DOI: 10.14393/DL32-v11n5a2017-21

Tradução Automática: estratégias e limitações

Machine Translation: approaches and limitations

Helena de Medeiros Caseli*

RESUMO: A Tradução Automática é uma das principais subáreas e aplicações do Processamento Automático de Línguas Naturais (PLN). Em um sistema de tradução automática, a informação em uma língua fonte, fornecida como entrada para o sistema, é transformada em uma versão equivalente na língua alvo. Apesar de mais de 70 anos de pesquisas em tradução automática, as principais estratégias propostas apresentam limitações. Neste artigo, são discutidas três dessas estratégias: a tradução automática baseada em regras, a tradução automática estatística e a tradução automática neural. Neste artigo, apresentamos uma breve descrição de cada estratégia, acompanhada de exemplos que ajudam a compreender as limitações citadas.

PALAVRAS-CHAVE: Tradução automática. Processamento Automático de Línguas Naturais. Tradução automática baseada em regras. Tradução automática estatística. Tradução automática neural.

ABSTRACT: Machine Translation is one of the main fields and applications of Computational Linguistics (CL). In a machine translation system, the information in a source language, provided as input to the system, is transformed into an equivalent version in the target language. Despite more than 70 years of researches regarding machine translation field, the main approaches proposed have limitations. In this paper, we discuss three of these approaches: rule-based machine translation, statistical machine translation, and neural machine translation. In this article, we present a brief description of each approach, accompanied by examples that help to understand the limitations mentioned.

KEYWORDS: Machine Translation. Computational Linguistics. Rule-based machine translation. Statistical machine translation. Neural machine translation.

1. Introdução

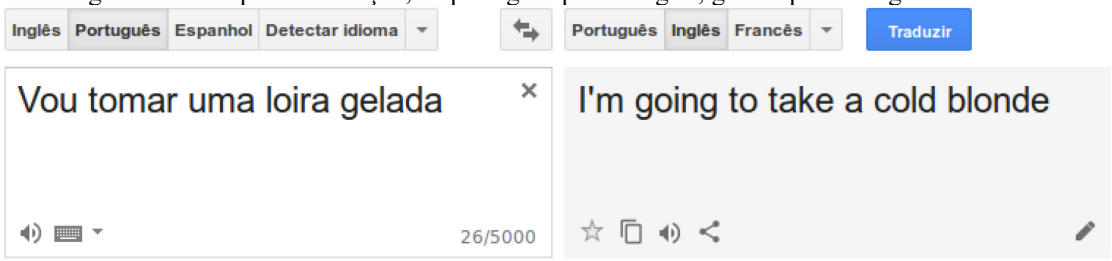
A Tradução Automática (TA) é tanto uma das principais subáreas do Processamento Automático de Línguas Naturais (PLN) como uma aplicação computacional disponível em sites[1] e sistemas/aplicativos para computadores e dispositivos móveis (celulares, *tablets*, etc.). Na TA, a partir da informação fornecida como entrada em um idioma original (língua fonte), um site/sistema/aplicativo gera uma versão equivalente em outro idioma (língua alvo).

* Doutora em Ciência da Computação, Departamento de Computação, Universidade Federal de São Carlos (UFSCar). E-mail: helenacaseli@dc.ufscar.br.

[1] Como o Google Translator. Disponível em: https://translate.google.com.br/. Acesso em: 25 jan. 2017.

Embora tenha surgido como uma área de pesquisa há mais de 70 anos, a TA ainda apresenta desafios, como tratar gírias e coloquialismos como ilustra o exemplo da Figura 1. Nesse exemplo, a tradução da expressão em português "loira gelada", tradicionalmente empregada para se referir a uma "cerveja gelada", foi equivocadamente traduzida em inglês para *cold blonde* pelo Google Translator. Esse exemplo demonstra uma das principais limitações dos sistemas automáticos de TA, que é lidar com gírias e coloquialismos. Essas limitações, aliadas à importância cada vez maior dos sistemas de TA em um cenário mundial com cada vez menos barreiras físicas e virtuais entre as pessoas falantes de diversos idiomas, fomentam as pesquisas atuais em TA.

Figura 1 - Exemplo de tradução, de português para o inglês, gerada pelo Google Translator.



Fonte: extraída do site do Google Translator (em 26/01/2017).

Desde seu surgimento, diversas estratégias de TA foram propostas. As estratégias mais tradicionais são: a tradução direta, a tradução por transferência e a tradução por interlíngua. Na tradução direta, ocorre o mapeamento direto das unidades lexicais fonte para as unidades lexicais alvo, ou seja, sem que nenhuma etapa de análise sintática ou semântica seja realizada. Na tradução por transferência, por sua vez, há uma análise sintática parcial ou completa da língua fonte e o mapeamento fonte-alvo se dá com base em regras de transferência sintática seguido da geração de uma saída equivalente na língua alvo. Por fim, na tradução por interlíngua, há o mapeamento completo da língua fonte para uma língua intermediária (representação abstrata do significado) e desta para a língua alvo (CASELI, 2015).

Além de caracterizar as pesquisas (ou os sistemas) de TA de acordo com a estratégia, também há a diferenciação de acordo com o paradigma empregado. O paradigma linguístico foi o mais investigado e empregado nos primórdios das pesquisas em TA, até a década de 1980 quando o paradigma empírico passou a ser o mais utilizado, até os dias de hoje. No paradigma linguístico, o conhecimento linguístico profundo das línguas fonte e alvo é mapeado,

geralmente na forma de regras, no que se conhece como TA baseada em regras (*Rule-based Machine Translation* ou RBMT).[2]

Contudo, a partir de 1989, o paradigma empírico (também conhecido como baseado em *corpus*) passou a dominar o cenário das pesquisas em TA (HUTCHINS, 2007), seguindo uma tendência das pesquisas em tradução de modo geral, as quais foram influenciadas pela linguística de corpus. No paradigma empírico, os sistemas de TA aprendem como gerar a sentença alvo equivalente à sentença fonte de entrada com base em um conjunto de treinamento, no caso, um *corpus* paralelo bilíngue. Um *corpus* paralelo bilíngue é uma coleção de pares de textos escritos em duas línguas sendo um texto a tradução do outro. Um exemplo de par de textos paralelos escritos em português e em inglês pode ser visto no Quadro 1.

