Universidade de São Paulo - USP
Universidade Federal de São Carlos - UFSCar
Universidade Estadual Paulista - UNESP

Tradução Automática Estatística baseada em Frases e Fatorada: Experimentos com os idiomas Português do Brasil e Inglês usando o *toolkit* Moses

Helena de Medeiros Caseli
Israel Aono Nunes

NILC-TR-09-07

Novembro, 2009

Série de Relatórios do Núcleo Interinstitucional de Linguística Computacional
NILC - ICMC-USP, Caixa Postal 668, 13560-970 São Carlos, SP, Brasil

Resumo

Este relatório apresenta uma descrição do uso do *toolkit* de tradução automática estatística Moses na construção e na avaliação de modelos de tradução baseados em frases (*phrase-based*) tradicionais (considerados o estado da arte) e fatorados (uma extensão dos modelos baseados em frases). Além de apresentar uma descrição da ferramenta utilizada, seu processo de instalação e utilização, também são relatados os resultados alcançados em vários experimentos desenvolvidos para testar a tradução automática estatística baseada em frases e a fatorada com um *corpus* paralelo de textos escritos em português do Brasil (pt) e inglês (en). Os experimentos demonstram que a tradução fatorada, na qual fatores adicionais (além das formas superficiais das palavras) são usados na geração dos modelos de tradução e língua, apresenta resultados melhores do que a tradução tradicional baseada em frases. Essa melhora no desempenho, verificada em termos das medidas de avaliação automática BLEU e NIST, mostrou-se estatisticamente significante em alguns experimentos no sentido de tradução en-pt, no qual as informações adicionais na língua alvo (o português nesse caso) possuem maior relevância por ser esta uma língua com maior variação morfológica do que a língua fonte (o inglês, nesse caso).

Índice

Tradução automática estatística baseada em frases e fatorada: experimentos com os idiomas português do Brasil e inglês usando o *toolkit* Moses[1]

1 Introdução

Este relatório apresenta uma descrição da aplicação do *toolkit* de código aberto Moses[2] na investigação do uso de informação adicional na tradução automática estatística, em especial o uso de informação morfossintática e sintática.

A *tradução estatística* começou e ser pesquisada mais profundamente a partir de meados dos anos 80 principalmente com o desenvolvimento do projeto Candide da IBM (Koerner & Asher, 1995). Inicialmente, os métodos aplicados mapeavam os textos palavra por palavra. Alguns anos depois, vários pesquisadores sugeriram e provaram que o mapeamento dos textos baseado em frases[3] era mais eficiente que o método que vinha sendo aplicado. Essa abordagem de tradução automática estatística que gera modelos de tradução para frases e não palavras (como nas primeiras versões), por exemplo (Koehn et al., 2003) e (Och & Ney, 2004), é considerado, hoje, o estado da arte.

A Figura 1 abaixo mostra como o modelo de tradução automática estatística baseado em frases (*phrase-based*) funciona para um par de sentenças alemão-inglês. Nesse exemplo, as frases (sequências de palavras) estão delimitadas pelos retângulos. Mais detalhes sobre a tradução automática estatística podem ser obtidos em (Lopez, 2008).

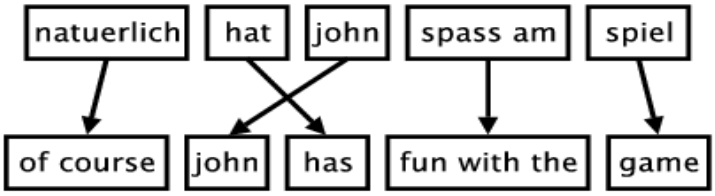

Figura 1 – Exemplo de uma tradução baseada em frases[4].

A tradução fatorada (do inglês, *factored translation* ou FT), prevê o uso de informação complementar (além da forma superficial das palavras normalmente usada) em sistemas de tradução automática estatística baseada em frase (*Phrase-based Statistical Machine Translation*). Na FT, os modelos estatísticos levam em consideração uma anotação,

[1] Este trabalho foi apoiado por PIADRD/PUIC/UFSCar e FAPESP.
[2] http://www.statmt.org/moses/
[3] O termo "frase" será usado nesse texto como uma tradução para o termo *phrase* do inglês que, em tradução automática estatística, não se refere necessariamente a um sintagma, mas sim a uma sequência qualquer de palavras.
[4] http://www.statmt.org/moses/?n=Moses.Background

em nível de palavras, enriquecida (além das formas superficiais) com diversos fatores como lemas, *part-of-speech*, características morfológicas (gênero, número, tempo verbal etc.) e informação sintática (etiquetas sintáticas superficiais e fatores que garantem concordância entre itens relacionados sintaticamente). Segundo Koehn e Hoang (2007), a principal diferença entre FT e SMT baseada em frases (o estado da arte na SMT) está na preparação dos dados de treinamento e nos tipos de modelos aprendidos a partir desses dados, ou seja, os modelos baseados em frases são, na verdade, um caso especial de FT.

Porém, como mencionado por Koehn e Hoang (2007), esses modelos se limitam a mapear pequenas porções de texto sem o uso explícito de nenhuma informação linguística seja ela morfológica, sintática ou semântica. Esses autores vão além e afirmam que embora o uso de informação adicional em passos de pré ou pós-processamento tenha apresentado melhoras no desempenho (Lee, 2004; Sadat & Habash, 2006; Och et al., 2004; entre outros), uma integração mais "forte" da informação linguística no modelo de tradução é desejável por duas razões:

1. Os modelos de tradução que operam com representações mais gerais – como lemas de palavras ao invés de suas formas superficiais – podem gerar estatísticas mais ricas e superar problemas de dados esparsos encontrados quando a quantidade de dados de treinamento é limitada.
2. Muitos aspectos da tradução podem ser melhor explicados em um nível morfológico, sintático ou semântico e a disponibilidade dessas informações, para o modelo de tradução, permite a modelagem direta desses aspectos. Por exemplo, a reordenação em nível sentencial é guiada principalmente por princípios sintáticos, restrições de concordância locais presentes na morfologia etc.

Assim, a tradução fatorada surge como uma extensão da abordagem baseada em frases: na qual uma palavra não é apenas um *token* mas um vetor de fatores que representam diferentes níveis de anotação. Entre os possíveis fatores pode-se citar: forma superficial, lema, *part-of-speech* (PoS), atributos morfológicos (gênero, número etc.), classes de palavras geradas automaticamente, etiquetas sintáticas superficiais, assim como fatores dedicados a garantir concordância entre itens relacionados sintaticamente. A tradução fatorada também está relacionada aos modelos de transferência baseados em árvore (Wu, 1997; Alshawi et al., 1998; Yamada & Knight, 2001; Melamed, 2004; Galley et al., 2006; entre outros) com a diferença de que, na tradução fatorada, o foco não está tão voltado para a estrutura sintática recursiva mas, sim, para uma anotação mais rica no nível lexical (Koehn & Hoang, 2007).

A tradução de representações fatoradas de palavras de entrada em representações fatoradas de palavras de saída é dividida em uma sequência de passos de mapeamento que (a) traduzem fatores de entrada em fatores de saída ou (b) geram fatores de saída adicionais a partir de fatores de saída existentes. Por exemplo, no modelo apresentado na Figura 2, o processo de tradução é dividido em três passos de mapeamento: (1) a tradução dos lemas de entrada em lemas de saída, (2) a tradução dos fatores morfológicos e de PoS e (3) a geração das formas superficiais a partir do lema e dos fatores linguísticos. Todos os passos de tradução operam no nível de frase enquanto os passos de geração operam no nível lexical.

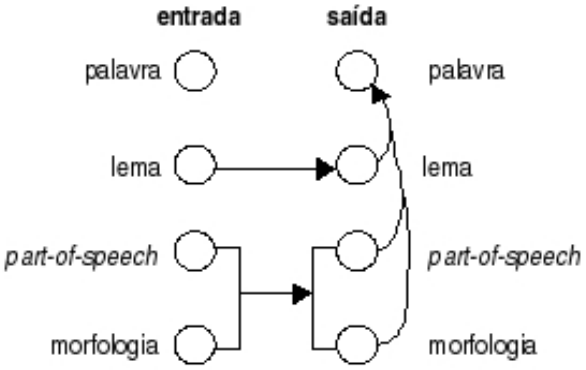

Figura 2 – Exemplo de uma tradução fatorada – adaptação de (Koehn & Hoang, 2007).

