O sonho da tradução perfeita

Estamos longe de uma tradução automática de qualidade, mas há **projetos portugueses** que querem lá chegar

VIRGÍLIO AZEVEDO

Wang Ling está a fazer uma pesquisa curiosa de microblogues no L2F, o Laboratório de Sistemas de Línguas Faladas do INESC-ID, um centro de investigação ligado ao Instituto Superior Técnico (IST), em Lisboa. O estudante de doutoramento chinês trabalha, no âmbito do programa entre a Universidade Carnegie Mellon (EUA) e Portugal, com traduções de *tweets* em inglês de gente famosa do cinema, música, desporto e outras áreas de projeção mediática, tendo acesso a uma base de dados de três milhões de traduções manuais detetadas automaticamente por *software* criado para o efeito.

"O nosso objetivo é termos material de teste que possa ser usado para melhorar os sistemas de tradução automática que hoje são mais populares", explica Wang Ling ao Expresso. Gente famosa como o cantor Justin Bieber, com 58 milhões de seguidores no Twitter, a atriz Paris Hilton (13,1 milhões) ou o *rapper* Snoop Dogg (11,8 milhões), fornecem matéria-prima vasta e rica para esses testes. "O problema principal que enfrentamos é que sistemas como o Google Tradutor não conseguem traduzir corretamente muita coisa."

A tradução automática do Twitter está longe da versão correta, como mostram estes exemplos de três celebridades com milhões de seguidores

Quebra-cabeças científico

A empresa portuguesa UNBabel, ligada a projetos nesta área, selecionou para o Expresso três casos exemplares no Twitter relacionados com aquelas três celebridades, aplicou o Google Tradutor e apresentou depois uma versão final corrigida (ver ilustrações). Traduzir automaticamente um *tweet* com precisão é um quebra-cabeças científico e tecnológico, porque é escrito rapidamente, muitas vezes com erros ortográficos ou gralhas, palavras abreviadas e linguagem coloquial. Mas não é mais difícil do que uma tradução noutros contextos, porque as tecnologias e as metodologias usadas são as mesmas.

"Os projetos de tradução automática do Parlamento Europeu (24 línguas oficiais) tiveram um impacto global enorme, porque permitiram as abordagens estatísticas que hoje são muito usadas, graças ao grande manancial de dados — exemplos de textos e gravações de cada língua — que foram recolhidos", salienta Isabel Trancoso, professora catedrática do IST e dirigente do L2F.

Quando passamos da língua escrita para a língua falada as dificuldades são ainda maiores. Isabel Trancoso mostra um vídeo de uma conferência de imprensa do treinador do Benfica, Jorge Jesus, que tem uma maneira muito típica de falar, porque "come" com frequência vogais e últimas sílabas das palavras. "Se os motores de tradução automática apanham à entrada erros de transcrição, mesmo pequenos, estes vão somar-se aos erros de tradução, há um efeito de propagação e é o descalabro total."

As diferenças entre o português europeu, brasileiro e africano também são um problema complexo. A professora do IST dá um exemplo típico: a palavra "historicamente". No português do Brasil pronuncia-se "hisstoricámentchi". E no africano "historicámenté", em boa parte "pelo efeito da segunda língua falada a nível local" (crioulo ou outra).

Como se atacam então estes problemas? "A melhor solução é criar sistemas de tradução para domínios muito limitados, mais fáceis de controlar como, por exemplo, as reservas de quartos de hotel ou a compra de bilhetes de transporte, onde o leque de palavras e frases típicas é muito mais restrito. E aqui conseguimos fazer traduções automáticas muito boas." É o que acontece também no Parlamento Europeu, "onde é mais fácil traduzir automaticamente intervenções preparadas dos deputados do que um diálogo espontâneo".

