Projetos sobre Tradução Automática do Português no Laboratório de Sistemas de Língua Falada do INESC-ID

Machine Translation Projects for Portuguese at INESC ID's Spoken Language Systems Laboratory

Anabela Barreiro
L2F - INESC-ID, Rua Alves Redol 9
1000-029, Lisboa
anabela.barreiro@inesc-id.pt

Wang Ling
L2F - INESC-ID Lisboa
Carnegie Mellon University
IST - Universidade de Lisboa

Luísa Coheur
L2F - INESC-ID
IST - Universidade de Lisboa

Fernando Batista
L2F - INESC-ID Lisboa
ISCTE-IUL

Isabel Trancoso
L2F - INESC-ID Lisboa
IST - Universidade de Lisboa

Resumo

As tecnologias da língua, de um modo especial as aplicações de tradução automática, têm o potencial de ajudar a quebrar barreiras linguísticas e culturais, apresentando um importante contributo para a globalização e internacionalização do português ao permitir que conteúdos linguísticos sejam partilhados 'a partir de' e 'para' esta língua. O presente artigo tem como objetivo apresentar o trabalho de investigação na área da tradução automática realizada pelo Laboratório de Sistemas de Língua Falada do INESC-ID, nomeadamente a tradução automática de fala, a tradução de microblogues e a criação de um sistema híbrido de tradução automática. Centraremos a nossa atenção na criação do sistema híbrido, que tem como objetivo a combinação de conhecimento linguístico, nomeadamente semântico-sintático, com conhecimento estatístico, de forma a aumentar o nível de qualidade da tradução.

Palavras Chave

Tradução Automática, Sistemas Híbridos, OpenLogos, Conhecimento Semântico-Sintático

Abstract

Language technologies, in particular machine translation applications, have the potential to help break down linguistic and cultural barriers, presenting an important contribution to the globalization and internationalization of the Portuguese language, by allowing content to be shared 'from' and 'to' this language. This article aims to present the research work developed at the Laboratory of Spoken Language Systems of INESC-ID in the field of machine translation, namely the automated speech translation, the translation of microblogs and the creation of a hybrid machine translation system. We will focus on the creation of the hybrid system, which aims at combining linguistic knowledge, in particular semantico-syntactic knowledge, with statistical knowledge, to increase the level of translation quality.

Keywords

Machine Translation, Hybrid Systems, OpenLogos, Semantico-Syntactic Knowledge

1 Introdução

A tradução automática, uma das mais complexas aplicações de língua natural, tem vindo a evoluir velozmente graças ao aumento quantitativo e qualitativo dos recursos linguísticos, nomeadamente dos dicionários electrónicos e corpos paralelos. Para além da evolução na construção de recursos, o desenvolvimento de algoritmos matemáticos que visam o mapeamento bilíngue e multilíngue de palavras, unidades lexicais multipalavra e expressões, e exploram a aprendizagem automática deste tipo de mapeamentos, vieram causar um impacto significativo nos sistemas estatísticos, especialmente em termos de velocidade e eficácia na aquisição de vocabulário. Com esta evolução de técnicas e recursos, a tradução automática tem vindo a afirmar-se no universo da tecnologia da língua, a conquistar a simpatia dos internautas e utilizadores das redes sociais e a estabelecer-se como ferramenta por excelência de globalização das línguas. No entanto, a tradução automática não é um problema resolvido e apesar da facilidade de acesso e utilidade das traduções produzidas por intermédio desta tecnologia, a sua qualidade é ainda limitada por falta de conhecimento linguístico inerente a um tradutor humano. Neste contexto, os utilizadores da tradução automática ainda a consideram "pouco fiável". A falta de precisão nas traduções obtidas através desta ferramenta tecnológica está na base do trabalho de investigação em tradução automática decorrente no INESC-ID, nomeadamente a tradução de fala (Secção 3.1), a tradução de microblogues (Secção 3.2) e a criação de um sistema híbrido de tradução automática, tema que será explorado com algum detalhe adicional (Secção 3.3). Este trabalho visa enriquecer os atuais sistemas de tradução automática de base estatística com conhecimento linguístico (nomeadamente morfológico e semântico-sintático), de modo a que o mapeamento de uma frase na língua-fonte numa frase na língua-alvo

reflita o modo como a linguagem é processada pelo cérebro humano.

2 Paradigmas da Tradução Automática

A tradução automática tem evoluído com base em dois paradigmas principais: um baseado em conhecimento linguístico e outro baseado em dados e métodos estatísticos de mapeamento desses dados.

Os sistemas construídos com base em conhecimento linguístico, conhecidos como sistemas por regras, era o paradigma vigente até ao final dos anos 90 (Nirenburg et al., 2003; Scott, 2003). Estes sistemas não precisam de corpos paralelos, produzem tradução de qualidade, e funcionam bem até com poucos dados e poucas regras em domínios especializados, desde que se baseiem em dicionários e terminologias de qualidade. Também funcionam bem em línguas com um sistema morfológico rico em que algumas regras de flexão são suficientes para traduzir um grande número de formas com o mesmo radical. No entanto, o desenvolvimento deste tipo de sistemas envolve um grande investimento de tempo e recursos humanos especializados, necessários ao desenvolvimento de recursos linguísticos avançados para cada par de línguas a ser traduzido.