Kohen e Hoang (2007) utilizaram o *toolkit* de código aberto de Moses[5] (Koehn et al., 2007a, 2007b) para o treinamento e a geração dos modelos estatísticos, bem como para a avaliação dos modelos gerados. Moses contém componentes para: (a) pré-processar dados, (b) treinar os modelos de língua e de tradução, (c) otimizar esses modelos usando o treinamento de menor taxa de erro (*minimum error rate* ou MER) proposto por Och (2003) e (d) avaliar as traduções resultantes usando as medidas BLEU (Papineni et al., 2002) e NIST (Doddington, 2002).

O treinamento é realizado a partir do *corpus* paralelo anotado com fatores adicionais e alinhado lexicalmente. O método de alinhamento de palavras pode operar nas formas superficiais das palavras ou em qualquer outro fator. A partir dessas informações, cada passo de mapeamento é um componente do modelo geral. Além disso, as tabelas de tradução e de geração devem ser geradas como descrito a seguir:

- *Tabelas de tradução*

  Para os fatores especificados na entrada (lado fonte) e na saída (lado alvo), determinam-se os mapeamentos de frases e suas pontuações com base em contagens relativas e probabilidades de tradução baseadas em palavras.

[5]http://www.statmt.org/moses/.

- *Tabelas de geração*

  As distribuições de geração são estimadas no lado alvo sem o uso de informações a respeito dos alinhamentos lexicais, mas utilizando, possivelmente, informações monolíngues adicionais. O modelo de geração é aprendido palavra-a-palavra.
- *Modelos de língua*

  Os modelos de língua, na tradução fatorada, são definidos para cada fator ou conjunto de fatores.

Os modelos de tradução automática estatística (baseada em frases ou fatorada) são construídos a partir de *corpus* paralelo alinhado sentencialmente, ou seja, um conjunto de sentenças que são traduções mútuas. O *corpus* utilizado nos experimentos relatados nesse documento é descrito na seção 3 juntamente com o relato dos experimentos realizados e seus respectivos resultados. A seguir, na Seção 2, apresenta-se uma descrição do Moses (*toolkit* de tradução automática empregado nos experimentos aqui descritos). Por fim, a Seção 4 traz as considerações finais e propostas de trabalhos futuros dessa pesquisa.

2 Moses

O Moses é um *toolkit* de tradução automática estatística de código aberto (*open source*) que permite realizar traduções de maneira automática entre quaisquer pares de línguas, para tanto é necessário apenas um *corpus de treinamento* composto por sentenças alinhadas em ambas as línguas a partir do qual serão gerados os modelos estatísticos usados na tradução.

O Moses foi desenvolvido a partir de uma ferramenta já existente chamada Pharaoh[6] e financiado pelo projeto EuroMatrix, P6-IST-5-034291-STP. Com o Moses também é possível realizar experimentos com *tradução fatorada* (Koehn & Hoang, 2007), que consiste em gerar modelos de tradução a partir de informações adicionais como:

- *Part-of-Speech* (categoria gramatical do tipo: substantivo, verbo, adjetivo, advérbio etc.)
- Lemas (forma base de uma palavra, por exemplo, o lema da palavra "cantou" é "cantar")
- Informações morfológicas (traços morfológicos que indicam, por exemplo, gênero – feminino ou masculino – ou número – singular ou plural – de uma certa palavra)
- entre outros

A tradução fatorada, assim como a tradução automática estatística baseada em frases, faz uso de um alinhador de palavras para gerar os modelos de tradução. Para tanto, Moses utiliza o alinhador lexical automático GIZA++[7] (Och & Ney, 2000), o qual determina quais são os melhores alinhamentos de palavras (e possivelmente sequências de palavras) a partir do qual as probabilidades de tradução serão geradas para o modelo.

A fim de permitir ao leitor a replicação de todos os experimentos descritos neste documento, bem como realizar seus próprios experimentos por meio do Moses, as próximas subseções apresentam um relato detalhado dos procedimentos necessários para utilização do Moses: (2.1) Instalação, (2.2) Treinamento, (2.3) Teste e (2.4) Avaliação. Juntamente com a descrição de cada etapa são apresentadas as linhas de comando pertinentes.

2.1 Instalação

Nesta seção são apresentados os comandos necessários para a realização da instalação do *toolkit* Moses com todos os seus componentes. Vale dizer que a instalação aqui descrita foi realizada tendo como base o sistema operacional Linux, distribuição Ubuntu, versão 8.04.

[6] http://www.isi.edu/licensed-sw/pharaoh/
[7] http://www.fjoch.com/GIZA++.html

Inicialmente, cria-se o diretório no qual as ferramentas auxiliares deverão ser armazenadas :

```
% mkdir tools
% cd tools
```

Em seguida, realiza-se o *download* e a compilação dos componentes (ferramentas auxiliares) GIZA++ e SRILM[8]. O primeiro, como já mencionado anteriormente, é a ferramenta de alinhamento lexical automática enquanto o segundo trata-se do componente responsável pela criação do modelo de língua (indispensável no processo de tradução como explicado adiante). Para *download* e instalação de GIZA++, os seguintes comandos foram usados:

```
% wget http://giza-pp.googlecode.com/files/giza-pp-v1.0.2.tar.gz
% curl -O http://giza-pp.googlecode.com/files/giza-pp-v1.0.2.tar.gz
% tar -xzvf giza-pp-v1.0.2.tar.gz
% cd giza-pp

% make
```

Após o *download* e a compilação dos arquivos referentes a GIZA++, os executáveis deve ser compilados para bin/pasta. Um procedimento semelhante foi realizado, então, para *download* e compilação do SRILM.

```
% mkdir srilm
% cd srilm
% tar -xzvf srilm.tgz
% chmod +w Makefile
```

O makefile do SRILM deve ser editado para apontar para sua pasta como apresentado a seguir, onde “<” indica o conteúdo original do arquivo e “>” o que deve ser alterado:

```
< # SRILM = /home/speech/stolcke/project/srilm/devel
---
> SRILM = /home/jschroe1/demo/tools/srilm
```

[8] http://www.speech.sri.com/projects/srilm/download.html

Em seguida SRILM deve ser compilado por meio do comando:

```
% make World
```

Caso ocorra um erro de o comando acima não reconhecer o arquivo machine-type, instale o pacote csh por meio do comando:

```
% sudo apt-get install csh
```

Também é importante, para que a instalação seja realizada com sucesso, que a versão do gcc++ seja a correta. Para tanto, pode-se modificar o arquivo Makefile.machine.<nome> (onde <nome> geralmente é i636) para alterar a versão do gcc para 4.1 por exemplo, como apresentado abaixo:

```
< CXX = $(GCC_PATH)g++ $(GCC_FLAGS) -DINSTANTIATE_TEMPLATES
---
> CXX = $(GCC_PATH)g++-4.1 $(GCC_FLAGS) -DINSTANTIATE_TEMPLATES
```

Outra tarefa importante para garantir a correta execução da ferramenta é adicionar o SRILM à variável de ambiente $PATH como apresentado abaixo:

```
% export
PATH=/home/jschroe1/demo/tools/srilm/bin/i686:/home/jschroe1/
demo/tools/srilm/bin:$PATH
% export
PATH=/Users/josh/demo/tools/srilm/bin/macosx:/Users/josh/demo/
tools/srilm/bin:$PATH
```

Em seguida o SRILM deve ser testado e, se nenhuma mensagem de erro for gerada apontando alguma pendência ou incompatibilidade, compilado por meio da execução dos comandos:

```
% cd test
% make all
```

Após a instalação das ferramentas auxiliares deve-se proceder com o *download* da última versão do Moses como mostrado a seguir:

```
% mkdir moses
% svn co
https://mosesdecoder.svn.sourceforge.net/svnroot/mosesdecoder/ trunk moses
```

Após o *download*, o Moses deve ser compilado executando-se:

```
% cd moses
% ./regenerate-makefiles.sh
% ./configure --with-srilm=/home/jschroe1/demo/tools/srilm \
–with- irstlm=/home/jschroe1/demo/tools/irstlm
% make -j 2
```

Vale mencionar algo muito importante neste ponto: se for realizada alguma modificação em qualquer pasta ou arquivo a partir desta etapa, os comandos acima deverão ser re-executados.