Quadro 1- Exemplo de um par de textos paralelos em português e em inglês.

Texto fonte (em português)	**Texto alvo (em inglês)**
Watson aprende português Aprender espanhol, português e japonês é a nova etapa da plataforma Watson, o sistema de computação cognitiva que a IBM lançou em 2011. Ele tem habilidade para interagir na linguagem do usuário, com voz, processar grandes quantidades de dados, aprender e adquirir conhecimento conforme é usado. "Para isso, é preciso uma adaptação a uma língua com vocabulários e regras semânticas", diz Fábio Gandour, cientista-chefe do Laboratório de Pesquisas da IBM Brasil. O Watson vai ser alimentado com mais de 300 mil palavras, além de ser dotado de um processamento que inclui o significado de cada palavra. "O sistema não é um produto pronto que a pessoa compra e instala no computador ou servidor, ele precisa ser alimentado com informações para que possa dar respostas adequadas a cada usuário.	**Watson learns Portuguese** Learning Spanish, Portuguese, and Japanese is the latest stage in the development of Watson, the cognitive computing system launched by IBM in 2011. The Watson platform is able to interact vocally with users in their own language, process large amounts of data, and acquire new knowledge on the fly. "This requires adaptation to the language, including vocabulary and semantic rules," says Fábio Gandour, chief scientist at the Brazil Research Lab at IBM Brazil. Watson will be "fed" over 300,000 words, as well as being given a processing feature that includes the meaning of each word. "The system is not a finished product that a person can buy and install on a computer or server, it needs to be fed information so that it can produce appropriate responses for each user.

Fonte: extraída do site da revista Pesquisa FAPESP (revistapesquisa.fapesp.br), edição 225 de novembro de 2014.[3]

[2] Um exemplo de um sistema de tradução automática baseada em regras é o Apertium (ARMENTANO-OLLER et al., 2006). Disponível em: http://www.apertium.org. Acesso em: 25 jan. 2017. Exemplos de regras presentes no sistema Apertium são apresentados no Quadro 2.

[3] O texto na íntegra pode ser consultado em: http://revistapesquisa.fapesp.br/2014/11/18/watson-aprende-portugues/ e http://revistapesquisa.fapesp.br/en/2014/11/23/watson-learns-portuguese/. Acesso em: 27 jan. 2016.

A TA empírica engloba estratégias como a baseada em exemplos (*Example-Based Machine Translation* ou EBMT)[4], a estatística (*Statistical Machine Translation* ou SMT)[5] e, mais recentemente, a neural (*Neural Machine Translation* ou NMT). Na TA baseada em exemplos, o aprendizado tem como base o reconhecimento de padrões recorrentes no *corpus* de treinamento, enquanto na TA estatística são as probabilidades de tradução (de palavras ou de frases[6]), calculadas com base no *corpus* de treinamento, que definem como a tradução alvo será gerada (CASELI, 2007). A TA estatística era considerada o estado da arte (e o paradigma usado nos tradutores do Google) até 2016, quando deu lugar à estratégia neural.[7] Na TA neural, redes neurais artificiais[8] são construídas com base nas características importantes nos dados de treinamento (como a morfologia das palavras, suas frequências e contextos de ocorrência, entre outros) e essas características são o que norteiam o mapeamento para gerar a saída apropriada na língua alvo.

2. Tradução automática na prática

Essa seção ilustra a utilização de três estratégias de TA na prática: TA baseada em regras, TA estatística baseada em frases e TA neural.

2.1. Tradução automática baseada em regras

Na TA baseada em regras, como já mencionado anteriormente, o conhecimento linguístico é mapeado na forma de regras, como as ilustradas no Quadro 2. Essas regras morfossintáticas geralmente são criadas manualmente por linguistas especialistas nas línguas fonte e alvo. Embora essa abordagem manual de criação de regras ofereça grande poder ao

[4] Um exemplo de um sistema de tradução automática baseada em exemplos é o Pangloss (BROWN, 1996).
[5] Exemplos de sistemas de tradução automática estatística baseada em frases (*Phrase-Based Statistical Machine Translation* ou PBSMT) são os propostos em (KOEHN et al., 2003) e em (OCH; NEY, 2004).
[6] Na SMT, duas ou mais palavras em sequência formam uma *phrase* (traduzida como frase para o português). Uma *phrase* (do inglês) recebe este nome mesmo que não seja um sintagma ou desempenhe qualquer papel sintático na sentença.
[7] Disponível em: https://mobile.nytimes.com/2016/12/14/magazine/the-great-ai-awakening.html. Acesso em: 26 jan. 2017.
[8] Redes neurais artificiais são técnicas computacionais, usadas na área de Aprendizado de Máquina, as quais estão baseadas no modelo neural de organismos inteligentes, ou seja, em uma rede neural artificial uma unidade de processamento (um neurônio) recebe entradas, as processa e propaga a saída para outras unidades de processamento (outros neurônios) organizadas em camadas. Após os processamentos nas diversas camadas, um resultado é produzido na camada final e dado como resposta.

projetista do sistema, uma vez que lhe permite explicitar o tratamento de casos específicos da tradução, essa é também a principal desvantagem de um sistema RBMT, por demandar um grande trabalho para a criação dessas regras. Para tentar contornar essa desvantagem, foram propostos métodos para gerar regras automaticamente (GALLEY et al., 2004, GÜVENIR; CICEKLI, 1998, CASELI et al., 2006, CASELI, 2007).

