipográfica;
- as projeções de transparências são substituídas por apresentações em Powerpoint;
- o correio eletrónico permite receber e enviar documentos de forma mais rápida que o fax;
- o Skype permite realizar chamadas de telefone gratuitas ou a preços reduzidos pela internet, assim como videoconferências;
- os formatos de codificação de áudio e vídeo facilitam a troca de conteúdos multimédia;
- os motores de busca permitem aceder a informação com base em palavras-chave;

- os serviços de tradução online, como o Google Translate, produzem traduções rápidas ainda que apenas aproximadas;
- as plataformas de redes sociais como o Facebook, o Twitter ou o Google+ facilitam a comunicação, a colaboração e a partilha de informação.

Apesar de estas ferramentas e aplicações serem úteis, ainda não são capazes de apoiar, de forma sustentada, uma sociedade europeia multilingue para todos, onde a informação e os bens possam circular livremente.

2.1 FRONTEIRAS LINGUÍSTICAS ENTRAVAM A SOCIEDADE DE INFORMAÇÃO EUROPEIA

Não podemos saber exatamente como será o futuro da sociedade de informação. Há porém uma forte probabilidade de que a revolução nas tecnologias da comunicação venha a aproximar, de forma inovadora, pessoas que falam diferentes línguas. Esta situação vai pressionar toda a gente a aprender novas línguas e pressiona sobretudo os criadores de software a desenvolverem novas aplicações que permitam a inter-compreensão entre falantes de diferentes idiomas e o acesso a conhecimento partilhado. Este espaço económico e de informação global envolve a interação entre línguas, falantes e conteúdos no âmbito de novos meios de comunicação. A recente popularidade das redes sociais (Wikipédia, Facebook, Twitter, YouTube e, mais recentemente, o Google+) é apenas a ponta visível de um iceberg.

A economia e o espaço de informação globais colocam-nos perante mais línguas, falantes e conteúdos.

Hoje, podemos transmitir gigabytes de texto para todo o mundo em poucos segundos antes ainda de nos con-

seguirmos aperceber de que o conteúdo está redigido numa língua que não entendemos. De acordo com um recente relatório da Comissão Europeia, 57% dos utilizadores da internet compram bens e serviços em línguas que não a sua (o inglês é a língua estrangeira mais usada, seguido pelo francês, alemão e espanhol). Por sua vez, 55% dos utilizadores leem conteúdos numa língua estrangeira, enquanto apenas 35% utilizam outra língua para escrever mensagens de correio eletrónico ou colocar comentários na internet [2].

Há alguns anos atrás, o inglês era a língua franca na internet – a maior parte dos conteúdos estavam de facto em inglês – mas agora a situação mudou radicalmente. A quantidade de conteúdos online noutras línguas europeias (assim como em línguas asiáticas e do Próximo Oriente) aumentou exponencialmente.

Surpreendentemente, esta divisão digital criada pelas fronteiras linguísticas não recebe muita atenção pública. Ainda assim, levanta uma questão premente:

Que línguas europeias vão prosperar na informação em rede e na sociedade do conhecimento, e quais estão condenadas a desaparecer?

2.2 AS NOSSAS LÍNGUAS EM RISCO

Embora a imprensa escrita tenha ajudado a intensificar a troca de informação na Europa, também levou à extinção de muitas línguas europeias. Línguas regionais e minoritárias raramente foram impressas, como o Cornish e o Dálmata, e foram reduzidas a formas orais de transmissão, o que limitou o seu uso.

No futuro, terá a internet o mesmo impacto nas nossas línguas?

As cerca de 80 línguas da Europa são um dos mais ricos e importantes patrimónios culturais e uma parte vital do seu modelo social, que é único [3]. Enquanto línguas como o inglês e o espanhol sobreviverão no mercado

digital emergente, muitas línguas europeias poderão tornar-se irrelevantes numa sociedade ligada em rede. Isso enfraqueceria a posição global da Europa e iria contra o objetivo estratégico da participação de todos os cidadãos europeus em igualdade de circunstâncias, independentemente da sua língua.