Em seguida, alguns scripts de suporte devem ser compilados como descrito abaixo:

```
% mkdir moses-scripts
% cd moses/scripts
```

É necessário, também, editar o arquivo Makefile do Moses como no exemplo abaixo, onde “<” indica o conteúdo original do arquivo e “>” o que deve ser alterado.

```
< TARGETDIR?=/home/s0565741/terabyte/bin
< BINDIR?=/home/s0565741/terabyte/bin
---
> TARGETDIR?=/home/jschroe1/demo/tools/moses-scripts
> BINDIR?=/home/jschroe1/demo/tools/bin

% make release
```

O comando acima é realizado para compilar os scripts complementares na pasta recém criada `moses-scripts` como indicam as alterações efetuadas no makefile modificado algumas linhas acima.

A variável de ambiente SCRIPTS_ROOTDIR deve ser definida para apontar para o diretório que contém os scripts do Moses:

```
% export SCRIPTS_ROOTDIR=/home/username/lab4/moses-scripts/scripts-YYYYMMDD-HHMM
```

Resta ainda um passo de instalação de scripts adicionais.

```
% cd ../../
% wget http://homepages.inf.ed.ac.uk/jschroe1/how-to/scripts.tgz
% curl -O http://homepages.inf.ed.ac.uk/jschroe1/how-to/scripts.tgz
% tar -xzvf scripts.tgz
```

Por fim, realiza-se a instalação do script para a avaliação automática da tradução que faz o cálculo das medidas BLEU (Papineni et al., 2002) e NIST (Doddington, 2002), as quais retornam um valor para uma tradução automática ao compararem os n-gramas da tradução gerada automaticamente com base nos modelos treinados e os n-gramas de uma (ou mais) traduções de referência (consideradas corretas).

```
% wget ftp://jaguar.ncsl.nist.gov/mt/resources/mteval-v11b.pl
```

2.2 Treinamento

Após a instalação com sucesso do *toolkit* de tradução automática Moses, conforme descrito na subseção anterior, pode-se proceder com o treinamento dos modelos de língua e tradução utilizando o Moses por meio da execução dos comandos descritos nessa subseção. O *corpus* paralelo português-inglês utilizado nos experimentos descritos nesse documento é apresentado na seção 3.

Antes de iniciar o treinamento propriamente dito deve-se criar um diretório no qual serão armazenados todos os arquivos utilizados e produzidos por Moses no processo de treinamento de um modelo de tradução.[9]

```
% mkdir work
% mkdir work/corpus
```

O primeiro passo no processo de treinamento está relacionado à tokenização (separação das unidades mínimas de processamento: palavras, caracteres de pontuação etc.)

[9]A partir desse ponto, todos os comandos serão executados a partir do diretório Moses/moses considerando-se o mesmo contém bin/moses-scripts/scripts-YYYYMMDD-HHMM e scripts . Considera-se, também que o diretório contendo a ferramenta para geração do modelo de língua, nesse caso o SRILM, está em Moses.

do *corpus* de treinamento representado pelos arquivos `treinamento.pt` e `treinamento.en` por meio dos comandos:

```
% scripts/tokenizer.perl -l pt \
< work/corpus/treinamento.pt \
> work/corpus/novo.tok.pt
% scripts/tokenizer.perl -l en \
< work/corpus/treinamento.en \
> work/corpus/novo.tok.en
```

Em seguida, conforme indicado pelas diretrizes de Moses, as sentenças longas são filtradas por meio da execução do comando:

```
%       bin/moses-scripts/scripts-YYYYMMDD-HHMM/training/clean-corpus-
n.perl \
work/corpus/novo.tok \
pt en work/corpus/novo.clean 1 40
```

O último passo de pré-processamento do *corpus* antes de proceder com o treinamento dos modelos é responsável pela conversão das letras para minúsculas (*lowercased*) com os comandos:

```
% scripts/lowercase.perl < work/corpus/novo.clean.pt \
> work/corpus/novo.lowercased.pt
% scripts/lowercase.perl < work/corpus/novo.clean.en \
> work/corpus/novo.lowercased.en
```

O treinamento propriamente dito tem início com a construção do modelo de língua (.lm). O modelo de língua é construído a partir do *corpus* de sentenças na língua alvo e é usado, durante a tradução, para ordenar as sentenças geradas automaticamente de acordo com suas probabilidades de serem sentenças alvo válidas. Esse modelo nada mais é do que um conjunto de n-gramas na língua alvo acompanhados de suas probabilidades de ocorrência de acordo com o *corpus* de treinamento. Como entrada para o comando tem-se o *corpus* alvo tokenizado, o qual é convertido para minúsculas e usado para geração de um modelo de língua de ordem 3, ou seja, composto de unigramas, bigramas e trigramas conforme indica o parâmetro `-order 3`.

```
% mkdir work/lm
% scripts/lowercase.perl \
< work/corpus/novo.tok.en \
> work/lm/novo.lowercased.en
% ../srilm/bin/i686/ngram-count -order 3 \
-interpolate -kndiscount -unk \
-text work/lm/novo.lowercased.en \
-lm work/lm/novo.lm
```

Por fim, o treinamento se completa com a geração do modelo de tradução baseado em frases que tem como entrada o *corpus* de treinamento em português (pt) e inglês (en). O modelo de tradução, relaciona frases fonte com frases alvo indicando a probabilidade de tradução de uma para outra de acordo com as ocorrências encontradas no *corpus* de treinamento.

```
% bin/moses-scripts/scripts-YYYYMMDD-HHMM/training/train-factored-
phrase-model.perl \
-scripts-root-dir bin/moses-scripts/scripts-YYYYMMDD-HHMM/ \
-root-dir work \
-corpus work/corpus/novo.lowercased \
-f pt -e en \
-alignment grow-diag-final-and \
-reordering msd-bidirectional-fe \
-lm 0:3:work/lm/novo.lm
```

Vale dizer que o comando de treinamento apresentado acima está composto pelos parâmetros padrão do Moses para treinamento do modelo de tradução baseado em frases. Nos experimentos descritos na seção 3, parâmetros adicionais foram utilizados com os quais melhores resultados foram alcançados.

2.3 Tradução

Após a instalação (descrita na seção 2.1) do Moses e de seu uso para treinamento (descrito na seção 2.2) dos modelos de língua e tradução, esta seção descreve como tais modelos são usados na tradução de um conjunto de sentenças inédito (corpus de teste). O processo de tradução, em tradução automática estatística, recebe o nome de *decodificação*. Esta etapa é relativamente a mais simples, mas depende de uma realização correta de todas as etapas descritas acima.

```
% moses-cmd/src/moses \
-f work/model/moses.ini < work/fapesp_pt.txt > work/novo.output
```

2.4 Avaliação

Por fim, após a tradução de um conjunto inédito de sentenças (descrita na seção 2.3) a partir dos modelos de língua e tradução treinados com o Moses e o *corpus* de treinamento (conforme descrito na seção 2.2), a avaliação automática é realizada. Para tanto, antes de avaliar a saída gerada automaticamente comparando-a com uma (ou mais) sentenças de referência (consideradas corretas), alguns passos de pré-processamento são necessários.