Quadro 2 - Exemplos de conhecimento linguístico mono e bilíngue mapeado em um sistema RBMT.[9]

Regras monolíngues (português)	**Regras bilíngues (espanhol-português)**
alunas:aluno<n><f><pl>	alumno<n>:aluno<n>
aluna:aluno<n><f><sg>	exterior<n>:exterior<n>
alunos:aluno<n><m><pl>	exterior<adj>:exterior<adj>
aluno:aluno<n><m><sg>	mandar<vblex>:mandar<vblex>

Fonte: adaptado de http://wiki.apertium.org/wiki/Bilingual_dictionary.

A partir da definição das regras mono e bilíngues – além de outras que especificam, por exemplo, o reordenamento das palavras – um sistema RBMT pode processar uma sentença fonte de entrada e gerar uma sentença alvo equivalente. Por exemplo, no caso do sistema Apertium, dada a entrada ilustrada no Quadro 3-a, o sistema realiza uma análise seguida de uma desambiguação (etiquetação) morfossintática gerando a versão apresentada no Quadro 3-b com base em regras como as do Quadro 2. Por exemplo, para a palavra "alunos" utiliza-se a regra monolíngue especificada no Quadro 2, a qual etiqueta "alunos" como "aluno<n><m><pl>", ou seja, a forma base "aluno" seguida das etiquetas: nome (n), masculino (m) e plural (pl).

Por fim, após a consulta ao dicionário bilíngue e às regras de reordenamento e geração na língua alvo, a saída produzida é apresentada no Quadro 3-c. Veja que para a geração da saída traduzida (no Quadro 3-c) são utilizadas regras bilíngues como as apresentadas no Quadro 2. Por exemplo, para "aluno<n>" verifica-se a correspondência com "alumno<n>" nas regras bilíngues do Quadro 2 e essa correspondência é utilizada na geração da sentença de saída no Quadro 3-c. Regras para flexão das palavras específicas em cada língua alvo são aplicadas para garantir a concordância correta da sentença completa gerada como saída pelo tradutor.

[9] Cada regra monolíngue especifica a forma superficial de uma palavra, seu lema e categoria gramatical (n: substantivo, adj: adjetivo, vblex, verbo, etc.), seguidos dos atributos morfológicos (f, m: gênero, sg, pl: número, etc.). Cada regra bilíngue especifica a forma superficial e categoria gramatical fonte associada (separada pelo ':') à forma superficial e categoria gramatical alvo. As regras bilíngues podem ser processadas tanto na direção fonte-alvo quanto na direção alvo-fonte.

Quadro 3 - Exemplo de sentença fonte (em português) de entrada (a), após a etiquetação morfossintática (b)[10] e a sentença alvo (em espanhol) correspondente gerada pelo sistema Apertium (c).

a	Mandamos alunos para o exterior rápido demais, diz presidente da Capes
b	^Mandamos/~~Mandar<vblex><ifi><p1><pl>~~/Mandar<vblex><pri><p1><pl>$ ^alunos/aluno<n><m><pl>$ ^para/para<pr>$ ^o/~~o<detnt>~~/o<prn><pro><p3><nt>/o<det><def><m><sg>/~~o<prn><pro><p3><m><sg>~~$ ^exterior/~~exterior<adj><mf><sg>~~/exterior<n><m><sg>$ ^rápido/rápido<adj><m><sg>$ ^demais/~~demais<adv>/demais<adj><mf><sp>~~/demais<n><mf><pl>$^,/,<cm>$ ^diz/dizer<vblex><pri><p3><sg>$ ^presidente/presidente<n><m><sg>$ ^da/de<pr>+o<det><def><f><sg>$ ^Capes/Capar<vblex><prs><p2><sg>$
c	Mandamos alumnos para el exterior rápido demás, dice presidente de la Capes

Fonte: elaborado pelo próprio autor.

Assim, a TA baseada em regras tem como vantagens o fato dos recursos linguísticos serem legíveis por humanos e do processamento (tradução de novas sentenças) ser muito simples. Como limitações principais desta estratégia de TA estão o alto custo no desenvolvimento dos recursos linguísticos e na sua extensão para outro idioma, e a cobertura lexical limitada, uma vez que não é possível traduzir uma palavra (ou construção gramatical) que não esteja presente no conjunto de regras.

2.2. Tradução automática estatística baseada em frases

A TA estatística, por sua vez, geralmente é definida como um processo de geração[11] de uma sentença na língua alvo que maximiza um modelo da adequação e da fluência da sentença alvo esperada. A adequação modela a equivalência de uma sentença alvo em relação à sentença fonte de entrada, ou seja, quanto ela preserva do significado originalmente codificado na sentença fonte. Enquanto a fluência modela quão natural (fluente) a sentença alvo gerada é no idioma alvo, ou seja, está relacionada à correção gramatical e à naturalidade das palavras e construções presentes na sentença alvo (CASELI, 2015). A adequação é modelada por um modelo de tradução gerado com base nas frequências de coocorrência de palavras fonte e suas correspondentes palavras alvo no *corpus* de treinamento. A fluência, por sua vez, é modelada

[10] Para gerar a saída apresentada no Quadro 2-b dois comandos foram executados: (1) lt-proc -a pt-es.automorf.bin (para atribuir, à cada palavra fonte todas as possíveis categorias gramaticais e respectivos atributos morfológicos) e (2) apertium-tagger -g pt-es.prob (para desambiguar as categorias deixando apenas aquela que é a mais provável, dado o contexto de ocorrência da palavra na sentença). O arquivo pt-es.automorf.bin é obtido a partir do arquivo de regras monolíngues do Apertium (ex: apertium-es-pt.pt.dix), como o ilustrado no Quadro 1, por meio da execução do comando: lt-comp lr apertium-es-pt.pt.dix pt-es.automorf.bin

[11] Além dos modelos gerativos existem também os modelos discriminativos. Para saber mais sobre esses modelos sugere-se a leitura do trabalho de Lopez (2008).

por um modelo de língua gerado com base nas frequências de coocorrência de palavras alvo no *corpus* de treinamento. Os modelos de língua e de tradução podem ser gerados com base apenas em palavras ou em sequências de palavras (frases ou, em inglês, *phrases*). O mais utilizado é o modelo de TA baseado em frases (*Phrase-Based Statistical Machine Translation*, ou PBSMT) de tamanho variável, mas que geralmente chegam a 5 ou 7 *tokens*[12]. O objetivo da TA estatística é gerar uma sentença alvo que maximize a adequação e a fluência.