A grande variedade de línguas na Europa é um dos seus patrimónios culturais mais ricos e importantes.

De acordo com um relatório da UNESCO sobre multilinguismo, as línguas são um meio essencial para o exercício dos direitos fundamentais, como a expressão política, a educação e a participação social [4].

2.3 A TECNOLOGIA DA LINGUAGEM É UMA TECNOLOGIA FACILITADORA

No passado, os esforços de investimento para a preservação das línguas concentraram-se no ensino e na tradução. De acordo com uma estimativa, o mercado europeu de tradução, interpretação, localização de software e preparação de websites para o mercado global foi de 8,4 mil milhões de euros em 2008 e deverá crescer 10% por ano [5]. No entanto, este número abrange apenas uma pequena parte das necessidades atuais e futuras da comunicação entre línguas.

A solução mais viável para garantir uma utilização ampla e continuada das várias línguas na Europa do futuro encontra-se no recurso a tecnologia apropriada, tal como recorremos a tecnologia apropriada para dar resposta às nossas necessidades, por exemplo, nas áreas da energia e dos transportes, ou para apoiar cidadãos com necessidades especiais, entre tantos outros casos.

A tecnologia da linguagem, dirigida a todas as formas de texto escrito e discurso falado, ajuda as pessoas a colabo-

rar, a concretizar negócios, a partilhar conhecimentos e a participar em debates sociais e políticos, independentemente das barreiras linguísticas e das aptidões informáticas de cada um.

A tecnologia da linguagem funciona muitas vezes “nos bastidores”, de forma invisível dentro de sistemas de software complexos, ajudando-nos já hoje em dia em tarefas como:

- encontrar informação com um motor de busca;
- verificar a ortografia e a gramática com um processador de texto;
- ver as recomendações para um produto numa loja online;
- seguir as indicações verbais de um sistema de navegação;
- traduzir páginas web com um serviço online.

A tecnologia da linguagem consiste num conjunto de aplicações nucleares que permitem uma série de procedimentos embebidos em sistemas mais amplos. Um dos objetivos desta coleção de Livros Brancos da META-NET é o de perceber o nível de desenvolvimento desta tecnologia para cada uma das línguas europeias.

A Europa precisa de tecnologia da linguagem robusta e económica para todas as línguas europeias.

Para manter a sua posição na linha da frente da inovação mundial, a Europa necessitará de tecnologia da linguagem que esteja adaptada a todas as línguas europeias e que seja igualmente robusta e económica, e bem integrada em ambientes de software-chave.

Sem tecnologia da linguagem suficientemente desenvolvida, não nos será possível alcançar uma experiência efetivamente interativa, multimédia e multilingue num futuro próximo.

2.4 OPORTUNIDADES PARA A TECNOLOGIA DA LINGUAGEM

O desenvolvimento da imprensa, com a duplicação rápida de uma imagem de texto, constituiu um avanço tecnológico fundamental. Mas os seres humanos continuam ainda a ter de fazer o trabalho árduo de buscar, apreciar, traduzir e resumir a informação.

A tecnologia da linguagem pode agora simplificar e automatizar muitos dos processos de tradução, produção de conteúdos e gestão de conhecimentos. Permite igualmente desenvolver interfaces de voz para eletrodomésticos, máquinas, veículos, computadores e robôs. As aplicações industriais e comerciais ainda estão num estágio inicial de desenvolvimento, mas os resultados em Investigação e Desenvolvimento estão a criar uma janela de oportunidade genuína. Por exemplo, a tradução automática já é razoavelmente precisa em certos domínios específicos e algumas aplicações experimentais já asseguram informação multilingue e gestão do conhecimento, assim como a possibilidade de produzir conteúdos, em várias línguas europeias.