O primeiro deles, que é opcional, está relacionado ao uso de um script que aplica à sentença traduzida automaticamente a mesma capitalização (maiúsculas e minúsculas) encontradas no *corpus* de treinamento, ou seja, o *recaser* é treinado a partir do *corpus* alvo por meio do comando:

```
% bin/moses-scripts/scripts-YYYYMMDD-HHMM/recaser/train-
recaser.perl \
-train-script  bin/moses-scripts/scripts-YYYYMMDD-HHMM/training/train-
factored-phrase-model.perl \
-ngram-count ../srilm/bin/i686/ngram-count \
-corpus work/lm/novo.tok.en \
-dir work/recaser
```

Após o treinamento do *recaser*, o modelo gerado é aplicado à tradução gerada automaticamente para “ajustar” a capitalização:

```
% bin/moses-scripts/scripts-YYYYMMDD-HHMM/recaser/recase.perl
-model \
work/recaser/moses.ini -in work/novo.output \
-moses moses-cmd/src/moses > work/novo.output.recased
```

A saída é então “detokenizada”:

```
% scripts/detokenizer.perl -l en < work/novo.output.recased >
work/novo.output.detokenized
```

Por fim, o último passo de pré-processamento é a geração dos arquivos no formato XML necessários como entrada para o script de avaliação automática:

```
% scripts/wrap-xml.perl \
work/novo.ref.sgm en \
< work/novo.output.detokenized  > work/novo.output.sgm
```

O comando a seguir gera os valores de NIST (Doddington, 2002) e BLEU (Papineni et al., 2002) para a saída automática da tradução estatística comparando-a com a referência.

```
% scripts/mteval-v11b.pl \
-s work/novo-src.pt.sgm \
-r work/novo-ref.en.sgm \
-t work/novo.output.sgm -c
```

No comando acima são passados três parâmetros, são eles:

- novo-src.pt.sgm – É o texto de origem em português, ele é identificado pela etiqueta “s” que indica fonte (*source* do inglês).
- novo-ref.en.sgm – É o texto de referência em inglês para a avaliação e é identificado pela etiqueta “r” que indica referência (*reference* do inglês).
- novo.output.sgm – É o texto gerado pelo MOSES a partir do texto fonte e é identificado por “t” que indica teste (test do inglês).

3 Experimentos e Resultados

3.1 Corpora de Treinamento e Teste

O *corpus* utilizado na realização dos experimentos apresentados nesse documento é resultado da compilação de textos retirados da revista científica *Pesquisa Fapesp*[10] escritos em português (pt) e depois traduzidos para o inglês (en). Os 646 pares de artigos dessa revista coletados para formar o *corpus de treinamento*, após o alinhamento sentencial e a eliminação de sentenças originais em pt sem tradução para en, resultou em 17.397 pares de sentenças paralelas num total de 1.026.512 *tokens* dos quais 494.391 são em português e 532.121, em inglês.

A Tabela 1 mostra a quantidade de *tokens*, e sentenças no *corpus de treinamento*.

Tabela 1 – Descrição do *corpus de treinamento* português-inglês usado nos experimentos descritos nesse documento (Caseli, 2007).

Idiomas	***Tokens***	**Sentenças**
Português	494.391	17.397
Inglês	532.121	17.397
Total	**1.026.512**	**34.794**

Para ser mais preciso, o *corpus* original conta com artigos de 9 seções de dita revista: ciência (205), editorial (11), estratégias (137), humanidades (40), linha de produção (111), memória (11), opinião (4), política (54) e tecnologia (73), escritos em estilos diferentes que vão desde relato de projetos até entrevistas com pesquisadores. Em média, cada arquivo desse *corpus* original contém 780 *tokens* em pt e 827 *tokens* em en, os tamanhos dos textos variam de 50 a 4.520 *tokens* na versão pt e de 54 a 4.927 *tokens* nos textos em en. O *token* mais frequente em pt é *","* (34.126 ocorrências) e o menos frequente é "*gaviões"* (1 ocorrência) enquanto o *token* mais frequente em en é "*the"* (45.478 ocorrências) e o menos frequente "*estereoscopic"* (1 ocorrência).

A Tabela 2 apresenta exemplos de sentenças paralelas português-inglês presentes no *corpus de treinamento*, já pré-processadas de acordo com a melhor configuração definida pelos experimentos descritos na seção 3.2, ou seja, com todas as letras convertidas para minúsculas e os caracteres de pontuação mantidos no texto. A separação dos *tokens*, conforme descrito anteriormente, é um passo prévio ao treinamento e, portanto, as palavras e caracteres de pontuação já aparecem delimitados por espaços nos exemplos da Tabela 2.

[10] Versão online da revista *Pesquisa Fapesp* está disponível em: http://revistapesquisa.fapesp.br/

Tabela 2 – Exemplos de pares de sentenças português-inglês do *corpus de treinamento* para a tradução automática estatística baseada em frases.

Português	**Inglês**
os dentes do mais antigo orangotango	the teeth of the oldest orangutan
uma nova espécie de hominídeo encontrado na tailândia , com estimados 12 milhões de anos , tornou - se o parente mais remoto dos atuais orangotangos (pongo pygmaeus) .	a new species of hominid found in thailand , with an estimated age of 12 million years , has become the most distant relative of today 's orangutans (pongo pygmaeus) .
um grupo de pesquisadores franceses ligados ao laboratório europeu de radiação síncrotron (esrf) chegou a essa conclusão comparando os 18 dentes do fóssil com a dentição de outros primatas antigos .	a group of french researchers connected with the european synchrotron radiation facility (esrf) arrived at this conclusion by comparing the 18 teeth of the fossil with the dentition of other ancient primates .

O corpus de treinamento pt-en usado nos experimentos aqui descrito já contava com informações complementares para cada forma superficial (a palavra da maneira como ocorre no texto, ex: meninos) das palavras em pt e en – lema (a forma base da palavra, ex: menino), part-of-speech (categoria gramatical ou part-of-speech da palavra, ex: substantivo) e traços morfológicos (valores dos atributos das PoS da palavra, ex: masculino plural) – conforme descrito em (Caseli, 2007). Porém, um novo nível de informação precisou ser adicionado para permitir o estudo do impacto também da informação sintática na tradução estatística. Assim, o corpus original foi pré-processado por meio da análise sintática das sentenças fonte e alvo. Para tanto, os analisadores sintáticos selecionados para tal tarefa foram o PALAVRAS (Bick, 2000) para o idioma português do Brasil e o parser do Collins (1999)[11] para o inglês, por serem estes os de melhor desempenho para os idiomas em questão.

Ao final desse processamento, para cada palavra nas sentenças paralelas português-inglês do *corpus* usado para treinamento dos modelos estatísticos, estavam associadas até 5 informações: (1) forma superficial, (2) lema, (3) categoria de *part-of-speech*, (4) traços morfológicos e (5) informações sintáticas. Dessas informações, apenas a primeira (formas superficiais) é usada para treinamento dos modelos estatísticos de TA baseada em frases enquanto os outros níveis foram usados nos experimentos com a tradução fatorada, sendo considerados fatores adicionais para aprendizado dos modelos. Enquanto a tradução fatorada aprende os modelos de tradução para cada fator separadamente, em vários passos de tradução, e os combina, ao final, possivelmente por meio de vários passos de geração; a tradução

[11]A implementação do *parser* do Collins (1999) utilizada no pré-processamento do *corpus* em inglês usado nos experimentos aqui descritos foi desenvolvida por Dan Bikel e está disponível em: http://www.cis.upenn.edu/~dbikel/software.html#stat-parser

estatística baseada em frases utiliza apenas as informações de formas superficiais para o aprendizado dos modelos de tradução.

O *corpus* paralelo pt-en com as informações adicionais em ambos os idiomas, gerado nesta etapa do projeto, encontra-se disponível para a utilização por todas as técnicas de TA a serem investigadas no projeto maior do qual esse faz parte, não apenas a tradução estatística fatorada, bem como por outros projetos na área e por toda a comunidade.

Assim, a Tabela 3 apresenta as mesmas sentenças da Tabela 2, porém, com os níveis de informação adicionais, separados pelo caractere "|", respectivamente: forma superficial, lema, categoria de PoS e essa categoria com os traços morfológicos. Esse é o formato de entrada da ferramenta Moses para treinamento dos modelos de tradução fatorada.