Os custos e morosidade envolvidos na construção dos sistemas por regras conduziram ao aparecimento de um novo paradigma, nos finais dos anos 90, os modelos estatísticos de tradução automática. Inicialmente baseados no alinhamento de n-gramas, estes modelos foram evoluindo com o tempo e começando a estender-se a expressões, de natureza não linguística (Koehn, 2007)[*]. Os sistemas estatísticos de tradução automática criam-se com base em corpos paralelos aos quais são aplicados algoritmos e técnicas de alinhamento dos vários elementos ou expressões da frase. Os algoritmos permitem encontrar padrões e prever a probabilidade de uma palavra ser a tradução de outra baseado no número de ocorrências dessa palavra em contexto nas duas línguas. O tempo de desenvolvimento dos sistemas estatísticos é rápido, desde que existam corpos paralelos para os pares de línguas que se pretendam traduzir, e por este motivo, a sua construção é muito mais económica do que a dos sistemas por regras. Até recentemente, os sistemas estatísticos envolviam conhecimento linguístico muito limitado, resultando em erros crassos há muito resolvidos pelos sistemas por regras. Mesmo com uma quantidade avultada de dados, como a que é utilizada por sistemas como o Google Translate, é necessária a pós-edição de erros simples, como os de concordância entre substantivo e adjetivo qualificativo, entre sujeito e verbo, entre outros. Para além disso, esses sistemas estão totalmente dependentes da quantidade e qualidade dos corpos paralelos que utilizam para os seus alinhamentos. Há línguas para os quais os corpos paralelos abundam e são de qualidade aceitável, como é o caso do inglês-mandarim, mas para outras línguas, como o basco, os corpos paralelos são escassos ou de qualidade reduzida, o que torna difícil a extração de generalizações necessária à tradução (Labaka et al., 2007). Apesar de terem sido propostos modelos mais sofisticados para a tradução de línguas com sistemas morfológicos complexos (Chahuneau et al., 2013), os sistemas por regras apresentam-se como a solução mais viável para traduzir estas línguas por necessitarem de uma quantidade menor de dados para traduzir as diferentes formas flexionadas de uma palavra.

2.1 Os Sistemas OpenLogos e Google Translate

Um dos modelos de tradução mais antigos é o sistema Logos (Scott, 2003), um sistema comercial baseado em regras, desenvolvido entre 1970 e 2001, e agora explorado na sua versão em código aberto: OpenLogos (Barreiro et al., 2011), adaptado pelo DFKI e disponível no SourceForge. O OpenLogos é considerado um sistema de tradução automática de qualidade, que comporta oito pares de línguas, contemplando a tradução de inglês para alemão, francês, espanhol, italiano e português, e de alemão para inglês, francês e italiano. A qualidade da tradução do OpenLogos resulta da sua componente semântico-sintática e da análise da língua de forma a que esta seja "entendida" pelo sistema computacional, tal como será descrito na Secção 5. A aproximação Logos assemelha-se, em espírito, à aproximação estatística, na medida em que as regras de base gramatical (semântico-sintática) são aplicadas a padrões em contexto. O conhecimento linguístico envolvido no sistema permite colmatar dificuldades e fraquezas apresentadas pelos métodos estatísticos, colocando-o na posição de plataforma ideal para uma solução híbrida.

Um dos sistemas de tradução automática mais populares é o Google Translate, disponível gratuitamente através da internet. O Google Translate utiliza um método estatístico para traduzir, que tem como alicerce o sistema em código aberto Moses (Koehn et al., 2003), usado por uma larga comunidade de investigadores e por alguns sistemas comerciais, como o Asia Online, entre outros. O Go-

[*] A tradução automática estatística é um paradigma com diversas vertentes, que não exploraremos neste artigo. Aconselhamos a leitura do seguinte estudo: http://www.cs.jhu.edu/~alopez/papers/survey.pdf

ogle Translate tem a forte vantagem de aceder a uma quantidade muito grande de corpos paralelos recolhidos da web, o que lhe permite a tradução de um número elevado de pares de línguas (cerca de 80), que varia em qualidade dependendo de fatores como a proximidade entre a língua-fonte e a língua-alvo, ou a quantidade e qualidade dos corpos disponíveis para a tradução de cada par de línguas. O Google Translate é um sistema comercial, pelo que não se sabe que módulos integra e se algum desses módulos contém conhecimento semântico necessário à tradução de qualidade.

2.2 Modelos Híbridos

Como o objetivo último dos investigadores e desenvolvedores de sistemas de tradução automática é o de criar sistemas que produzam tradução de qualidade comparável à que é produzida por tradutores humanos, a necessidade de um paradigma mais robusto e linguisticamente mais avançado tem vindo a afirmar-se nos últimos anos. Deste modo, surgiram os sistemas híbridos, que têm a vantagem de poder usufruir do trabalho de investigação de várias décadas, donde resultou, por um lado, a invenção e aperfeiçoamento das técnicas estatísticas que aceleram o processo de aquisição lexical e de tradução, e por outro lado, o desenvolvimento de maior quantidade de recursos linguísticos de melhor qualidade e para mais línguas. Os progressos alcançados nos paradigmas de vertente linguística e matemática da tradução automática tornaram-se, assim, um campo fértil para o desenvolvimento da nova geração de sistemas de tradução que ambicionam uma tradução de qualidade mais próxima à do tradutor humano. Vários métodos têm sido propostos para combinar tradução automática baseada em regras com tradução automática estatística.

Alguns sistemas processam estatisticamente as traduções obtidas a partir de um sistema por regras, enquanto que outros utilizam regras de base gramatical para processar previamente os dados, regras essas que ajudam a guiar o sistema estatístico. Um modelo de hibridização simples é o que combina traduções do mesmo texto por dois sistemas conceptualmente distintos (Heafield e Lavie, 2011; Eisele et al., 2003). Para além da combinação de sistemas, os modelos híbridos podem assentar, por um lado, na aplicação de técnicas estatísticas de alinhamento de expressões ou exemplos para melhorar a cobertura num sistema de tradução por regras (Eisele et al., 2003; Sánchez-Martínez et al., 2009) ou no melhoramento da qualidade de tradução de um sistema por regras através da utilização de métodos estatísticos de pós-edição (Simard et al., 2007; Elming, 2006; Dugast et al., 2007; Terumasa, 2007). Por outro lado, também tem sido proposta a integração de conhecimento linguístico em sistemas de tradução automática estatística (Satoshi et al., 1997). O trabalho de Niessen e Ney (2004) usa modelos lexicais hierárquicos para induzir a forma base das palavras em alemão para a tradução de termos compostos. Por outro lado, Ueffing et al. (2003) usa conhecimento linguístico para obter a forma correta das palavras quando a tradução é feita para línguas morfologicamente ricas como é o caso do espanhol e do catalão. Finalmente, a informação linguística pode também ser usada para melhorar a qualidade dos alinhamentos através do uso de informação acerca da categoria gramatical, como foi feito por Koehn e Night (2001) para o par de línguas alemão-inglês.