"Tradução fiável, de alta qualidade e generalizada"

Mas será que a rápida evolução das tecnologias da linguagem vai permitir em breve uma tradução automática perfeita? "Estamos ainda longe de uma tradução fiável, de alta qualidade e generalizada para todos os pares de línguas", constata António Branco. O professor do Departamento de Informática da Faculdade de Ciências da Universidade de Lisboa (FCUL), dirige o Grupo de Linguagem Natural (NLX), onde trabalham 20 investigadores.

E lidera o QTLeap, um projeto financiado pela Comissão Europeia que envolve sete universidades e uma empresa em Portugal, Alemanha, Bulgária, Espanha, Holanda e República Checa. O objetivo é investigar e desenvolver uma metodologia inovadora para a tradução automática de oito línguas — português, alemão, checo, búlgaro, castelhano, basco, holandês e inglês — explorando novas soluções para alcançar traduções de maior qualidade. Arrancou em 2013 e tem um orçamento de três milhões de euros para três anos.

"Atualmente a tecnologia de tradução automática baseia-se em texto corrido, mas este projeto europeu pretende associar aos textos naquelas oito línguas a sua análise sintática e semântica", afirma o coordenador do QTLeap. Ou seja, já não são apenas as palavras, "mas também as estruturas gramaticais associadas às frases que são objeto de investigação". Resultado: esta nova metodologia de tradução automática permite "uma maior abstração, um afastamento dos detalhes específicos de cada língua, em benefício de uma representação mais universal de todas as línguas e de uma tradução de melhor qualidade".

Trata-se, assim, "de um dos projetos de investigação científica mais ambiciosos dos últimos anos nesta área", sublinha o professor da FCUL. E já há resultados no primeiro ano: "A aplicação de um sistema de tradução automática num cenário de uso real, nomeadamente num serviço de atendimento informático fornecido pela empresa portuguesa de sistemas informáticos inteligentes parceira do projeto", a Higher Functions.

Um utilizador de equipamentos ou serviços que precise de resolver um problema pode fazer uma pergunta através de um canal de *chat*. Caso exista um problema semelhante na base de dados do sistema, a resposta é fornecida sem intervenção humana, na mesma língua em que a pergunta foi feita. "Neste teste conseguimos ganhos de 60% na produtividade do serviço, isto é, na capacidade de resposta aos clientes sem intervenção humana", revela António Branco.

vazevedo@expresso.impresa.pt

O PAR IDEAL

Português/Inglês

■ Não faz sentido "classificar as línguas em geral de fáceis ou difíceis de traduzir por sistemas automáticos, mas apenas os pares de línguas", explica António Branco, investigador da Faculdade de Ciências da Universidade de Lisboa. "Assim, o par de línguas inglês/português é relativamente fácil na tradução, porque são línguas configuracionais, isto é, em que a ordem das palavras é relativamente rígida numa frase, entre sujeito, verbo, objeto direto e outras funções gramaticais"

■ Mas "o paradigma flexional do português é muito mais rico do que o de línguas como o inglês, em especial no que diz respeito aos verbos", afirma o Livro Branco "A Língua Portuguesa na Era Digital", lançado em 2012. Um verbo em português "pode ter diferentes marcas para aspeto, tempo, modo, pessoa, número, género ou polaridade, atingindo mais de 160 formas flexionadas diferentes, incluindo as simples e compostas"

■ Isabel Trancoso, professora catedrática do Instituto Superior Técnico, destaca também que "os chamados pronomes clíticos no português (se, te, lhe, lho, lo, la, nos, etc.) e nas línguas românicas dificultam a tradução automática para o inglês, porque o número de clíticos associado a cada verbo é muito grande" (são mais de 20). A posição dos pronomes clíticos na frase coloca, assim, desafios específicos no processamento automático da língua portuguesa

Português/Alemão

■ No alemão ou no finlandês, que têm declinações (ou sistemas de caso gramatical), a tradução automática para português (e o contrário) enfrenta algumas barreiras, "porque o sujeito é marcado pela declinação e não pela posição relativa na frase", o que significa que é mais livre, pode estar em posições diferentes, tal como o verbo ou o objeto, esclarece António Branco