Tabela 3 – Sentenças português-inglês originais enriquecidas com as informações adicionais citadas anteriormente: forma superficial, lema, PoS e PoS+traços morfológicos. Essas sentenças fazem parte do *corpus de treinamento* usado na tradução automática estatística fatorada.

Português	**Inglês**
os|o|det|det.def.m.pl dentes|dente|n|n.m.pl do| de+o|pr+det|pr+det.def.m.sg mais|mais|adv| adv antigo|antigo|adj|adj.m.sg orangotango| orangotango|n|n.m.sg	the|the|det|det.def.sp teeth|tooth|n|n.pl of|of|pr| pr the|the|det|det.def.sp oldest|old|adj| adj.sint.sup orangutan|orangutan|n|n.sg
uma|um|det|det.ind.f.sg nova|novo|adj|adj.f.sg espécie|espécie|n|n.f.sg de|de|pr|pr hominídeo| hominídeo|n|n.m.sg encontrado|encontrar| vblex|vblex.pp.m.sg na|em+o|pr+det| pr+det.def.f.sg tailândia|tailândia|np|np.loc ,|,| cm|cm com|com|pr|pr estimados|estimar| vblex|vblex.pp.m.pl 12|12|num|num milhões| milhão|n|n.m.pl de|de|pr|pr anos|ano|n| n.m.pl ,|,|cm|cm tornou|tornar|vblex| vblex.ifi.p3.sg -|nc|nc|nc se|se|cnjadv|cnjadv o|o|det|det.def.m.sg parente|parente|n|n.m.sg mais|mais|adv|adv remoto|remoto|adj|adj.m.sg dos|de+o|pr+det|pr+det.def.m.pl atuais|atual| adj|adj.mf.pl orangotangos|orangotango|n| n.m.pl (|(|lpar|lpar pongo|pongar|v|v.pri.p1.sg pygmaeus|pygmaeus|nc|nc)|)|rpar|rpar .|.|sent| sent	a|a|det|det.ind.sg new|new|adj|adj.sint species| species|n|n.sg of|of|pr|pr hominid|hominid|n| n.sg found|find|vblex|vblex.pp in|in|pr|pr thailand|thailand|np|np.loc.sg ,|,|cm|cm with| with|pr|pr an|a|det|det.ind.sg estimated| estimated|adj|adj age|age|n|n.sg of|of|pr|pr 12| 12|num|num million|million| num|num.sp years|year|n|n.pl ,|,|cm|cm has| have|vbhaver|vbhaver.pri.p3.sg become| become|vblex|vblex.pp the+most|the+most| det+preadv|det.def.sp+preadv distant|distant| adj|adj relative|relative|n|n.sg of|of|pr|pr today|today|adv|adv 's|'s|gen|gen orangutans| orangutan|n|n.pl (|(|lpar|lpar pongo|pongo|n| n.sg pygmaeus|pygmaeus|nc|nc)|)|rpar|rpar .|.| sent|sent
um|um|det|det.ind.m.sg grupo|grupo|n|n.m.sg de|de|pr|pr pesquisadores|pesquisador|n| n.m.pl franceses|francês|adj|adj.m.pl ligados| ligar|vblex|vblex.pp.m.pl ao|a+o|pr+det| pr+det.def.m.sg laboratório|laboratório|n| n.m.sg europeu|europeu|adj|adj.m.sg de|de|pr|	a|a|det|det.ind.sg group|group|n|n.sg of|of|pr|pr french|french|adj|adj researchers|researcher|n| n.pl connected|connect|vblex|vblex.past with| with|pr|pr the|the|det|det.def.sp european| european|adj|adj synchrotron|synchrotron|n| n.sg radiation|radiation|n|n.sg facility|facility|

pr radiação|radiação|n|n.f.sg síncrotron| síncrotron|nc|nc (|(|lpar|lpar esrf|esrf|nc|nc)|)| rpar|rpar chegou|chegar|vblex|vblex.ifi.p3.sg a|a|pr|pr essa|esse|det|det.dem.f.sg conclusão| conclusão|n|n.f.sg comparando|comparar| vblex|vblex.ger os|o|prn|prn.pro.p3.m.pl 18| 18|num|num dentes|dente|n|n.m.pl do|de+o| pr+det|pr+det.def.m.sg fóssil|fóssil|n|n.m.sg com|com|pr|pr a|o|det|det.def.f.sg dentição| dentição|n|n.f.sg de|de|pr|pr outros|outro|det| det.ind.m.pl primatas|primata|n|n.m.pl antigos|antigo|adj|adj.m.pl .|.|sent|sent	n|n.sg (|(|lpar|lpar esrf|esrf|nc|nc)|)|rpar|rpar arrived|arrive|vblex|vblex.pp at|at|pr|pr this| this|det|det.dem.sg conclusion|conclusion|n| n.sg by|by|pr|pr comparing|compare|vblex| vblex.ger the|the|det|det.def.sp 18|18|num| num teeth|tooth|n|n.pl of|of|pr|pr the|the|det| det.def.sp fossil|fossil|n|n.sg with|with|pr|pr the|the|det|det.def.sp dentition|dentition|n|n.sg of|of|pr|pr other|other|det|det.ind.sp ancient| ancient|adj|adj primates|primate|n|n.pl .|.|sent| sent

A informação sintática foi separada dos demais fatores devido ao processamento realizado pelas ferramentas automáticas de análise sintática usadas no enriquecimento do *corpus*, as quais unem ou separam *tokens* durante o processo de análise. Por exemplo, o *parser* PALAVRAS separa a contração de preposição+artigo "*do*" em dois *tokens*: "*de*" e "*o*". Um pós-processamento da saída dos analisadores sintáticos poderá ser realizado no futuro para tentar contornar esses casos e, assim, agrupar todas as informações adicionais em um único arquivo. Contudo, vale ressaltar, que no momento tal processamento não se mostrou essencial. As informações sintáticas das sentenças na Tabela 2 são apresentadas na Tabela 4 na seguinte sequência de fatores: forma superficial, PoS e informação sintática referente à etiqueta mais interna na árvore sintática.

Tabela 4 – Sentenças português-inglês originais enriquecidas com as informações sintáticas. Essas sentenças fazem parte do *corpus de treinamento* usado na tradução automática estatística fatorada.

Português	**Inglês**
os|art|np dentes|n|np de|prp|pp o|art|np mais| adv|adjp antigo|adj|adjp orangotango|n|np	the|dt|np teeth|nns|np of|in|pp the|dt|np oldest| jjs|np orangutan|nn|np
uma|art|np nova|adj|np espécie|n|np de|prp|pp hominídeo|n|np encontrado|v-pcp|icl em|prp| pp a|art|np tailândia|prop|np ,|pu|nc com|prp| np estimados|v-pcp|np 12|num|np milhões|n| np de|prp|pp anos|n|np ,|pu|nc tornou|v-fin| fcl :|pu|nc se|conj-s|acl o|art|np parente|n|np mais|adv|adjp remoto|adj|adjp de|prp|pp os|art| np atuais|adj|np orangotangos|n|np (|pu|nc pongo|prop|np pygmaeus|adj|np)|pu|nc	a|dt|np new|jj|np species|nns|np of|in|pp hominid|nnp|np found|vbd|vp in|in|pp thailand|nnp|np ,|,|vp with|in|sbar an|dt|np estimated|vbn|np age|nn|np of|in|pp 12|cd|qp million|cd|qp years|nns|np ,|,|s has|vbz|vp become|vbn|vp the|dt|np most|rbs|np distant|jj| np relative|nn|np of|in|pp today|nn|np 's|pos|np orangutans|nnp|np (|nnp|np pongo|nnp|np pygmaeus|nnp|np)|nnp|np .|.|s
um|art|np grupo|n|np de|prp|pp pesquisadores| n|np franceses|adj|np ligados|v-pcp|icl a|prp| pp o|art|np	a|dt|np group|nn|np of|in|pp french|jj|np researchers|nns|np connected|vbn|vp with|in| pp the|dt|np european|nnp|np synchrotron|nnp|

laboratório_europeu_de_radiação_síncrotron|prop|np (|pu|nc esrf|prop|np)|pu|nc chegou|v-fin|fcl a|prp|pp essa|pron-dem|np conclusão|n|np comparando|v-ger|advp os|art|np 18|num|np dentes|n|np de|prp|pp o|art|np fóssil|n|np com|prp|pp a|art|np dentição|n|np de|prp|pp outros|pron-indef|np primatas|n|np antigos|adj|np	np radiation|nnp|np facility|nnp|np (|nnp|np esrf|nnp|np)|nnp|np arrived|vbd|vp at|in|pp this|dt|np conclusion|nn|np by|in|pp comparing|vbg|vp the|dt|np 18|cd|np teeth|nns|np of|in|pp the|dt|np fossil|nn|np with|in|pp the|dt|np dentition|nn|np of|in|pp other|jj|np ancient|jj|np primates|nns|np .|.|s