Para ilustrar o processo de TA estatística baseada em frases, considere o trecho do *corpus* paralelo português-inglês da FAPESP (AZIZ; SPECIA, 2011)[13] apresentado no Quadro 4.

Quadro 4 - Sentenças fonte (em português) presentes no *corpus* FAPESP usadas no treinamento de um modelo de língua (em português).

***Corpus* monolíngue em português**
Para o diretor científico da FAPESP, as perspectivas que se abrem com o depósito da patente do Evasin "são um início auspicioso" para o Cepid.
O aumento da escala de produção, de novos materiais e a necessidade de geração de maiores níveis de eletricidade sem utilizar os combustíveis fósseis abrem um largo caminho para a energia solar.
Ainda sem aplicação clínica, esses achados abrem novos caminhos para se compreender como o cérebro se forma e como surgem certas doenças neurológicas, além da obesidade.
Como avalia o atual patamar de cooperação entre Brasil e Argentina e as possibilidades que se abrem?
Franchini acredita que a D5 ou outra molécula equivalente, se agirem apenas nos fibroblastos e passarem nos testes seguintes de eficácia e toxicidade, abrem perspectivas de uso futuro para eliminar a fibrose em doenças como cirrose hepática e pulmonar e esquistossomose.
Mas ele não é impermeável, existem frestas, e é necessário estar de olho nas oportunidades que se abrem.
Grupos de pesquisa abrem perspectivas para prevenção e tratamento mais eficaz
Estudos na Unicamp abrem caminho para a produção de isoflavonas extraídas da soja e de novos usos da própolis

Fonte: adaptado de Caseli (2015).

Na geração do modelo de língua considera-se, por exemplo, que a probabilidade de geração da frase “grupos de pesquisa abrem perspectivas” é maior do que a probabilidade de geração da frase “grupos de investigação abrem perspectivas”, uma vez que a probabilidade de

[12] *Tokens* é o nome que se dá, na computação, a uma sequência de caracteres delimitada por espaços. No contexto da TA um *token* pode ser uma palavra, um caractere, um símbolo de pontuação ou um número.
[13] Disponível em: http://www.nilc.icmc.usp.br/nilc/tools/Fapesp%20Corpora.htm. Acesso em: 26 jan. 2017.

"grupos de pesquisa" é maior do que a probabilidade de "grupos de investigação", no *corpus* FAPESP e na língua portuguesa falada no Brasil, de modo geral. Além disso, a probabilidade de geração da frase "grupos de pesquisa abrem perspectivas" é maior do que a probabilidade de geração da frase "de pesquisa grupos perspectivas abrem", por exemplo, já que a frase "de pesquisa grupos perspectivas abrem" provavelmente é muito rara!

A Figura 2 traz um trecho do modelo de língua aprendido considerando-se frases de tamanho máximo igual a 5 no *corpus* em português da FAPESP.[14] Exemplos de frases de tamanho 1 (1-grama) desse modelo são "abrem" e "grupos", enquanto exemplos de frases de tamanho 2 (2-grama) são "abrem perspectivas" e "grupos de", e assim por diante até os exemplos de frases de tamanho 5 (5-grama) como "prevenção e tratamento de doenças".

Figura 2 - Trecho do modelo de língua (em português) aprendido a partir do *corpus* FAPESP.

Modelo de língua em português

```
\data\
ngram 1=123825
ngram 2=1323945
ngram 3=549676
ngram 4=354401
Ngram 5=211555

\1-grams:
-3.664527   !                   -1.068139
-2.112817   "                   -0.6644497
-99         <s>                 -1.2639
-2.200827   </s>
-4.635522   abrem               -0.2803798
-1.707354   e                   -0.6711801
-3.747774   grupos              -0.4765508
-4.766913   prevenção           -0.4181228
-2.493415   se                  -0.4916143
-4.136694   tratamento          -0.3972652
…
\2-grams:
-2.074136   ! !                 -0.07787027
-0.6743945  ! "                 -0.5657353
-4.808689   <s> prevenção       -0.0777496
-1.219755   abrem perspectivas  -0.1640183
-0.7952127  grupos de           -0.341452
-3.501996   e tratamento        -0.2975546
-3.934279   pesquisa abrem
-1.01534    prevenção e         -0.1688237
-3.331075   se abrem            -0.1299266
-2.198522   tratamento </s>
...
```

```
\3-grams:
-0.766511   ! ! !
-0.4400599  abrem perspectivas para
-0.7030438  grupos de pesquisa      -0.3448678
-1.45863    e tratamento </s>
-0.8142314  prevenção e tratamento  -0.139513
-1.108147   tratamento mais eficaz
…
\4-grams:
-0.1434542  é possível ! </s>
-0.1847344  , sim ! </s>
-0.3916683  para prevenção e tratamento
…
\5-grams:
-0.1665249  <s> nossa Senhora ! </s>
-0.1665249  brasileiro e carioca ! </s>
-0.5688003  prevenção e tratamento de doenças
…
```

Fonte: extraído de Caseli (2015).

A geração do modelo de tradução segue uma estratégia similar de cálculo de probabilidades, mas agora realizada com base em um *corpus* paralelo bilíngue e não apenas em

[14] Detalhes sobre a geração desse modelo de língua para o português estão descritos e podem ser consultados no trabalho de Caseli (2015).

um *corpus* monolíngue. Assim, para o *corpus* paralelo inglês-português, conforme trecho ilustrado no Quadro 5, o trecho de um modelo de tradução aprendido é apresentado na Figura 3. Por exemplo, como pode ser visto na Figura 3, a probabilidade de traduzir "groups" para "grupos" é de cerca de 87%, enquanto probabilidade de tradução de "research groups" para "grupos de pesquisa" é de aproximadamente 76%. Quanto maior a frase (n-grama) considerada no modelo, menor será sua probabilidade, pois menos frequente ela é no *corpus* de treinamento. Por exemplo, a probabilidade de tradução da frase de tamanho 4 "a more efficient method" para "mais eficaz" é de apenas 0,07%.