Embora não exista ainda indicação de que abordagem híbrida seja mais eficaz e conduza a um aumento do nível da qualidade da tradução, uma junção dos pontos mais fortes de cada paradigma ajuda a melhorar a tradução alcançada pelos novos sistemas e, como tal, a hibridização de sistemas de tradução automática continua a representar uma linha de investigação promissora. É no âmbito dessa aposta que se enquadra o trabalho de investigação descrito na Secção 3.3, onde exploramos uma aproximação nova à tradução automática híbrida, partindo da integração do conhecimento semântico-sintático do sistema OpenLogos em modelos estatísticos. Mas, antes de chegarmos a esse trabalho, abordaremos outros desafios na área da tradução automática, que são o da tradução de fala para fala, apresentado na secção 3.1, e o da tradução da linguagem usada em microblogues, descrito na secção 3.2.

3 Tradução Automática no INESC-ID

3.1 Tradução Automática de Fala

A tarefa de tradução ganha uma utilidade adicional se tiver como objetivo traduzir fala para fala. No entanto, a tarefa em si torna-se igualmente muito mais complexa, pelo que a investigação na área da tradução de fala para fala enfrenta desafios adicionais para além dos usualmente associados às tarefas de tradução de texto para texto. Este processo pode ser visto como uma sequência de três etapas: reconhecimento automático de fala (passagem de fala para texto na língua-fonte), tradução (passagem do texto na língua-fonte para texto na língua-alvo) e síntese (passagem do texto na língua-alvo para fala). Assim, quaisquer erros que ocorram em cada um destes módulos dificultam seriamente as tarefas dos módulos subsequentes. Neste cenário, o primeiro grande desafio consiste na tradução de fala espontânea. Dado que os módulos de

reconhecimento e de tradução são normalmente treinados com base em textos escritos, torna-se extremamente difícil processar hesitações, repetições, pausas preenchidas e expressões não gramaticais, muito frequentes em fala espontânea.

O projeto PT-STAR (financiado pela FCT, no quadro do programa Carnegie Mellon – Portugal e recentemente terminado) teve como objetivo melhorar os sistemas de tradução de fala para fala 'de' e 'para' português, focando-se na interligação entre os três principais módulos destes sistemas. Fizeram parte deste consórcio o Laboratório de Sistemas de Língua Falada (L2F), do INESC-ID Lisboa, o Instituto de Tecnologias da Língua (LTI) da Universidade de Carnegie Mellon, o Centro de Linguística da Universidade de Lisboa e a Universidade da Beira Interior. O estado da arte atual em tradução de fala para fala mostra uma integração relativamente fraca entre os três módulos, não explorando as sinergias existentes entre o reconhecimento e a tradução, entre a tradução e a síntese e ainda entre o reconhecimento e a síntese. Por exemplo, o módulo de reconhecimento escolhe normalmente uma única hipótese de transcrição que será a entrada do módulo de tradução. Se, em alternativa a esta hipótese, for oferecida ao módulo de tradução uma lista de possíveis transcrições, este pode decidir qual a mais adequada aos modelos de tradução. Por outro lado, o módulo de síntese assume que receberá como entrada texto fluente, o que usualmente não acontece quando essa entrada resulta de um módulo de tradução automática. Assim, de maneira a que a interligação entre estes dois módulos seja mais robusta, a estratégia de síntese tem que ser modificada a fim de evitar a produção não compreensível de voz, por exemplo, eliminando palavras com baixa confiança. Ser capaz de transferir o foco principal (ou ênfase) de entrada da língua-fonte para a língua-alvo é outra tarefa extremamente desafiante e que obriga a que exista uma ligação entre os módulos de reconhecimento e de síntese.

3.1.1 Cenários de Aplicação

Dois grandes cenários foram palco do grosso dos avanços no projeto PT-STAR: a tradução de TED Talks e de notícias televisivas. Na tradução das TED Talks, a adaptação de domínio e de conversão de voz foram os principais desafios, tendo o sistema de lidar com aplausos ocasionais e risos da plateia. A tradução das notícias televisivas tornou-se também um cenário de grande interesse, pois permitiu, para além de testar técnicas de adaptação ao domínio, trabalhar sobre fala controlada (por exemplo, dos pivôs do telejornal), bem como fala espontânea. O leitor pode encontrar uma demonstração, feita em tempo real, do sistema de tradução de fala para fala desenvolvido no projeto PT-STAR no seguinte endereço: www.l2f.inesc-id.pt/wiki/index.php/Demos. Nesta demonstração o reconhecedor foi treinado com textos de jornais e o tradutor com os habituais dados do Europarl (Koehn, 2005). A qualidade da tradução do sistema PT-STAR deve-se à proximidade entre os dois domínios.

3.1.2 Desafios em Destaque

Dos vários trabalhos de investigação abordados no projeto PT-STAR, destacamos dois que ilustram a problemática em mãos: o primeiro na área do reconhecimento, o segundo na fronteira entre a tradução e a síntese.

Enriquecimento de transcrições

A saída de um reconhecedor consiste apenas numa sequência de palavras. A capacidade de enriquecer esta sequência com a pontuação apropriada afeta não só a qualidade da transcrição, mas também a possibilidade de melhorar o passo de tradução, pois uma segmentação adequada é fundamental para o sucesso da tarefa de tradução (por exemplo, conseguiu-se uma melhoria de 2 pontos BLEU numa experiência que consistiu em passar imediatamente ao tradutor todos os segmentos terminados com qualquer sinal de pontuação e não apenas os segmentos terminados com um ponto final (Grazina, 2010). Assim, uma das tarefas deste projeto consistiu no desenvolvimento de módulos capazes de inserir vários sinais de pontuações em transcrições automáticas. Duas estratégias diferentes foram exploradas no que diz respeito ao ponto final e à vírgula. O primeiro fez uso de fontes de informação que podem ser encontradas em textos; o segundo consiste na introdução de características prosódicas: além de pistas lexicais, foram utilizadas pistas baseadas no tempo e em características do falante. Ambas as estratégias melhoraram os resultados iniciais. Por exemplo, para a saída do reconhecedor para português, a pontuação foi melhorada em cerca de 5,6% (Batista et al., 2012).