Além do *corpus* usado para treinamento dos modelos de tradução, outro *corpus* também foi necessário para o teste dos modelos treinados. O *corpus de teste* utilizado é composto de 667 sentenças paralelas pt-en. As sentenças desse *corpus* têm a mesma origem do *corpus* utilizado no treinamento, ou seja, também são provenientes de textos de edições da revista Pesquisa Fapesp. Porém, como esperado, trata-se de um conjunto inédito de sentenças e, portanto, diferente daquelas encontradas no *corpus* de treinamento.

Além dos *corpora* de treinamento e teste, os experimentos aqui relatados mencionam um outro *corpus* usado na avaliação automática das traduções geradas automaticamente: o *corpus* de referência. Esse *corpus*, nos experimentos aqui descritos, na verdade, trata-se do mesmo *corpus* usado para teste. Assim, por exemplo, ao traduzir do português para o inglês, as 667 sentenças em português são usadas como *corpus* de teste enquanto suas respectivas traduções coletadas da revista Pesquisa Fapesp são usadas como referência na comparação com a tradução em inglês gerada pelos modelos treinados. Os papéis se invertem quando o sentido da tradução vai do inglês (*corpus* de teste) para o português (*corpus* de referência).

3.2 Experimentos com tradução automática estatística baseada em frases

O Moses oferece a possibilidade de realizar experimentos de tradução com os textos na forma superficial (tradução estatística baseada em frases) e com informações adicionais (tradução fatorada), como já mencionado anteriormente. Essa seção descreve os experimentos realizados com os textos puros, ou seja, na forma superficial. A seção seguinte (3.3) descreve os experimentos realizados com a tradução fatorada.

Para a realização dos experimentos com tradução automática estatística foram adicionados os seguintes parâmetros à configuração padrão do treinamento do Moses, derivados de experimentos prévios envolvendo o português e o espanhol (Aziz et al., 2008):

```
-giza-option m1=5,m2=0,mh=5,m3=3,m4=3 \
-alignment grow-diag-final-and \
-max-phrase-length 7 \
-reordering msd-bidirectional-fe \
```

Essas configurações especificam os seguintes parâmetros:

- `-giza-option m1=5,m2=0,mh=5,m3=3,m4=3` – GIZA++ deve ser executado com 5 iterações do modelo IBM1, 3 dos modelos IBM3 e IBM4 (Brown et al., 1993) e 5 iterações de HMM (Vogel et al. 1996)[12].
- `-alignment grow-diag-final-and` – Especifica a opção de alinhamento que deve ser usada.
- `-max-phrase-length 7` – Opção que define 7 como tamanho máximo da frase a ser considerada na geração do modelo.
- `-reordering msd-bidirectional-fe` – Gera modelo de reordenação de *tokens*.

Assim, a linha de comando utilizada para treinamento de um modelo de tradução estatística tradicional baseado em frases, ou seja, em todos os experimentos apresentados nessa seção foi[13]:

```
% bin/moses-scripts/scripts-YYYYMMDD-HHMM/training/train-factored-phrase-model.perl \
-scripts-root-dir bin/moses-scripts/scripts-YYYYMMDD-HHMM/ \
-root-dir work \
-corpus work/corpus/novo.lowercased \
-f pt -e en \
-giza-option m1=5,m2=0,mh=5,m3=3,m4=3 \
-alignment grow-diag-final-and \
-max-phrase-length 7 \
-reordering msd-bidirectional-fe \
-lm 0:3:/work/lm/novo.lm:0
```

Vale dizer que as etapas de preparação do *corpus* para o treinamento, citadas na seção 2.2 (tokenização, conversão para minúsculas) deve ser realizada antes do comando apresentado acima.

Há também a opção de realizar um *tuning* (otimização), por meio do script oferecido pelo Moses, com o intuito de melhorar o desempenho do tradutor. Esse *tuning* consiste em adicionar mais um *corpus* de treinamento ao programa e realizar um novo treinamento. Esse processo é opcional e é realizado na tentativa de melhorar o desempenho do tradutor.

A etapas para a realização do *tuning* estão descritas nas linhas de comando a seguir:

[12] http://jedlik.phy.bme.hu/~gerjanos/HMM/node2.html

[13] Supõe-se aqui, que o modelo de língua para formas superficiais já foi gerado anteriormente (com SRILM) e encontra-se disponível em work/lm.

% mkdir work/tuning

```
% bin/moses-scripts/scripts-YYYYMMDD-HHMM/training/mert-moses.pl \
work/tuning/tuning.lowercased.pt \
work/tuning/tuning.lowercased.en \
moses-cmd/src/moses work/model/moses.ini \
--working-dir work/tuning/mert \
--rootdir bin/moses-scripts/scripts-YYYYMMDD-HHMM/ \
--decoder-flags "-v 0"
```

No comando acima é adicionado o *corpus* que contém os textos `fapesp.pb-ia.lowercased` em português e inglês para a realização do novo treinamento. Após a realização desse novo treinamento com *tuning* deve-se se executar o seguinte comando para substituir o arquivo moses.ini pelo com melhor desempenho encontrado.

```
% cp run2.moses.ini moses.ini
```

Nesta etapa, o arquivo gerado anteriormente moses.ini é substituído pelo novo arquivo gerado pelo *tuning* run2.moses.ini.

Os valores de BLEU e NIST alcançados com o treinamento e teste da tradução automática estatística baseada em frases *com* e *sem* o *tuning*, publicados em (Nunes & Caseli, 2009), estão apresentados na Tabela 5 a seguir.

Tabela 5 – Valores de BLEU e NIST para os experimentos com tradução automática estatística baseada em frases *com* e *sem* o *tuning* (Nunes & Caseli, 2009).

	BLEU	**NIST**
Sem *tuning*	0,3589	7,8312
Com *tuning*	0,3209	7,5745

Como já mencionado, os principais parâmetros usados nos experimentos descritos acima foram: 5 iterações dos modelos IBM-1 e HMM e 3 iterações dos modelos IBM-3 e IBM-4 para GIZA++, tamanho máximo de frase igual a 7 com a opção de reordenamento. Infelizmente, os resultados obtidos com esse experimento não foram os desejados, uma vez que esperava-se que o uso do *tuning* melhorasse a tradução o que não ocorreu. Contudo, acredita-se que esse pior desempenho do treinamento com o passo de otimização se deve ao tamanho limitado do *corpus* usado no *tuning* – apenas cerca de 46.000 palavras – o que

poderá ser comprovado com experimentos futuros que utilizem uma versão maior de tal *corpus*.

Além dos experimentos *com* e *sem* o *tuning*, foram realizados também experimentos diferentes variando o pré-processamento do *corpus* de treinamento e teste no que diz respeito à utilização de pontuação e a opção de conversão para letras minúsculas (*lowercase*). Para testar qual configuração ofereceria melhores resultados foram realizados quatro experimentos – E1, E2, E3, E4 – conforme sumarizado na Tabela 6.

Tabela 6 – Descrição dos experimentos (E1-E4) de acordo com o conteúdo dos *corpora* de treinamento e teste (Caseli & Nunes, 2009).