Quadro 5 - Sentenças paralelas inglês-português presentes no *corpus* FAPESP usadas no treinamento de um modelo de tradução.

***Corpus* paralelo inglês-português**	
For the Scientific Director of FAPESP, the perspectives that open up with the registering of the patent of Evasin "are an auspicious beginning" for the Cepids.	Para o diretor científico da FAPESP, as perspectivas que se abrem com o depósito da patente do Evasin "são um início auspicioso" para o Cepid.
Research groups open up prospects for a more efficient method of prevention and treatment	Grupos de pesquisa abrem perspectivas para prevenção e tratamento mais eficaz
Ways open up for strengthening actions on gender in the social sphere", says Maria Cecilia Comegno, from Seade, who is the project's coordinator.	Abrem-se caminhos para o fortalecimento de ações de gênero em âmbito social", diz Maria Cecilia Comegno, do Seade, coordenadora do projeto.
As more sodium enters, more pores open up, and the nerve impulse propagates in a single direction, like a wave, until, in thousandths of a second, it hits the extremity of the neuron, releasing chemical messengers, called neurotransmitters, which pass the information on to the next cell.	À medida que entram mais sódio, mais poros se abrem e o impulso nervoso se propaga num único sentido como uma onda até atingir, em milésimos de segundo, a extremidade do neurônio, liberando mensageiros químicos chamados de neurotransmissores, que passam a informação para a célula seguinte.
This is the case of Camarea hirsuta , with its rounded yellow petals, Passiflora clathrata , a plant that is a relative of the passionflower, with its purple leaves, hidden amongst the bushes, and a species with a whitish flower, Alophia sellowiana , whose petals only open up at night.	É o caso da Camarea hirsuta , com suas pétalas redondas e amareladas, da Passiflora clathrata , uma planta aparentada do maracujá com flores violeta, escondida entre os arbustos, e de uma espécie com uma flor esbranquiçada, a Alophia sellowiana , cujas pétalas só se abrem à noite.

Fonte: adaptado de Caseli (2015).

Figura 3 - Trecho do modelo de tradução inglês-português aprendido a partir do *corpus* FAPESP.[15]

Tabela de tradução de frases inglês-português

```
groups ||| grupos ||| 0.875878 0.936047 0.743539 0.889503 2.718 ||| ||| 1281 1509
groups of ||| grupos ||| 0.00156011 0.0461955 0.0102487 0.889503 2.718 ||| ||| 1281 195
groups of ||| grupos de ||| 0.355353 0.388355 0.8 0.388285 2.718 ||| ||| 439 195
research groups ||| grupos de pesquisa ||| 0.768786 0.82872 0.738889 0.189696 2.718 ||| ||| 173 180
open up ||| abrem ||| 0.180797 0.112345 0.121571 0.0258495 2.718 ||| ||| 39 58
prospects ||| perspectivas ||| 0.577889 0.661017 0.646067 0.735849 2.718 ||| ||| 199 178
prospects for ||| perspectivas para ||| 0.62963 0.234406 0.191011 0.280476 2.718 ||| ||| 27 89
a more efficient method ||| mais eficaz ||| 0.000725015 5.57683e-07 0.0311756 0.10198 2.718 ||| ||| 86 2
…
the conclusions ||| as conclusões ||| 0.745098 0.632046 0.655172 0.0399798 2.718 ||| ||| 51 58
the possibility of ||| a possibilidade de ||| 0.708543 0.190833 0.661972 0.0675447 2.718 ||| ||| 398 426
indicating ||| indicar ||| 0.141509 0.132743 0.137615 0.141509 2.718 ||| ||| 106 109
the ||| os ||| 0.835612 0.71112 0.112016 0.0675026 2.718 ||| ||| 27277 203480
most suitable ||| mais adequado ||| 0.277778 0.0420202 0.344828 0.138171 2.718 ||| ||| 36 29
most suitable ||| mais adequados ||| 0.124906 0.039875 0.0689138 0.0614095 2.718 ||| ||| 16 29
exercises ||| exercícios ||| 0.595238 0.574468 0.704225 0.72973 2.718 ||| ||| 84 71
for ||| para ||| 0.36472 0.354614 0.4057 0.38116 2.718 ||| ||| 39422 35440
specific ||| certas ||| 0.0327663 0.0388889 0.00670135 0.0078918 2.718 ||| ||| 181 885
specific ||| certos ||| 0.00550579 0.0116959 0.0010825 0.0022548 2.718 ||| ||| 174 885
specific ||| certo ||| 0.00011611 0.001385 7.04534e-05 0.0011274 2.718 ||| ||| 537 885
diseases , ||| doenças , ||| 0.662651 0.618291 0.578947 0.788998 2.718 ||| ||| 83 95
something ||| algo ||| 0.790454 0.778539 0.632216 0.513168 2.718 ||| ||| 859 1074
until ||| até ||| 0.221516 0.221303 0.739904 0.859229 2.718 ||| ||| 4880 1461
now ||| agora ||| 0.725012 0.790323 0.571703 0.497706 2.718 ||| ||| 2051 2601
done ||| feito ||| 0.315104 0.268908 0.368902 0.305246 2.718 ||| ||| 1152 984
only ||| somente ||| 0.77724 0.795566 0.0553353 0.0538423 2.718 ||| ||| 413 5801
on the basis ||| a partir ||| 0.0095518 0.000204552 0.100775 0.00350264 2.718 ||| ||| 1361 129
of ||| de ||| 0.591588 0.414888 0.593874 0.436519 2.718 ||| ||| 132239 131730
intuition ||| intuição ||| 0.727273 0.857143 0.571429 0.666667 2.718 ||| ||| 22 28
, ||| , ||| 0.861116 0.868727 0.909177 0.883301 2.718 ||| ||| 253146 239764
without any ||| sem nenhuma ||| 0.666667 0.197121 0.136691 0.0471304 2.718 ||| ||| 57 278
experimental evidence ||| evidência experimental ||| 0.319336 0.700538 0.239502 0.106604 2.718 ||| ||| 3 4
. ||| . ||| 0.951828 0.99432 0.971535 0.991891 2.718 ||| ||| 144151 141227
...
```

Fonte: extraída de Caseli (2015).