Melhoria da síntese resultante de um texto traduzido automaticamente

A saída do módulo de tradução automática é muitas vezes inadequado para ser passada diretamente ao módulo de síntese, o que representa um problema na interface entre a tradução automática e a síntese. Como os modelos do sintetizador são geralmente treinados com texto fluente, este sintetizará a saída do tradutor assumindo a sua fluência,

o que tornará a fala resultante difícil de entender. Assim sendo, um dos desafios deste projeto prendeu-se com a tentativa de otimizar o sintetizador de modo a que a compreensão do texto fosse a melhor possível, apesar dos erros de tradução. Várias técnicas foram testadas (Parlikar et al., 2010), tais como a utilização de material de preenchimento de pausas que provaram ser de utilidade para melhorar a inteligibilidade da fala.

3.2 Tradução Automática de Microblogues

Outra problemática da tradução automática abordada em projetos desenvolvidos no INESC-ID é o da tradução de linguagem de microblogues. Na última década, os microblogues, como o Facebook, o Twitter, o Youtube ou o Sina Weibo (versão chinesa do Twitter), têm sido alvo de uma atenção especial pela comunidade científica por razões que se prendem com a quantidade de pessoas que as utilizam e com o volume de informação existente neste domínio. No entanto, os conteúdos textuais abrangidos pelos microblogues caracterizam-se por incluírem termos pouco formais e linguagem não padronizada. Alguns exemplos incluem a presença de abreviaturas como a da expressão em inglês *r u still following me or what?*. Estas expressões são geralmente problemáticas para os sistemas de tradução automática, por dois motivos principais, que passaremos a descrever.

No primeiro caso, os modelos de tradução não são treinados com dados deste domínio, simplesmente porque eles não existem. Em consequência, os sistemas de tradução treinados com dados fora do domínio dos microblogues não estão aptos para traduzir a linguagem nele usada. Este problema dá origem a erros de tradução, tais como os que são evidenciados na tradução para português do exemplo acima pelo Google Translate: *r u ainda me seguindo ou o quê?*, em que as abreviaturas *r* e *u* simplesmente não são traduzidas. Em resposta a este problema, foi construído um sistema de extração automática de traduções do Twitter e Sina Weibo. Este sistema é motivado pela observação que há alguns utilizadores que traduzem os seus tuítes e estes podem ser extraídos e usados para melhorar significativamente os sistemas de tradução no domínio. Por exemplo, o tuíte *Male Body Painting - Pintura Corporal Masculina (Essa Moda Pega?)* contém a tradução do nome composto em inglês *Male Body Painting* para o nome composto em português *Pintura Corporal Masculina*. Estas traduções no tuíte publicado podem ser encontradas através de um algoritmo de "emparelhamento baseado em conteúdo" (Resnik e Smith, 2003), que representa o estado da arte (Ling et al., 2013b). Este algoritmo explora técnicas para a extração de corpos paralelos da web, onde dois documentos são identificados como traduções um do outro se existir uma grande percentagem de palavras que são consideradas como boas traduções entre esses documentos. O contributo principal desse trabalho consiste na extensão do algoritmo aos casos em que a tradução se encontra no mesmo documento, como acontece no caso dos microblogues. Com este algoritmo é possível obter uma grande quantidade de traduções para várias línguas, incluindo o português. As experiências feitas com 3 milhões de frases paralelas para chinês-inglês mostram que a utilização desse corpo pode fazer aumentar consideravelmente a qualidade da tradução. A melhoria deve-se essencialmente ao facto de os sistemas existentes serem incapazes de traduzir termos frequentes como *u*, *thx* e *r*, responsáveis pela degradação da qualidade da tradução.

O segundo problema encontrado nos atuais sistemas de tradução automática estatística está relacionado com facto de estes sistemas modelarem a linguagem como sequências de palavras. Por exemplo, os modelos consideram as formas *hellllo* e *hello* como dois códigos (*tokens*) diferentes. Assim sendo, os modelos de tradução apenas conseguem traduzir a forma *hellllo* se esta constar no conjunto de dados de treino do sistema. Este problema não pode ser tratado através da extração de mais corpos por não ser possível obter traduções para todas as formas que aparecem no corpo (e.g. *helo*, *heeeeello*, *ello*, etc.). É necessário que o modelo aprenda a generalizar o processo de tradução de maneira a que reconheça todas estas formas como variantes da palavra *hello*. Para solucionar o problema, Ling et al. (2013a) propõem um sistema de normalização que aprende a converter as frases informais em paráfrases com a mesma informação, mas representada de forma padronizada. Este sistema de normalização converte, por exemplo, a frase *r u still following me or what??* em *Are you still with me or what?*. Este parafraseamento permite um processamento mais fácil destas mensagens, quer por sistemas de tradução, quer por outros sistemas de processamento de linguagem natural.

O sistema de tradução de microblogues assenta em tecnologias de parafraseamento construídas a partir da tradução (Callison-Burch, 2007). O sistema usa um corpo paralelo inglês-chinês, onde a porção chinesa é traduzida de forma automática para inglês, gerando um corpo de paráfrases que constitui uma versão alternativa do corpo original. É possível usar esse corpo recolhido automaticamente para obter um mapeamento das frases onde podem ocorrer variações estilísticas entre a frase original e a tradução. Uma vantagem que o método apresenta é tornar possível o uso de frases paralelas