■ Nas chamadas línguas aglutinativas como o alemão ou o turco, que juntam várias palavras numa só, a tradução para português (e o contrário) também pode ser mais difícil

Português/Castelhano

■ Nas línguas mais próximas do português (latinas, inglês), a tradução automática é de facto mais fácil e de melhor qualidade. A surpresa é que no castelhano a facilidade pode ser por vezes aparente. Tudo por causa da semântica lexical, isto é, do significado das palavras comuns às duas línguas, que por vezes é diferente, apesar de terem a mesma raiz etimológica (origem). Os especialistas chamam a estas palavras "falsos amigos" e há mais de 100. É o caso típico de "acordar", que em castelhano significa "lembrar", de "cena" ("jantar"), "padre" ("pai"), "berro" ("agrião"), "borrar" ("apagar") ou "palco" ("camarote")

PROJETOS EUROPEUS

17

centros de investigação de todo o país e o Instituto Camões estão envolvidos no projeto CLARIN Portugal, que pretende criar uma grande infraestrutura de investigação para a ciência e tecnologia da língua portuguesa, integrada na rede europeia CLARIN

3

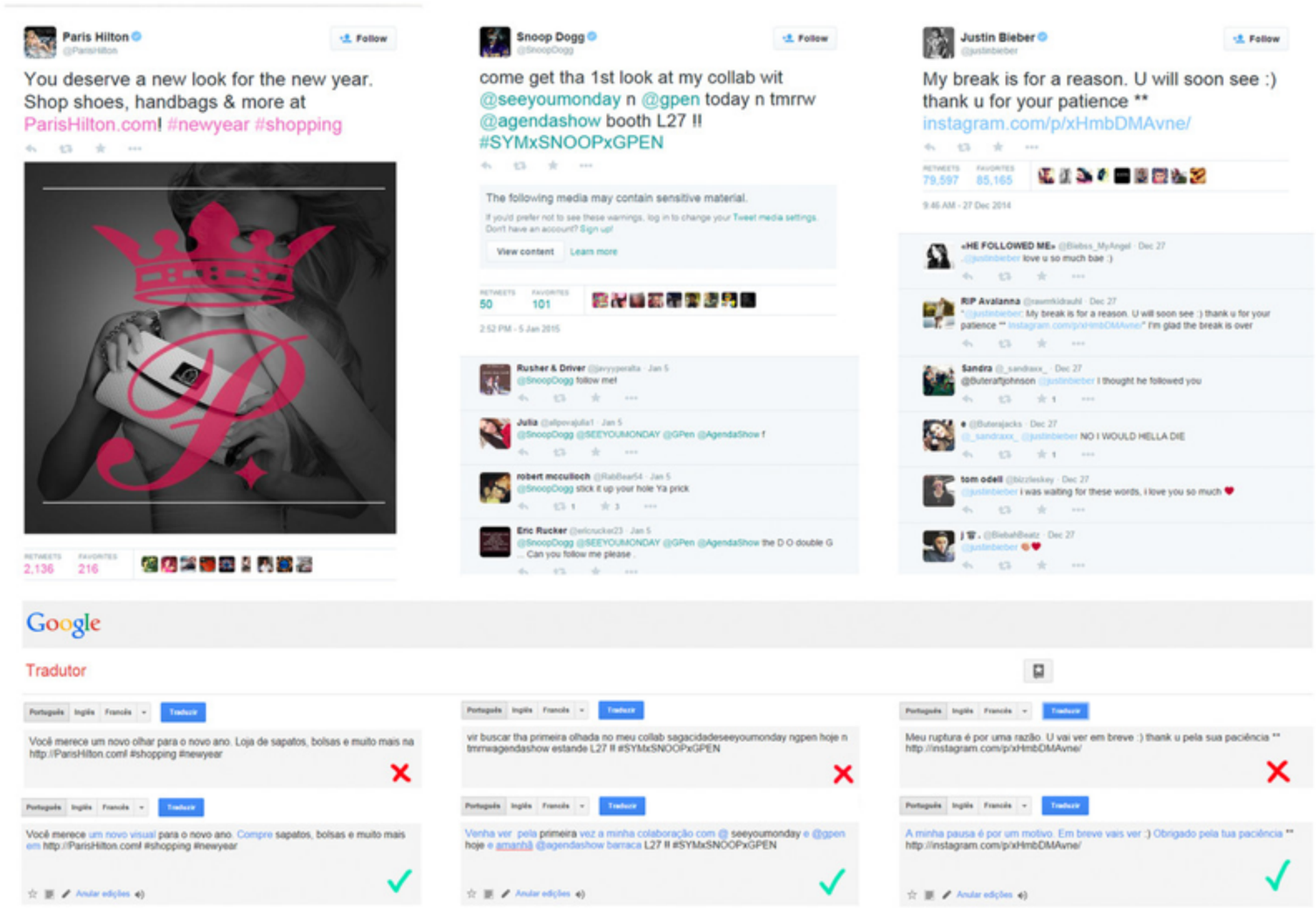
milhões de euros é o orçamento do projeto QTLeap, destinado a desenvolver metodologias inovadoras na tradução automática de oito línguas europeias. Tem oito parceiros e é liderado por Portugal