Tal como a maioria das tecnologias, as primeiras aplicações para a linguagem humana, como as interfaces com o utilizador baseadas na voz ou os sistemas de diálogo, foram desenvolvidas para domínios altamente especializados, e em regra apresentam limitações de desempenho. Contudo, existem imensas oportunidades de mercado nas indústrias da educação e do entretenimento para a integração da tecnologia da linguagem em jogos, pacotes de jogos educativos, bibliotecas, ambientes de simulação ou programas de formação. Os serviços de informação móveis, os programas de aprendizagem de uma língua assistida por computador, os ambientes de e-learning, as ferramentas de autoavaliação e os programas de deteção de plágio são apenas alguns dos exemplos onde esta tecnologia pode desempenhar um papel importante. A popularidade das redes sociais, como o Twitter e o Facebook, sugerem uma maior neces-

sidade de sofisticação da tecnologia da linguagem para se poder monitorizar mensagens, resumir discussões, sugerir tendências de opinião, detetar respostas emocionais, identificar infrações aos direitos de autor ou encontrar usos indevidos.

A tecnologia da linguagem ajuda a superar os obstáculos colocados pela diversidade linguística.

A tecnologia da linguagem representa uma enorme oportunidade para a União Europeia. Pode ajudar a resolver a complexa questão do multilinguismo na Europa, nomeadamente ajudando a que diferentes línguas coexistam naturalmente nos negócios, nas organizações e nas escolas. Os cidadãos têm a necessidade de comunicar para além destas fronteiras linguísticas que cruzam o Mercado Comum Europeu e a tecnologia da linguagem pode assim ajudar a superar os obstáculos que ainda existem, permitindo o uso livre e ilimitado do idioma de cada um.

Pensando a longo prazo, a tecnologia da linguagem multilingue europeia poderá ser inclusive uma referência inovadora para os nossos parceiros globais e as suas comunidades multilingues.

A tecnologia da linguagem pode ser vista como uma forma de “tecnologia de apoio” que ajuda a ultrapassar os obstáculos da diversidade linguística e tornar as comunidades linguísticas mais acessíveis umas às outras.

2.5 DESAFIOS PARA A TECNOLOGIA DA LINGUAGEM

Apesar do progresso assinalável na área da tecnologia da linguagem nos últimos anos, o atual ritmo de progresso tecnológico e de inovação em termos de produtos é demasiado lento. As tecnologias com maior utilização,

como os corretores ortográficos e gramaticais em processadores de texto, são normalmente monolíngues e estão disponíveis apenas para um pequeno número de idiomas. Os serviços de tradução automática online, apesar de serem úteis para gerar rapidamente uma aproximação razoável ao conteúdo de um documento, veem-se enredados em imensa dificuldade quando lhe são pedidas traduções mais precisas e completas.

○ ritmo atual do progresso da tecnologia da linguagem é demasiado lento.

Devido à complexidade da linguagem humana, providenciar a modelação computacional dos nossos idiomas e testá-la no mundo real é um processo longo e oneroso, que exige compromissos de financiamento sustentados. A Europa tem, por isso, de manter o seu papel pioneiro de lidar com os desafios tecnológicos colocados por uma comunidade multilíngue, inventando novos métodos para acelerar o desenvolvimento de forma pervasiva.

2.6 AQUISIÇÃO DA LINGUAGEM POR SERES HUMANOS E POR MÁQUINAS

Para ilustrar como os computadores lidam com a linguagem natural e as razões pelas quais é difícil programá-los para esse efeito, vamos-nos centrar, muito brevemente, na forma como os seres humanos adquirem as suas primeira e segunda línguas, e depois ver como funcionam os sistemas de tecnologia da linguagem.