	Fenómeno Linguístico	Original em Inglês	OpenLogos	Google Translate
(1)	Conc. SUJ-V PRON OI	Kennedy interviewed you.	Kennedy entrevistou-o.	*Kennedy entrevistei.
(2)	PE vs PB	Kennedy interviewed me.	Kennedy entrevistou-me. (PE)	Kennedy me entrevistou. (PB)
(3)	Conc. SUJ-V PRON OI	Kennedy interviewed us.	Kennedy entrevistou-nos.	*Kennedy nos entrevistaram.
(4)	Conc. SUJ-V	Me and her interviewed Kennedy.	Eu e ela entrevistámos Kennedy.	*Eu e ela entrevistou Kennedy.
(5)	Conc. N-ADJ	Kennedy has a bookcase that is heavy.	Kennedy tem uma estante que é pesada.	*Kennedy tem uma estante que é pesado.
(6)	V-Aux	Kennedy hired women who were competent.	Kennedy contratou mulheres que foram competentes.	*Kennedy contratou mulheres que estavam competente.
(7)	V-Sem	She manages whom?	Ela dirige quem?	*Ela consegue quem?
(8)	HOMO N-V HOMO V-N	Managers work.	Os gerentes trabalham.	*Gerentes de trabalho.
(9)	HOMO V-N	List women who have bookcases.	Enumere mulheres que têm estantes.	*Lista de mulheres que têm estantes.
(10)	V-PT vs V-PP PRON REFL	The women evaluated themselves.	As mulheres avaliaram-se.	*As mulheres avaliadas si.

Tabela 1 – Tradução de aspetos da gramática traduzidos pelos sistemas OpenLogos e Google Translate.

em línguas para as quais é viável extrair uma grande quantidade de dados, como é o caso do par inglês-chinês e criar sistemas de normalização para estas línguas, estendendo-a a pares de línguas para os quais existem poucos dados traduzidos, como é o caso do inglês-português ou do chinês-português. Outra vantagem consiste na possibilidade de utilizar as ferramentas de parafraseamento e normalização para outras tarefas de processamento de linguagem natural. Xu et al. (2014) apresentam a extração, por via de métodos estatísticos, de pares de tuítes semelhantes que constituem paráfrases. No entanto, para a finalidade de normalização não há garantia que os tuítes colecionados tenham o mesmo nível de língua não-padrão e as paráfrases podem não ser adequadas para a criação de um sistema de normalização.

3.3 Tradução Automática com Conhecimento Semântico-Sintático

O trabalho de investigação em tradução automática híbrida atualmente em vigor no INESC-ID consiste na criação de um novo modelo que combina conhecimento linguístico com tradução automática estatística. Para o efeito, partimos do sistema OpenLogos, um sistema com características peculiares que lhe permitem servir de plataforma para a criação desse novo modelo híbrido. Embora a hibridização exija um esforço significativo, os seus princípios basilares assentam na integração de conhecimento linguístico, que já deu provas de sucesso na tradução de muitos aspetos da gramática. A Tabela 1 apresenta a tradução de frases com diferentes níveis de gramaticalidade/naturalidade extraídas de um corpo construído para testar fenómenos linguísticos no sistema OpenLogos contrastando-a com a tradução obtida através do sistema Google Translate. Estas frases apresentam fenómenos variados como a concordância entre o sujeito e o verbo (exemplos (1), (3) e (4)) ou a concordância entre o nome e o adjetivo qualificativo numa construção relativa (exemplo (5)), os diferentes tipos de pronome (objeto indireto, reflexo, etc. (exemplos (1), (3) e (10)), a escolha do verbo auxiliar (*ser*, *estar* (exemplo (6)), a semântica do verbo (exemplo (7)), as palavras homógrafas, como nome-verbo (exemplos (8) e (9)), as formas terminadas em *-ed*, que podem ser formas do pretérito perfeito ou do particípio passado (exemplo (10)), entre outros. As variantes europeia (PE) e brasileira (PB) do português (exemplo (2)) também surgem em contraste nas traduções obtidas através dos dois sistemas. As frases traduzidas pelo sistema Google Translate refletem a dificuldade e imprevisibilidade dos sistemas estatísticos quando confrontados com alguns fenómenos gramaticais, principalmente em frases curtas e ambíguas para os quais não foram treinados. Em contrapartida, estes sistemas conseguem uma cobertura lexical mais abrangente, o que, por um lado representa uma grande vantagem, mas por outro, pode provocar nos utilizadores uma sensação de que os sistemas estatísticos traduzem melhor do que os sistemas por regras, mesmo que essa tradução se deva essencialmente ao acesso a grandes quantidades de dados que permitem treinar os sistemas estatísticos em domínios específicos de interesse e de utilidade para os seus utilizadores.

A avaliação desenvolvida em trabalho anterior (Barreiro et al., 2013) mostra que tanto os modelos linguísticos como os estatísticos apresentam uma baixa qualidade na tradução de unidades lexicais multipalavra e que este fenómeno linguístico continua a representar um desafio significativo para a tradução automática, independentemente do tipo de paradigma. Na mesma linha de pensamento, a avaliação quantitativa e qualitativa anteriormente realizada (Barreiro et al., 2014b) das traduções de construções com verbos-suporte pelos sistemas OpenLogos e Google Translate reforça a necessidade de criação de sistemas híbridos que tirem partido da robustez dos sistemas por regras para melhorar a tradução automática de unidades lexicais multipalavra. A proposta consiste na hibridização por via da integração de conhecimento semântico-sintático nos sistemas estatísticos.