500

milhões de euros era a verba que a Aliança Tecnológica para a Europa Multilingue (META) queria ver cativada para o multilinguismo e tecnologia da linguagem no programa Horizonte 2020 da UE. Não conseguiu, porque o ambiente político geral é de falta de solidariedade entre os Estados-membros, que acham que cabe a cada país tratar da sua própria língua

CIÊNCIA

O sonho da tradução perfeita



A TRADUÇÃO AUTOMÁTICA DO TWITTER ESTÁ LONGE DA VERSÃO CORRETA, COMO MOSTRAM ESTES EXEMPLOS DE TRÊS CELEBRIDADES COM MILHÕES DE SEGUIDORES

Estamos longe de uma tradução automática de qualidade, mas há **projetos portugueses** que querem lá chegar

VIRGÍLIO AZEVEDO

O PAR IDEAL

PORTUGUÊS/INGLÊS

Não faz sentido "classificar as línguas em geral de fáceis ou difíceis de traduzir por sistemas automáticos, mas apenas os pares de línguas", explica António Branco, investigador da Faculdade de Ciências da Universidade de Lisboa. "Assim, o par de línguas inglês/português é relativamente fácil na tradução, porque são línguas configuracionais, isto é, em que a ordem das palavras é relativamente rígida numa frase, entre sujeito, verbo, objeto direto e outras funções gramaticais"

Mas "o paradigma flexional do português é muito mais rico do que o de línguas como o inglês, em especial no que diz respeito aos verbos", afirma o Livro Branco "A Língua Portuguesa na Era Digital", lançado em 2012. Um verbo em português "pode ter diferentes marcas para aspeto, tempo, modo, pessoa, número, género ou polaridade, atingindo mais de 160 formas flexionadas diferentes, incluindo as simples e compostas"

Isabel Trancoso, professora catedrática do Instituto Superior Técnico, destaca também que "os chamados pronomes clíticos no português (se, te, lhe, lho, lo, la, nos, etc.) e nas línguas românicas dificultam a tradução automática para o inglês, porque o número de clíticos associado a cada verbo é muito grande" (são mais de 20). A posição dos pronomes clíticos na frase coloca, assim, desafios específicos no processamento automáticoda língua portuguesa

Wang Ling está a fazer uma pesquisa curiosa de microblogues no L2F, o Laboratório de Sistemas de Línguas Faladas do INESC-ID, um centro de investigação ligado ao Instituto Superior Técnico (IST), em Lisboa. O estudante de doutoramento chinês trabalha, no âmbito do programa entre a Universidade Carnegie Mellon (EUA) e Portugal, com traduções de *tweets* em inglês de gente famosa do cinema, música, desporto e outras áreas de projeção mediática, tendo acesso a uma base de dados de três milhões de traduções manuais detetadas automaticamente por *software* criado para o efeito.