Assim como ocorre nos sistemas RBMT, após a fase de desenvolvimento do sistema, que no caso dos sistemas PBSMT equivale ao treinamento dos modelos de língua e de tradução, o sistema está pronto para traduzir uma sentença nova. Por exemplo, no caso do sistema PBSMT inglês-português treinado como descrito acima, dada a entrada ilustrada no Quadro 6-a, o sistema gera a tradução apresentada no Quadro 5-b com base nos modelos de língua (ilustrado na Figura 2) e de tradução (ilustrado na Figura 3).[16] Para se ter uma ideia da qualidade da tradução gerada, uma tradução de referência (gerada por um humano e, portanto, considerada correta) é apresentada no Quadro 6-c. As diferenças entre a sentença produzida pelo sistema PBSMT (Quadro 6-b) e a sentença de referência (Quadro 6-c) aparecem em destaque (sublinhadas).

[15] Cada linha desta tabela contém informações separadas pelos caracteres '|||'. Essas informações são: a frase fonte (em inglês), a frase alvo (em português), as probabilidades de tradução de frase e de palavras, entre outros.

[16] Detalhes sobre o processo de treinamento do sistema PBSMT inglês-português podem ser consultados no trabalho de Caseli (2015).

Quadro 6 - Exemplo de sentença fonte (em inglês) de entrada (a), a sentença alvo (em português) gerada pelo sistema PBSMT (b) e a respectiva tradução de referência (c).

a	The conclusions open up the possibility of indicating the most suitable exercises for specific diseases, something until now done only on the basis of intuition, without any experimental evidence.
b	As conclusões abrem a possibilidade de indicar os exercícios mais adequado para certas doenças, algo feito até agora somente a partir de intuição, sem nenhuma evidência experimental.
c	As conclusões abrem a possibilidade de indicar os exercícios mais apropriados para doenças específicas, algo feito até agora com base apenas na intuição, sem evidências experimentais.

Fonte: adaptado de Caseli (2015).

Assim, a TA estatística tem como vantagens: o baixo custo de geração do sistema, que pode ser treinado em algumas horas usando um dos *toolkits* disponíveis gratuitamente[17]; sua aplicabilidade a, possivelmente, qualquer par de línguas e tipo de *corpus*; sua independência de código em relação à língua, uma vez que os mesmos scripts e programas de computador podem ser aplicados para quaisquer línguas; e sua simplicidade de processamento (tradução de novas sentenças). Como limitações principais estão: a dificuldade de interpretar/editar o conhecimento de tradução representado nos modelos estatísticos (como pode ser constatado nos trechos apresentados nas Figuras 2 e 3); a dependência em relação ao *corpus* de treinamento e, muitas vezes, à língua ponte[18]; e a incapacidade de generalização e de modelagem de aspectos estruturais e sintáticos da língua (KITAMURA, 2004).

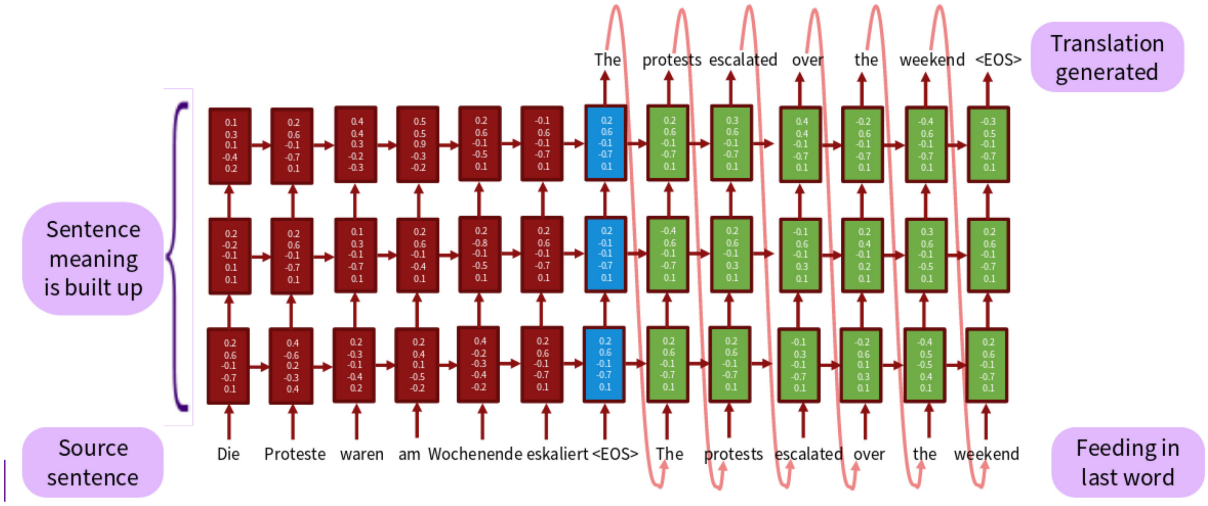
2.3. Tradução automática neural

Considerando-se as limitações das estratégias de TA ilustradas anteriormente, recentemente uma nova proposta surgiu com ênfase na captura dos aspectos estruturais da língua: a TA neural (KALCHBRENNER; BLUNSOM, 2013; CHO et al., 2014; SUTSKEVER et al., 2014). A TA neural já apresentou resultados promissores para diversos pares de línguas (LUONG et al., 2015a; JEAN et al., 2015; LUONG et al., 2015b; SENNRICH et al., 2016; LUONG; MANNING, 2016). Diferentemente da PBSMT, na qual as probabilidades governam a tradução, na NMT a tradução é aprendida usando redes neurais como ilustra a Figura 4.