	Pontuação	**Letras minúsculas**
E1	NÃO	SIM
E2	NÃO	NÃO
E3	SIM	SIM
E4	SIM	NÃO

Na configuração E1 todos os sinais de pontuação foram removidos e todas as palavras foram passadas para letras minúsculas (*lowercase*). Na configuração E2 os sinais de pontuação também foram removidos mas opção *lowercase* não foi acionada. Na configuração E3 os sinais de pontuação continuaram nos textos e as palavras foram passadas para letras minúsculas. Por fim, nos experimentos com a configuração E4 os textos possuíam sinais de pontuação e a opção *lowercase* não foi acionada, ou seja, nenhum pré-processamento foi aplicado neste caso. Os resultados de pontuação BLEU e NIST para esses experimentos, conforme publicado em (Caseli & Nunes, 2009), estão na Tabela 7 a seguir.

Tabela 7 – Valores de BLEU e NIST para os experimentos com tradução estatística baseada em frases pt-en e en-pt de acordo com as configurações apresentadas na Tabela 6, aplicadas aos *corpora* de treinamento e teste (Caseli & Nunes, 2009).

	pt-en		**en-pt**	
	BLEU	**NIST**	**BLEU**	**NIST**
E1	0,3624	8,2096	0,3247	7,6220
E2	0,3523	8,0838	0,3095	7,4354
E3	**0,3903**	**8,3008**	**0,3589**	**7,8312**
E4	0,3826	8,1971	0,3485	7,6656

A melhor dessas configurações, E3, comparada à configuração sem nenhum pré-processamento (E4) gerou um ganho considerado estatisticamente significante apenas na

direção de tradução en-pt como verificado por meio de *bootstrapping* (Zhang et al., 2004). Na direção de tradução pt-en o ganho de performance em termos de BLEU e NIST de E3 quando comparado a E4 não foi considerado estatisticamente significante. Segundo Zhang et alli (2004), quando o conjunto de teste é pequeno (nos experimentos descritos nesse documento são usados apenas 667 segmentos paralelos pt-en) e há apenas uma referência, características dos experimentos aqui apresentados, então é preciso ser mais rigoroso na definição do intervalo de confiança. Esses autores também relatam que o número de amostras (*sample*) geradas no processo de *bootstrapping* deve ser maior do que 2000. Portanto, em todos os testes de significância estatística apresentados nesse documento e realizados com os scritps disponibilizados por esses autores[14], foram usados como parâmetros 99% de confiança e 10000 amostras.

A configuração E3 gerou um modelo de língua de ordem 3 (`../srilm/bin/i686/ngram-count -order 3`) para formas superficiais com entradas como as apresentadas a seguir para pt:

Trecho do modelo de língua de ordem 3 para formas superficiais

```
…
-3.826149   busca                   -0.2257756
-5.161891   buscado                 -0.1163499
-4.482761   buscam                  -0.1163499
-5.161891   buscamos                -0.1163499
-4.897127   buscando                -0.1163499
…
-3.453879   A busca                 -0.1104146
-0.992283   Bonagamba busca
-2.994005   Em busca                -0.2415103
-2.325608   Embraer busca
-2.76017    Essa busca
…
-1.177598   Hoje a busca
-1.389082   começar a busca
-2.75612    e a busca
-0.5772818  A busca da
-0.2941761  Em busca de
```

Nesse processo também é construída uma tabela de frases com tamanho máximo 7 (de acordo com o parâmetro de treinamento `-max-phrase-length 7`) com entradas como as apresentadas a seguir nas quais as frases (com no máximo 7 palavras) são seguidas de sua tradução, dos alinhamentos em ambos os sentidos e alguns valores definidos no treinamento.

[14] O script para realização de *bootstrapping* está disponível (data de acesso: 24/11/2009) em: http://projectile.sv.cmu.edu/research/public/tools/bootStrap/tutorial.htm

Trecho da tabela de frases com tamanho máximo 7

```
…
you seek to analyze a poem in ||| o senhor busca analisar o
poema em ||| (0,1) (2) (3) (3) (4) (5) (6) ||| (0) (0) (1) (2,3)
(4) (5) (6) ||| 1 3.1827e-06 1 6.07269e-06 2.718
you seek to analyze a poem ||| o senhor busca analisar o poema
||| (0,1) (2) (3) (3) (4) (5) ||| (0) (0) (1) (2,3) (4) (5) |||
1 4.95682e-06 1 1.85913e-05 2.718
you seek to analyze a ||| o senhor busca analisar o ||| (0,1)
(2) (3) (3) (4) ||| (0) (0) (1) (2,3) (4) ||| 1 6.6091e-06 1
1.85913e-05 2.718
you seek to analyze ||| o senhor busca analisar ||| (0,1) (2)
(3) (3) ||| (0) (0) (1) (2,3) ||| 1 0.000662241 1 0.00155662
2.718
you seek ||| o senhor busca ||| (0,1) (2) ||| (0) (0) (1) ||| 1
0.00856837 1 0.00490274 2.718
```

Uma tabela de reordenamento também é gerada (segundo o parâmetro `-reordering msd-bidirectional-fe`) conforme apresentado a seguir.

Trecho da tabela de reordenamento

```
…
you seek to analyze a poem in ||| o senhor busca analisar o
poema em ||| 0.6 0.2 0.2 0.6 0.2 0.2
you seek to analyze a poem ||| o senhor busca analisar o poema
||| 0.6 0.2 0.2 0.6 0.2 0.2
you seek to analyze a ||| o senhor busca analisar o ||| 0.6 0.2
0.2 0.6 0.2 0.2
you seek to analyze ||| o senhor busca analisar ||| 0.6 0.2 0.2
0.6 0.2 0.2
you seek ||| o senhor busca ||| 0.6 0.2 0.2 0.6 0.2 0.2
```

Foram testados também quais seriam os resultados em aplicar as mudanças E1, E2 e E3 apenas no *corpus* de treinamento, nesse caso nenhuma alteração foi realizada no *corpus* de teste. Essas novas configurações foram chamadas E1', E2' e E3' e os resultados – também publicados em (Caseli & Nunes, 2009), são apresentados na Tabela 8.

Tabela 8 – Valores de BLEU e NIST para os experimentos com tradução estatística baseada em frases pt-en e en-pt de acordo com as configurações E1-E3 apresentadas na Tabela 6, aplicadas apenas ao *corpus* de treinamento (Caseli & Nunes, 2009).

	pt-en		**en-pt**	
	BLEU	**NIST**	**BLEU**	**NIST**
E1'	0,3029	7,0192	0,2903	6,8974
E2'	0,2326	6,5197	0,3301	7,5923
E3'	0,3380	7,4885	0,2957	6,9309

Como se pode notar pelos valores das Tabelas 7 e 8, o pré-processamento deve ser aplicado em ambos os *corpora*: treinamento e teste. Além disso, o melhor desempenho foi obtido na configuração E3 na qual os caracteres de pontuação estão presentes no texto e as palavras são convertidas para letras minúsculas. A diferença de valores de BLEU e NIST encontrada entre E3 e E3' mostrou-se estatisticamente significante de acordo com *boostrapping* (99% de confiança e 10000 amostras) em ambos os sentidos de tradução (pt-en e en-pt).

A Tabela 9 a seguir apresenta exemplos de uma sentença original em português traduzida usando a melhor configuração, E3, e também usando a versão dessa configuração aplicada apenas no *corpus* de treinamento, ou seja, E3'.

Tabela 9 – Exemplos de sentença em português (fonte) traduzida para o inglês (saída) nos experimentos E3 e E3', bem como a sentença de referência correspondente (Caseli & Nunes, 2009).

E3	fonte	o centro de pesquisa de vacinas do instituto nacional de saúde dos estados unidos conseguiu a primeira vitória contra o vírus ebola .
	saída	the research center of vaccines of the national institutes of health of the united states , has managed to the first victory against the ebola virus .
	referência	the research center into vaccines of the national institute of health of the united states managed the first victory against the ebola virus .
E3'	fonte	O Centro de Pesquisa de Vacinas do Instituto Nacional de Saúde dos Estados Unidos conseguiu a primeira vitória contra o vírus Ebola .
	saída	O Centro of Pesquisa of Vacinas of the Instituto Nacional of Saúde of the Estados Unidos managed to the first victory against the virus Ebola .
	referência	The Research Center into Vaccines of the National Institute of Health of the United States managed the first victory against the Ebola virus .