"O nosso objetivo é termos material de teste que possa ser usado para melhorar os sistemas de tradução automática que hoje são mais populares", explica Wang Ling ao Expresso. Gente famosa como o cantor Justin Bieber, com 58 milhões de seguidores no Twitter, a atriz Paris Hilton (13,1 milhões) ou o *rapper* Snoop Dogg (11,8 milhões), fornecem matéria-prima vasta e rica para esses testes. "O problema principal que enfrentamos é que sistemas como o Google Tradutor não conseguem traduzir corretamente muita coisa."

QUEBRA-CABEÇAS CIENTÍFICO

A empresa portuguesa UNBabel, ligada a projetos nesta área, selecionou para o Expresso três casos exemplares no Twitter relacionados com aquelas três celebridades, aplicou o Google Tradutor e apresentou depois uma versão final

PROJETOS EUROPEUS

17

centros de investigação de todo o país e o Instituto Camões estão envolvidos no projeto CLARIN Portugal, que pretende criar uma grande infraestrutura de investigação para a ciência e tecnologia da língua portuguesa, integrada na rede europeia CLARIN

3

milhões de euros é o orçamento do projeto QTLeap, destinado a desenvolver metodologias inovadoras na tradução automática de oito línguas europeias. Tem oito parceiros e é liderado por Portugal

PORTUGUÊS/ALEMÃO

No alemão ou no finlandês, que têm declinações (ou sistemas de caso gramatical), a tradução automática para português (e o contrário) enfrenta algumas barreiras, "porque o sujeito é marcado pela declinação e não pela posição relativa na frase", o que significa que é mais livre, pode estar em posições diferentes, tal como o verbo ou o objeto, esclarece António Branco

Nas chamadas línguas aglutinativas como o alemão ou o turco, que juntam várias palavras numa só, a tradução para português (e o contrário) também pode ser mais difícil

PORTUGUÊS/CASTELHANO

Nas línguas mais próximas do português (latinas, inglês), a tradução automática é de facto mais fácil e de melhor qualidade. A surpresa é que no castelhano a facilidade pode ser por vezes aparente. Tudo por causa da semântica lexical, isto é, do significado das palavras comuns às duas línguas, que por vezes é diferente, apesar de terem a mesma raiz etimológica (origem). Os especialistas chamam a estas palavras "falsos amigos" e há mais de 100. É o caso típico de "acordar", que em castelhano significa "lembrar", de "cena" ("jantar"), "padre" ("pai"), "berro" ("agrião"), "borrar" ("apagar") ou "palco" ("camarote")

500

milhões de euros era a verba que a Aliança Tecnológica para a Europa Multilingue (META) queria ver cativada para o multilinguismo e tecnologia da linguagem no programa Horizonte 2020 da UE. Não conseguiu, porque o ambiente político geral é de falta de solidariedade entre os Estados-membros, que acham que cabe a cada país tratar da sua própria língua

corrigida (ver ilustrações). Traduzir automaticamente um *tweet* com precisão é um quebra-cabeças científico e tecnológico, porque é escrito rapidamente, muitas vezes com erros ortográficos ou gralhas, palavras abreviadas e linguagem coloquial. Mas não é mais difícil do que uma tradução noutros contextos, porque as tecnologias e as metodologias usadas são as mesmas.

"Os projetos de tradução automática do Parlamento Europeu (24 línguas oficiais) tiveram um impacto global enorme, porque permitiram as abordagens estatísticas que hoje são muito usadas, graças ao grande manancial de dados — exemplos de textos e gravações de cada língua — que foram recolhidos", salienta Isabel Trancoso, professora catedrática do IST e dirigente do L2F.