[17] *Toolkits* são conjuntos de ferramentas computacionais criados para o desenvolvimento de algum sistema ou recurso computacional. O *toolkit* de geração de tradutor automático estatístico mais utilizado é o Moses (KOEHN et al., 2007), disponível em: http://www.statmt.org/moses. Acesso em: 26 jan. 2017.

[18] Língua ponte é a língua usada como intermediária na tradução. Por exemplo, na versão antiga dos tradutores do Google (disponível até setembro/2016), ao traduzir a sentença "Hoje eu quero matar dois coelhos" (em português) para o espanhol, o sistema gerava "Hoy quiero matar dos pájaros". Isso ocorria, provavelmente, por uma influência da língua ponte inglês que fez com que o sistema aprendesse, equivocadamente, que a tradução de "matar dois coelhos" é "kill two birds".

Figura 4 - Exemplo de uma rede neural recorrente em um sistema de TA neural.



Fonte: Luong et al. (2016, slide 23).

A TA neural tem como principais limitações: a alta complexidade computacional para treinamento das redes neurais e, em decorrência dessa complexidade, sua incapacidade, no momento, de lidar com grandes vocabulários, ou seja, para um conjunto muito grande de palavras pode ser intratável a geração de um sistema de TA neural. Segundo Luong et al. (2016), embora a TA neural tenha apresentado ganhos em relação aos sistemas PBSMT, ela também apresenta as mesmas limitações como o fato de não utilizar explicitamente as informações sintáticas e semânticas. Suas principais vantagens são: o pouco conhecimento linguístico necessário para gerar o tradutor, a otimização conjunta de toda a rede (e não de modelos separados, como ocorre na TA estatística) e o fato de gerar um conhecimento mais compacto (do que o conjunto de regras usado na TA baseada em regras, por exemplo).

3. Considerações finais

Neste artigo foi apresentada uma explanação sobre as estratégias aplicadas à TA e suas limitações. Essas limitações, somadas à importância e à utilidade cada vez maior dos sistemas de TA, fomentam as pesquisas atuais em TA.

Entre as principais estratégias de TA, três receberam especial atenção neste artigo: a TA baseada em regras, a TA estatística baseada em frases e a TA neural. Enquanto a TA baseada em regras é muito cara em termos de tempo e pessoal especializado para gerar as bases de conhecimento, a TA estatística e a TA neural baseiam-se no conhecimento codificado em *corpora* paralelos para aprenderem como traduzir. A TA estatística baseada em frases era o estado-da-arte "absoluto" até o surgimento da TA neural há cerca de 5 anos atrás. Essa nova

abordagem vem para superar problemas inerentes da TA estatística, como sua incapacidade de generalização e de modelagem da estrutura (hierarquia) da língua.

Após muitos anos de pesquisa em TA, com o surgimento de estratégias inovadoras que pareciam ser a solução para todos os desafios da TA, o que se pode concluir é que seja qual for a estratégia selecionada para realizar a TA o importante é ter em mente que ela deve ser avaliada considerando-se, sempre: (1) a utilidade e (2) o domínio para o qual a tradução é gerada. Nenhuma estratégia, proposta até o presente momento, mostrou-se capaz de atingir as ambiciosas metas estabelecidas nos primórdios da TA: produzir traduções completamente automáticas de boa qualidade para domínios irrestritos.

Referências

ARMENTANO-OLLER, C.; CARRASCO, R. C.; CORBÍ-BELLOT, A. M.; FORCADA, M. L.; GINESTÍ-ROSELL, M.; ORTIZ-ROJAS, S.; PÉREZ-ORTIZ, J. A.; RAMÍREZ-SÁNCHEZ, G.; SÁNCHEZ-MARTÍNEZ, F.; SCALCO, M. A. Open-source Portuguese-Spanish machine translation. In: INTERNATIONAL WORKSHOP ON COMPUTATIONAL PROCESSING OF WRITTEN AND SPOKEN PORTUGUESE, 7., 2006, Itatiaia. **Lecture Notes in Computer Science**. Itatiaia: PROPOR, 2006, p. 50-59. https://doi.org/10.1007/11751984_6

AZIZ, W.; SPECIA, L. Fully Automatic Compilation of Portuguese-English and Portuguese-Spanish Parallel Corpora. In: BRAZILIAN SYMPOSIUM IN INFORMATION AND HUMAN LANGUAGE TECHNOLOGY, 8., 2011, Cuiabá. **Proceedings…**, Cuiabá: BSIHLT, 2011, p. 234-238.

BROWN, R. D. Example-Based Machine Translation in the Pangloss System. In: INTERNATIONAL CONFERENCE ON COMPUTATIONAL LINGUISTICS, 16., 1996, Copenhagen. **Proceedings…**, Copenhagen: COLING, 1996, p. 169-174. https://doi.org/10.3115/992628.992660

CASELI, H. M. **Indução de léxicos bilíngües e regras para a tradução automática**. Maio 2007. 158 p. Tese (Doutorado em Computação e Matemática Computacional) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo. São Paulo, SP, 2007. https://doi.org/10.11606/T.55.2007.tde-29082007-090905

CASELI, H. M. Tradução automática: o uso de corpora paralelos para a criação de um tradutor automático estatístico. In: VIANA, V.; TAGNIN, S. E. O. (Org.). **Corpora na tradução**. 1ed. São Paulo: Hub editorial, 2015, p. 243-267.