Os seres humanos adquirem competências linguísticas de dois modos diferentes. Os bebés aprendem uma língua interagindo linguisticamente e ouvindo as interações entre os pais, irmãos e outros membros da família. Por volta dos dois anos de idade, as crianças começam a produzir as suas primeiras palavras e frases curtas. Isto

só é possível porque os seres humanos têm uma predisposição genética para imitar e racionalizar o que ouvem. Aprender uma segunda língua numa idade mais avançada exige um maior esforço cognitivo, sobretudo quando quem aprende não está inserido numa comunidade de falantes dessa língua. Na escola, as línguas estrangeiras são normalmente adquiridas através do ensino da estrutura gramatical, vocabulário e ortografia, utilizando exercícios que descrevem conhecimentos linguísticos em termos de regras abstratas, tabelas e exemplos.

Os seres humanos adquirem aptidões linguísticas de dois modos diferentes: aprendendo a partir de exemplos e aprendendo as regras subjacentes.

Passando agora para a tecnologia da linguagem, os dois tipos principais de sistemas adquirem capacidades linguísticas de forma similar. As abordagens estatísticas permitem obter conhecimentos linguísticos a partir de vastas coleções de exemplos concretos de textos. Embora seja suficiente usar textos numa única língua para, por exemplo, treinar um corretor ortográfico, são necessários textos paralelos em duas ou mais línguas para o treino de um sistema de tradução automática. O algoritmo de aprendizagem automática pode então adquirir os padrões quanto ao modo como as palavras, expressões e frases completas são traduzidas.

Em regra, esta abordagem estatística requer milhões de frases para se obter um acréscimo significativo da qualidade no seu desempenho. Esta é uma das razões por que os fornecedores de motores de busca pretendem recolher o máximo de material escrito possível. Por exemplo, a correção ortográfica em processadores de texto ou serviços como o Google Search ou o Google Translate depende de abordagens estatísticas. A grande vantagem da estatística é que a máquina realiza uma rápida aprendizagem em séries contínuas de ciclos de treino.

Uma outra abordagem na tecnologia da linguagem, em geral, e na tradução automática, em particular, consiste na construção de sistemas baseados em regras. Peritos nas áreas da Linguística, Linguística Computacional e Engenharia Informática têm de, primeiro, codificar a análise gramatical (regras gramaticais) e compilar listas de vocabulário (léxicos). Isto requer imenso tempo e trabalho. Alguns dos principais sistemas de tradução automática baseados em regras têm estado em constante desenvolvimento desde há mais de 20 anos. A grande vantagem de sistemas baseados em regras é que os peritos têm um controlo mais pormenorizado sobre o processamento da linguagem. Isto torna possível corrigir de forma sistemática os erros no software e dar uma resposta detalhada ao utilizador, especialmente quando os sistemas baseados em regras são usados para a aprendizagem de línguas. Contudo, devido ao alto custo deste trabalho, a tecnologia da linguagem baseada em regras tem sido desenvolvida apenas para alguns idiomas até agora.

Como os pontos fortes e fracos de sistemas baseados em estatística e em regras tendem a ser complementares, a investigação atual concentra-se em abordagens híbri-

das que combinem as duas metodologias. No entanto, até agora, estas abordagens têm tido menos sucesso nas aplicações industriais do que nos laboratórios de investigação.

Os dois principais tipos de tecnologia da linguagem adquirem capacidades de processamento de uma forma algo similar à forma como os seres humanos o fazem.

Como vimos neste capítulo, muitas aplicações amplamente utilizadas na atual sociedade de informação dependem fortemente da tecnologia da linguagem. Devido à sua comunidade multilingue, isto é particularmente verdadeiro no espaço económico e de informação da Europa. Embora a tecnologia da linguagem tenha obtido progressos assinaláveis nos últimos anos, há ainda um enorme potencial para melhorar os resultados alcançados. Nos próximos capítulos, vamos descrever o papel do português na sociedade europeia de informação e no mundo e avaliar o estado atual da tecnologia da linguagem para a língua portuguesa.

The Portuguese Language in the Digital Age

Rehm, G.; Uszkoreit, H. (Eds.)

2012, VI, 68 p. 24 illus. in color., Softcover

ISBN: 978-3-642-29592-8