Quando passamos da língua escrita para a língua falada as dificuldades são ainda maiores. Isabel Trancoso mostra um vídeo de uma conferência de imprensa do treinador do Benfica, Jorge Jesus, que tem uma maneira muito típica de falar, porque 'come' com frequência vogais e últimas sílabas das palavras. "Se os motores de tradução automática apanham à entrada erros de transcrição, mesmo pequenos, estes vão somar-se aos erros de tradução, há um efeito de propagação e é o descalabro total."

As diferenças entre o português europeu, brasileiro e africano também são um problema complexo. A professora do IST dá um exemplo típico: a palavra "historicamente". No português do Brasil pronuncia-se "hisstoricámentchi". E no africano "historicámentê", em boa parte "pelo efeito da segunda língua falada a nível local" (crioulo ou

VOLTAR AO TOPO

outra).

Como se atacam então estes problemas? “A melhor solução é criar sistemas de tradução para domínios muito limitados, mais fáceis de controlar como, por exemplo, as reservas de quartos de hotel ou a compra de bilhetes de transporte, onde o leque de palavras e frases típicas é muito mais restrito. E aqui conseguimos fazer traduções automáticas muito boas.” É o que acontece também no Parlamento Europeu, “onde é mais fácil traduzir automaticamente intervenções preparadas dos deputados do que um diálogo espontâneo”.

“TRADUÇÃO FIÁVEL, DE ALTA QUALIDADE E GENERALIZADA”

Mas será que a rápida evolução das tecnologias da linguagem vai permitir em breve uma tradução automática perfeita? “Estamos ainda longe de uma tradução fiável, de alta qualidade e generalizada para todos os pares de línguas”, constata António Branco. O professor do Departamento de Informática da Faculdade de Ciências da Universidade de Lisboa (FCUL), dirige o Grupo de Linguagem Natural (NLX), onde trabalham 20 investigadores.

E lidera o QTLeap, um projeto financiado pela Comissão Europeia que envolve sete universidades e uma empresa em Portugal, Alemanha, Bulgária, Espanha, Holanda e República Checa. O objetivo é investigar e desenvolver uma metodologia inovadora para a tradução automática de oito línguas — português, alemão, checo, búlgaro, castelhano, basco, holandês e inglês — explorando novas soluções para alcançar traduções de maior qualidade. Arrancou em 2013 e tem um orçamento de três milhões de euros para três anos.

“Atualmente a tecnologia de tradução automática

baseia-se em texto corrido, mas este projeto europeu pretende associar aos textos naquelas oito línguas a sua análise sintática e semântica", afirma o coordenador do QTLeap. Ou seja, já não são apenas as palavras, "mas também as estruturas gramaticais associadas às frases que são objeto de investigação". Resultado: esta nova metodologia de tradução automática permite "uma maior abstração, um afastamento dos detalhes específicos de cada língua, em benefício de uma representação mais universal de todas as línguas e de uma tradução de melhor qualidade".

Trata-se, assim, "de um dos projetos de investigação científica mais ambiciosos dos últimos anos nesta área", sublinha o professor da FCUL. E já há resultados no primeiro ano: "A aplicação de um sistema de tradução automática num cenário de uso real, nomeadamente num serviço de atendimento informático fornecido pela empresa portuguesa de sistemas informáticos inteligentes parceira do projeto", a Higher Functions.

Um utilizador de equipamentos ou serviços que precise de resolver um problema pode fazer uma pergunta através de um canal de *chat*. Caso exista um problema semelhante na base de dados do sistema, a resposta é fornecida sem intervenção humana, na mesma língua em que a pergunta foi feita. "Neste teste conseguimos ganhos de 60% na produtividade do serviço, isto é, na capacidade de resposta aos clientes sem intervenção humana", revela António Branco.

VOLTAR AO TOPO