O sistema OpenLogos foi o escolhido por oferecer duas importantes vantagens em relação a outros sistemas baseados em regras. Uma diz respeito às regras não serem dependentes de algoritmos (ou guiadas por meta-regras), libertando o sistema da saturação da lógica quando confrontado com um problema tão complexo como o da tradução. A outra, está relacionada com a representação simbólica da língua natural, que no sistema OpenLogos, é transposta para um nível mais abstrato do que as palavras. Normalmente, os sistemas algorítmicos tendem a estar limitados à capacidade do algoritmo, e qualquer sequência lógica deixa de funcionar a determinada altura ao processar a língua natural. Não tendo a limitação do algoritmo de supervisão, o OpenLogos tem capacidade ilimitada para melhorar a tradução e as melhorias são fácil e imediatamente implementáveis. Por outro lado, as regras do sistema OpenLogos são baseadas em padrões de língua natural organizadas numa taxonomia de segunda ordem, chamada linguagem de abstração semântico-sintática, de ora em diante, SAL[†], descrita em Barreiro et al. (2011) e no Tutorial SAL[‡]. Estas duas importantes características do sistema posicionam o OpenLogos num patamar semelhante, em espírito, ao da tradução automática estatística e, em relação a esta, lhe concedem a vantagem de não sofrer do problema típico de escassez dos dados de treino, possibilitando o processamento de texto e a sua tradução. O analisador gramatical do sistema permite gerar árvores em diferentes níveis de análise, como descrito em Barreiro et al. (2011). Em cada nível de análise, as sequências de língua natural (palavras ou expressões) são representados por unidades SAL, que podem substituir os elementos comuns de mapeamento dos sistemas estatísticos para a finalidade da tradução. Com esta representação do conhecimento linguístico, os n-gramas evoluem de palavras e expressões para sequências de elementos SAL (ou seja, sequências de palavras ou expressões com propriedades semântico-sintáticas), permitindo que este mapeamento ocorra a um nível mais abstrato. Esta técnica conduz a um aumento da capacidade de mapeamento de sinónimos e expressões semanticamente equivalentes, aumentando a capacidade de encontrar paráfrases e traduções mais adequadas. Em consonância com SAL, o sistema usa regras semântico-sintáticas (designadas por SEMTAB) para realizar transformações monolíngues e multilíngues que funcionam com elevada eficácia na tradução das áreas mais frágeis dos sistemas estatísticos, como a tradução de todo o tipo de homógrafos (Barreiro et al., 2005), de verbos com traduções diferentes dependendo dos seus argumentos (e.g. *to raise a child - criar/educar um(a) criança/filho*; *to raise awareness – consciencializar*; *to raise concerns - suscitar preocupação*; *to raise funds – angariar / arranjar / obter financiamentos/fundos*) e de unidades lexicais multipalavra (incluindo as unidades não adjacentes, ou descontínuas, tais como *was in no way related to - não estava de forma alguma relacionado com* ou *is falling far short of - está bem aquém de*), entre outros.

O modelo integra um módulo de parafraseamento que permite melhorar o mapeamento de termos e expressões semanticamente idênticos. Este modelo funciona independentemente da quantidade e qualidade de corpos disponível e assenta numa metodologia repetível, que pode ser usada em diferentes aplicações de processamento de linguagem natural e tarefas multilíngues. O módulo de parafraseamento do modelo híbrido ajuda a tradução automática, permitindo o mapeamento de unidades lexicais multipalavra com palavras ou expressões com o mesmo significado. Por exemplo, as construções com verbos-suporte podem ser transformadas em verbos (*fazer a apresentação de - apresentar* ou *dar um abraço a - abraçar*), porque os nomes predicativos *apresentação* e *abraço* estão semanticamente relacionados com os verbos *apresentar* e *abraçar* não apenas a um nível abstrato (SAL) independente da categoria gramatical, mas também ao nível morfossintático, através de regras de derivação.

Em suma, o modelo híbrido em desenvolvimento aplica conhecimento linguístico na análise e resolução de ambiguidades por meio de regras e técnicas estatísticas de mapeamento de unidades linguísticas em vez de n-gramas, ou seja, palavras ou sequências de palavras, onde a informação semânti-

[†] SAL é o acrónimo de Semantico-Syntactic Abstraction Language.

[‡] O Tutorial SAL está disponível no seguinte endereço da internet: http://www.l2f.inesc-id.pt/~abarreiro/openlogos-tutorial/new_A2menu.htm

```
<Entry source="pipe" target="tubo">
   <source head_word="1" homograph="no" word_type="01">
     <pos description="Noun" wclass="01"/>
     <morphology num_id="1" number="singular">
       <inflection description="like book, books" example="book" id="16"/>
     </morphology>
     <sal code="3,34,745" mnemonic="COcond" set="functional" subset="conduit"
superset="concrete"/>
   </source>
   <target head_word="1" word_type="01">
     <pos description="Noun" wclass="01"/>
     <morphology gen_id="1" gender="masculine" num_id="1" number="singular">
       <inflection description="plural adding -s" example="tinteiro" gender_code="1" id="99"/>
     </morphology>
   </target>
</Entry>
```

Figura 1 - Entrada lexical para o nome concreto, funcional, objeto condutor: EN *pipe* - PT *tubo*.

co-sintática está disponível em cada etapa da análise da frase. A técnica permite respeitar a representação científica da linguagem, como, por exemplo, a unicidade de unidades lexicais multipalavra, para as quais os atuais sistemas de alinhamento não apresentam uma solução científica e tecnicamente viável.

3.3.1 Dicionário Inglês-Português

Os dicionários bilíngues representam um recurso importante na tradução automática, e quanto mais conhecimento envolverem, melhor poderão contribuir para a qualidade da tradução. O dicionário de inglês-português, tal como os restantes dicionários da OpenLogos apresentados em Barreiro et al. (2014a), contém conhecimento semântico-sintático e conhecimento ontológico (SAL), desenvolvido ao longo de várias décadas pela equipa de linguistas da Logos. Estes dicionários integravam o antigo produto comercial de tradução automática LogosMT, agora disponível em código aberto no sistema OpenLogos[§]. Nos dicionários do OpenLogos existem mais de 1.000 categorias distribuídas por quatro níveis de abstração: o nível sintático (categoria gramatical) e três níveis de conceito abstrato SAL, designados de superconjunto (superset), conjunto (set) e subconjunto (subset). Essas categorias compreendem tanto classificações gramaticais, tais como verbo bitransitivo, nome próprio, etc.), como classificações conceptuais (método, instrumento, etc.), que estão sistematicamente relacionadas entre si.

Os verbos intransitivos, por exemplo, estão classificados semanticamente em três conjuntos: os *existenciais*, os *operacionais* e os *de movimento*, cada um destes, por sua vez, subdivididos em vários subconjuntos, de acordo com as suas propriedades sintáticas. Por outro lado, os substantivos *funcionais*, um conjunto do superconjunto *concreto*, subdivide-se em vários subconjuntos, tais como *ferramentas/dispositivos*, *objetos em tecido*, *objetos condutores*, entre outros.