CASELI, H. M.; NUNES, M. G. V.; FORCADA, M. L. Automatic induction of bilingual resources from aligned parallel corpora: application to shallow-transfer machine translation. **Machine Translation**, Amsterdam, v. 20, p. 227-245, 2006. https://doi.org/10.1007/s10590-007-9027-9

CHO, K.; MERRIENBOER, B. V.; GULCEHRE, C.; BAHDABAU, D.; BOUGARES, F.; SCHWENK, H.; BENGIO, Y. Learning phrase representations using RNN encoder-decoder for statistical machine translation. In: CONFERENCE OF EMPIRICAL METHODS IN NATURAL LANGUAGE PROCESSING, Doha, 2014. **Proceedings…**, Doha: EMNLP, 2014, p. 1724-1734. https://doi.org/10.3115/v1/D14-1179

GALLEY, M.; HOPKINS, M.; KNIGHT, K.; MARCU, D. What's in a translation rule? In: HUMAN LANGUAGE TECHNOLOGY CONFERENCE AND MEETING OF THE NORTH AMERICAN CHAPTER OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS, 4., Edmonton, 2004. **Proceedings…**, Edmonton: HLT-NAACL, 2004, p. 273-280. https://doi.org/10.21236/ADA460212

GÜVENIR, H. A.; CICEKLI, I. Learning translation templates from examples. **Information Systems**, v. 23, n. 6, p. 353-363, 1998. https://doi.org/10.1016/S0306-4379(98)00017-9

HUTCHINS, W. J. Machine translation: A concise history. In: WAI, C. S. (Ed.) **Computer Aided Translation**: Theory and Practice. Hong Kong: Chinese University of Hong Kong, 2007. https://doi.org/10.1007/s10590-006-9003-9

JEAN, S.; CHO, K.; MEMISEVIC, R.; BENGIO, Y. On using very large vocabulary for neural machine translation. In: ANNUAL MEETING OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS, 53., Beijing, 2015. **Proceedings…**, Beijing: ACL, 2015, p. 1-10. https://doi.org/10.3115/v1/P15-1001

KALCHBRENNER, N.; BLUNSOM, P. Recurrent continuous translation models. In: CONFERENCE OF EMPIRICAL METHODS IN NATURAL LANGUAGE PROCESSING, Washington, 2013. **Proceedings…**, Washington: EMNLP, 2013, p. 1700-1709.

KITAMURA, M. **Translation knowledge acquisition for pattern-based machine translation**. 2004, 114 f. Thesis (Doctorate) – Department of Information Processing, Graduate School of Information Science, Nara Institute of Science and Technology, 2004.

KOEHN, P.; OCH, F. J.; MARCU, D. Statistical phrase-based translation. In: HUMAN LANGUAGE TECHNOLOGY CONFERENCE AND MEETING OF THE NORTH AMERICAN CHAPTER OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS, 3., Edmonton, 2003. **Proceedings…**, Edmonton: HLT-NAACL, 2003, p. 127-133. https://doi.org/10.3115/1073445.1073462

KOEHN, P.; HOANG, H.; BIRCH, A.; CALLISON-BURCH, C.; FEDERICO, M.; BERTOLDI, N.; COWAN, B.; SHEN, W.; MORAN, C.; ZENS, R.; DYER, C.; BOJAR, O.; CONSTANTIN, A.; HERBST, E. Moses: Open Source Toolkit for Statistical Machine Translation. In: ANNUAL MEETING OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS, 45., 2007, Prague. **Proceedings…**, Prague: ACL, 2007, p. 177-180.

LOPEZ, A. Statistical Machine Translation. **ACM Computing Surveys**, New York, v. 40, n. 3, p. 1-49, 2008. https://doi.org/10.1145/1380584.1380586

LUONG, M.; SUTSKEVER, I.; LE, V. Q.; VINYALS, O.; ZAREMBA, W. Addressing the rare word problem in neural machine translation. In: ANNUAL MEETING OF THE ASSOCIATION FOR COMPUTACIONAL LINGUISTICS, 53., 2015, Lisbon. **Proceedings…**, Lisbon: ACL, 2015a, p. 11-19.

LUONG, M.; PHAM, H.; MANNING, C. D. Effective approaches to attention-based neural machine translation. In: ANNUAL MEETING OF THE ASSOCIATION FOR COMPUTACIONAL LINGUISTICS, 53., 2015, Lisbon. **Proceedings…**, Lisbon: ACL, 2015b, p. 1412-1421. https://doi.org/10.18653/v1/D15-1166

LUONG, M.; MANNING, C. D. Achieving open vocabulary neural machine translation with hybrid word-character models. In: ANNUAL MEETING OF THE ASSOCIATION FOR COMPUTACIONAL LINGUISTICS, 54., 2016, Berlin. **Proceedings…**, Berlin: ACL, 2016, p. 1054-1063. https://doi.org/10.18653/v1/P16-1100

LUONG, T.; CHO, K.; MANNING, C. **Neural Machine Translation**. Disponível em: http://nlp.stanford.edu/projects/nmt/Luong-Cho-Manning-NMT-ACL2016-v4.pdf. Acesso em: 17 out. 2016.

OCH, F. J.; NEY, H. The Alignment Template Approach to Statistical Machine Translation. **Computational Linguistics**, London, v. 30, n. 4, p. 417-449, 2004. https://doi.org/10.1162/0891201042544884

SENNRICH, R.; HADDOW, B.; BIRCH, A. Improving neural machine translation models with monolingual data. In: ANNUAL MEETING OF THE ASSOCIATION FOR COMPUTACIONAL LINGUISTICS, 54., 2016, Berlin. **Proceedings…**, Berlin: ACL, 2016, p. 86-96. https://doi.org/10.18653/v1/P16-1009

SUTSKEVER, I.; VINYALS, O.; LE, Q. V. Sequence to sequence learning with neural networds. In: INTERNATIONAL CONFERENCE ON NEURAL INFORMATION PROCESSING SYSTEMS, 27., Montréal, 2014. **Proceedings…**, Montréal: NIPS, 2014, p. 3104-3112.

Artigo recebido em: 13.04.2017 Artigo aprovado em: 08.05.2017