Categoria	id	Frequência
Substantivo	1	28270
Verbo	2	34985
Advérbio (locativo)	3	458
Adjetivo	4	24503
Pronome	5	121
Advérbio (modo, grau, etc.)	6	2189
Preposição (não locativa)	11	140
Auxiliar/modal	12	34
Preposição (locativa)	13	148
Artigo definido	14	194
Artigo indefinido	15	66
Número em aposição	16	208
Negação	17	2
Pronome relativo e interrogativo	18	23
Conjunção	19	160
Pontuação	20	30
Total		91531

Tabela 2 - Total de entradas por categoria.

A Figura 1 apresenta a entrada lexical EN *pipe* – PT *tubo*, classificada como um nome concreto, funcional, objeto condutor. Para além da classificação SAL, a entrada lexical tem informação morfológica para a palavra em inglês e em português. *Pipe* segue o paradigma flexional #16 para substantivos que formam o plural em *-s*, como *book*. *Tubo* segue o paradigma flexional #99 para substantivos masculinos que formam o plural em *-s*, como *tinteiro*. Outro tipo de informação pode ser extraída dos dicionários, mas, por motivos de simplicidade, não será ilustrada neste artigo. A Tabela 2 apresenta o número de entradas por categoria gramatical. Estes recursos linguísticos estão disponíveis para integração em aplicações de processamento de linguagem natural, incluindo a tradução automática 'de' e 'para' português e serão usados no sistema híbrido em desenvolvimento.

[§] http://logos-os.dfki.de, openlogos-mt.sourceforge.net

4 Conclusão e Trabalho Futuro

A tradução automática veio para revolucionar a comunicação no mundo, permitindo o acesso à informação em várias línguas estar ao alcance de um simples clique. Resta ainda um trabalho significativo para melhorar a qualidade linguística dos atuais sistemas, para que esta tecnologia alcance a sua plenitude. Neste artigo, descrevemos o trabalho de investigação em tradução automática desenvolvido no INESC-ID. Este trabalho começa a revelar sólidos progressos em várias frentes, nomeadamente na tradução de fala para fala, na tradução de microblogues e na evolução da tecnologia de tradução automática híbrida com qualidade melhorada. Os esforços futuros continuarão a incidir sobre o desenvolvimento e aperfeiçoamento do modelo híbrido que visa ser linguisticamente mais avançado e tecnicamente mais rápido e fácil de explorar para outras aplicações. Integrará um módulo parafrástico resultante do projeto eSPERTo[**] (Sistema de Parafraseamento para Edição e Revisão de texto) que visa servir de ferramenta de pré-edição e de auxílio à tradução. Para além disso, o sistema híbrido ambiciona tirar proveito da computação em nuvem, de grandes volumes de dados e de técnicas de alinhamentos linguisticamente motivados, que contribuem, no seu conjunto, para um desenvolvimento mais fácil e rápido de novos pares de línguas. O modelo deverá também poder ser combinado com a tradução participativa ou colaborativa (*crowdsourcing*) que permite aumentar exponencialmente o volume dos corpos paralelos existentes para as várias línguas e posteriormente beneficiar do apoio de tradução coletiva especializada para aumentar a qualidade das traduções adquiridas colaborativamente. A língua portuguesa precisa de se manter viva e relevante para o mundo moderno, o que só acontecerá se se posicionar na linha da frente da tecnologia em tradução automática. Se os esforços forem canalizados nesse sentido e um maior investimento contemplar a melhoria da qualidade e quantidade dos recursos linguísticos orientados para a tradução, no futuro a tradução automática tornar-se-á a principal amiga da língua portuguesa.

Agradecimentos

Agradecemos aos organizadores do colóquio *Português, Lingua Global*, promovido pelo Centro de Estudos Lusíadas do Instituto de Letras e Ciências Humanas da Universidade do Minho, onde foi apresentada a comunicação *Contributos da Tecnologia da Língua na Globalização do Português*, e à sua respetiva audiência, a oportunidade de desenvolver o trabalho que esteve na base deste artigo.

O trabalho de Anabela Barreiro foi financiado pela FCT (bolsa de pós-doutoramento SFRH / BPD / 91446 / 2012). Este trabalho também contou com o apoio da FCT, através do projeto PEst-OE/EEI/LA0021/2013.

Referências

Barreiro, Anabela e Elisabete Ranchhod. 2005. Machine Translation Challenges for Portuguese, *Lingvisticæ Investigationes* 28:1, pp. 3-18. In Sylviane Cardey, Peter Greenfield, Séverine Vienney (eds), Machine Translation, Controlled Languages and Specialised Languages.

Barreiro, Anabela, Bernard Scott, Walter Kasper e Bernd Kiefer. 2011. OpenLogos Rule-Based Machine Translation: Philosophy, Model, Resources and Customization. *Machine Translation*, 25(2):107– 126.

Barreiro, Anabela, Johanna Monti, Brigitte Orliac e Fernando Batista. 2013. When Multiwords Go Bad in Machine Translation, in *Proceedings of the Workshop on Multi-word Units in Machine Translation and Translation Technology, Machine Translation Summit XIV*.

Barreiro, Anabela, Fernando Batista, Ricardo Ribeiro, Helena Moniz e Isabel Trancoso. 2014a. OpenLogos Semantico-Syntactic Knowledge-Rich Bilingual Dictionaries, in *Proceedings of the 9th edition of the LREC conference*.

Barreiro, Anabela, Johanna Monti, Brigitte Orliac, Susanne Preuss, Kutz Arrieta, Wang Ling, Fernando Batista e Isabel Trancoso. 2014b. Linguistic Evaluation of Support Verb Construction Translations by OpenLogos and Google Translate, in *Proceedings of the 9th edition of the LREC conference*.

Batista, Fernando, Helena Moniz, Isabel Trancoso e Nuno J. Mamede. 2012. *Bilingual Experiments on Automatic Recovery of Capitalization and Punctuation of Automatic Speech Transcripts.* IEEE Transactions on Audio, Speech and Language Processing, Special Issue on New Frontiers in Rich Transcription, 20(2):474-485.

Brown, Ralf D. 1996. Example-Based Machine Translation in the Pangloss System, in *COLING* vol. 1, pp. 169-174.

Callison-Burch, Chris. 2007. Paraphrasing and Translation, PhD Thesis, University of Edinburgh.

** https://esperto.l2f.inesc-id.pt

Chahuneau, Victor, Smith, Noah A. e Dyer, Chris. 2013. "Knowledge-Rich Morphological Priors for Bayesian Language Models.", in *HLT-NAACL* (ACL), pp. 1206-1215.

Dugast, Löıc, Senellart, Jean e Koehn, Philipp. 2007. Statistical Post-editing on SYSTRAN's Rule-based Translation System, in *Proceedings of the Second Workshop on Statistical Machine Translation* (Stroudsburg, PA, USA: ACL), pp. 220-223.

Eisele, Andreas, Christian Federmann, Hans Uszkoreit, Hervé Saint-Amand, Martin Kay, Michael Jellinghaus, Sabine Hunsicker, Teresa Herrmann e Yu Chen. 2008. Hybrid Machine Translation Architectures within and beyond the EuroMatrix project, in J. Hutchins and W.V. Hahn, ed., *Hybrid MT Methods in Practice: Their Use in Multilingual Extraction, Cross-Language Information Retrieval, Multilingual Summarization, and Applications in Hand-Held Devices. Proceedings of the European Machine Translation Conference* (HITEC e. V 2008), pp. 27-34.

Elming, Jakob. 2006. Transformation-based correction of rule-based MT., in *Proceedings of EAMT 2006*, pp. 219--226.

Grazina, Nuno. 2010. Tradução Automática de Fala. Tese de Mestrado. IST.

Heafield, Kenneth e Lavie, Alon. 2011. CMU System Combination in WMT 2011, in *Proceedings of the Sixth Workshop on Statistical Machine Translation* (Stroudsburg, PA, USA: ACL), pp. 145-151.

Koehn, Philipp e Kevin Knight. 2002. ChunkMT: Statistical Machine Translation with Richer Linguistic Knowledge.

Koehn, Philipp, Franz Josef Och e Daniel Marcu. 2003. Statistical Phrase-Based Translation, in *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology (HLT-NAACL '03)* (Morristown, NJ, USA: ACL), pp. 48-54.

Koehn, Philipp. 2005. Europarl: A parallel corpus for statistical machine translation. MT Summit.

Koehn, Philipp, Hoang, Hieu, Birch, Alexandra, Callison-Burch, Chris, Federico, Marcello, Bertoldi, Nicola, Cowan, Brooke, Shen, Wade, Moran, Christine, Zens, Richard, Dyer, Chris, Bojar, Ondřej, Constantin, Alexandra e Herbst, Evan. 2007. Moses: Open Source Toolkit for Statistical Machine Translation, in *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions* (Stroudsburg, PA, USA: ACL), pp. 177-180.

Labaka, Gorka, Stroppa, Nicolas, Way, Andy e Sarasola, Kepa. 2007. "Comparing rule-based and data-driven approaches to Spanish-to-Basque machine translation", in *Machine Translation Summit XI* (Copenhagen, Denmark).

Ling, Wang, Dyer, Chris, Black, Alan W. e Trancoso, Isabel. 2013a. Paraphrasing 4 Microblog Normalization, in *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing* (Seattle, Washington, USA: ACL), pp. 73-84.

Ling, Wang, Xiang, Guang, Dyer, Chris, Black, Alan e Trancoso, Isabel. 2013b. Microblogs as Parallel Corpora, in *Proceedings of the 51st Annual Meeting on Association for Computational Linguistics* (ACL).

Niessen, Sonja e Hermann Ney. 2004. Statistical Machine Translation with Scarce Resources Using Morpho-syntactic Information, *Computational Linguistics* 30, 2, pp. 181--204.

Nirenburg, Sergei, Harold Somers e Yorick Wilks. 2003. *Readings in Machine Translation* (Five Cambridge Center, Cambridge, MA: The MIT Press).

Parlikar, Alok, Alan W. Black e Stephan Vogel. 2010. Improving Speech Synthesis of Machine Translation Output, Interspeech (2010), Makuhari, Japan.

Resnik, Philip e Noah A. Smith. 2003. The Web as a parallel corpus. *Computational Linguistics* 29, 3 (September 2003), 349-380.

Sánchez-Martínez, Felipe, Forcada, Mikel L. e Way, Andy. 2009. Hybrid rule-based - example-based MT: feeding Apertium with sub-sentential translation units", in *EBMT 2009 - 3rd Workshop on Example-Based Machine Translation* (Dublin, Ireland: DORAS).

Scott, Bernard. 2003. The Logos Model: An Historical Perspective, *Machine Translation* 18, 1, pp. 1-72.

Shirai, Satoshi, Francis Bond e Yamato Takahashi. 1997. A Hybrid Rule and Example-based Method for Machine Translation, in *Recent Advances in Example-Based Machine Translation* (Kluwer Academic Publishers), pp. 211-224.

Simard, Michel, Ueffing, Nicola, Isabelle, Pierre e Kuhn, Roland. 2007. Rule-Based Translation with Statistical Phrase-Based Post-Editing", in *Proceedings of the Second Workshop on Statisti-*

cal Machine Translation (Prague, Czech Republic: ACL, pp. 203-206.

Terumasa, Ehara. 2007. Rule Based Machine Translation Combined with Statistical Post Editor for Japanese to English Patent Translation, in *Proceedings of the MT Summit XI Workshop on Patent Translation* vol. 11, pp. 13--18.

Ueffing, Nicola e Ney, Hermann. 2003. Using POS Information for Statistical Machine Translation into Morphologically Rich Languages, in *Proceedings of the Tenth Conference on European Chapter of the Association for Computational Linguistics - Volume 1* (Stroudsburg, PA, USA: ACL), pp. 347--354.

Xu, Wei, Ritter, Alan, Callison-Burch, Chris, Dolan, William, e Ji, Yangfeng. 2014. Extracting Lexically Divergent Paraphrases from Twitter. *Transactions Of The Association For Computational Linguistics, 2*, 435-448.