Como se pode notar pelos valores da Tabela 9, a saída obtida por meio do experimento E3 difere ligeiramente da sentença de referência (apenas *tokens* diferentes, os que aparecem sublinhados). Por outro lado, a saída gerada automaticamente no experimento E3' levou a 14 *tokens* diferentes em relação à sentença de referência.

Além dos experimentos aqui descrito, em (Caseli & Nunes, 2009) também são apresentados resultados com o treinamento do *recase*. O uso de tal estratégia opcional, assim como no caso do *tuning*, não apresentou melhora no desempenho da tradução, mas sim uma piora nos valores de BLEU e NIST, sendo inclusive pior do que E3' como mostram os valores da Tabela 10.

Tabela 10 – Valores de BLEU e NIST para os experimentos com tradução estatística baseada em frases pt-en e en-pt para as configurações E3, E3' e a versão usando *recase* (Caseli & Nunes, 2009).

	pt-en		**en-pt**	
	BLEU	**NIST**	**BLEU**	**NIST**
E3	**0,3903**	**8,3008**	**0,3589**	**7,8312**
E3'	0,3380	7,4885	0,2957	6,9309
recase	0,3053	7,0531	0,2689	6,6323

Assim, ao final desses experimentos com a tradução automática estatística baseada em frases para o *corpus* português-inglês descrito na seção 3.1, pode-se concluir que a melhor configuração é a E3. Os valores de BLEU e NIST obtidos com essa configuração são, a partir de agora, considerados como *baseline* para os experimentos mais complexos realizados com a tradução fatorada, como descrito na próxima seção (3.3).

3.3 Experimentos com tradução fatorada

Além dos experimentos com a tradução automática estatística baseada em frase apresentados na seção anterior (3.2), o Moses também foi utilizado para realizar experimentos com a tradução fatorada português-inglês.[15] Como já mencionado, a tradução fatorada, proposta por Koehn e Hoang (2007), surgiu como uma alternativa para superar uma limitação dos modelos baseados em frase: a de mapear pequenas porções de texto sem o uso explícito de nenhuma informação linguística seja ela morfológica, sintática ou semântica.

Assim, a tradução fatorada surge como uma extensão da abordagem baseada em frases, na qual uma palavra não é representada apenas por sua forma superficial (a palavra da maneira como ocorre no texto), mas sim por um conjunto de fatores que representam diferentes níveis de anotação. Entre os possíveis fatores pode-se citar:

- forma superficial,
- lema,
- *part-of-speech* (PoS),
- atributos morfológicos (gênero, número etc.),
- classes de palavras geradas automaticamente,
- etiquetas sintáticas superficiais,
- assim como fatores dedicados a garantir concordância entre itens relacionados sintaticamente.

[15] Até onde se tem notícia, esses são os primeiros experimentos com tradução fatorada para o idioma português do Brasil.

A tradução de representações fatoradas de palavras de entrada em representações fatoradas de palavras de saída é dividida em uma sequência de passos de mapeamento que

a) *traduzem* fatores de entrada em fatores de saída ou

b) *geram* fatores de saída adicionais a partir de fatores de saída existentes.

A Figura 2 apresenta a representação gráfica de uma possível configuração de tradução fatorada na qual o processo de tradução é dividido em três passos de mapeamento:

1. a tradução dos lemas de entrada em lemas de saída,
2. a tradução dos fatores morfológicos e de PoS e
3. a geração das formas superficiais a partir do lema e dos fatores linguísticos.

Todos os passos de tradução operam no nível de frase enquanto os passos de geração operam no nível lexical.

Para que o treinamento e a geração possam ser realizados aproveitando todo o potencial dos fatores adicionais, nesse tipo de tradução estatística, os textos são enriquecidos com informações adicionais e o treinamento é realizado baseado nesse novo *corpus*. Nos experimentos descritos nesse documento, os *corpora* de treinamento e teste descritos na seção 3.1 foram enriquecidos com as informações adicionais:

- lema,
- categoria de *part-of-speech*,
- informações morfológicas,
- informação sintática.

Exemplos de pares de sentenças português-inglês enriquecidas com essas informações são apresentados nas Tabelas 3 e 4. Esses são os mesmos pares de sentenças apresentados na Tabela 2, porém, com fatores adicionais delimitados pelo caractere "|".

Os primeiros experimentos com a tradução automática estatística fatorada foram desenvolvidos com o intuito de investigar o impacto do uso de um modelo de língua adicional no processo de tradução. Assim, além do modelo de língua normalmente gerado e utilizado para formas superficiais, realizou-se experimentos também com a tradução fatorada usando modelos de língua adicionais gerados para:

- *part-of-speech* (fator 2[16] das sentenças na Tabela 3) – modelos de língua de ordem 3 (até 3-gramas – LMP3) e 7 (até 7-gramas – LMP7);
- *part-of-speech* + traços morfológicos (fator 3 das sentenças na Tabela 3) – modelos de língua de ordem 3 (até 3-gramas – LMPM3) e 7 (até 7-gramas – LMPM7)

[16] Vale lembrar que os fatores são enumerados a partir de 0 (forma superficial).

Por exemplo, para treinar o modelo de tradução na configuração LMP7 usou-se 2 modelos de língua gerados previamente – um para formas superficiais (novo.lm) gerado como explicado na seção 2.2 e outro para PoS (novo_pos.lm) gerado de modo similar porém usando um método de *smoothing* distinto (Witten-Bell ao invés de Kneser-Ney)[17]. Para treinar o modelo com a configuração LMPM7 usou-se comandos similares alterando apenas o arquivo de entrada (novo_pos+mor). A seguir são apresentadas as linhas de comando para geração dos modelos de língua de PoS e de PoS+informação morfológica, ambos de ordem 7, bem como trechos dos arquivos referentes aos modelos de língua gerados como saídas.

```
% ../srilm/bin/i686/ngram-count -order 7 \
-interpolate -wbdiscount -unk \
-text work/lm/novo_pos.lowercased.en \
-lm work/lm/novo_pos.lm
```

Trecho do modelo de língua de ordem 7 para PoS

```
…
-4.361742   ij                              -0.6172999
-2.216037   lpar                            -2.001122
-0.6565934  n                               -0.1977885
…
-3.901653   n abr                           0.03293782
-0.8022929  n adj                           -0.874305
-1.63751    n adv                           -0.7174007
…
-0.6061094  lpar abr nc                     -0.60206
-0.2057383  lpar abr rpar                   -0.2237071
-0.3050794  n abr nc                        -0.1628366
…
-0.03117357 adj nc abr nc                   -0.07213339
-0.03117357 lpar nc abr nc                  -0.60206
-0.1907173  n nc abr nc                     0.07532421
…
-0.01014458 n lpar nc abr nc                -0.4771212
-0.2782157  num n nc abr sent
-0.04267756 det nc nc abr nc
…
-0.001658415n cnjcoo n lpar abr rpar        -0.4771212
-0.0007099739     adj pr n lpar abr rpar
-0.000425845n pr n lpar abr rpar            -0.1766713
…
-0.0005521016     pr n cnjcoo n lpar abr rpar
-0.000236529n adj pr n lpar abr rpar
-0.0001064221     pr+det n pr n lpar abr rpar
```

Comando usado para gerar o modelo de língua de ordem 7 para PoS+informação morfológica:

[17]A opção por um outro método de *smoothing* foi tomada de acordo com as instruções apresentadas em http://www.speech.sri.com/projects/srilm/manpages/srilm-faq.7.html

```
% ../srilm/bin/i686/ngram-count -order 7 \
-interpolate -wbdiscount -unk \
-text work/lm/novo_pos.lowercased.en \
-lm work/lm/novo_pos+mor.lm
```

Trecho do modelo de língua de ordem 7 para PoS+informação morfológica

```
…
-1.563244   n.f.pl                                    -1.944814
-1.114645   n.f.sg                                    -2.2477
-4.148203   n.f.sp                                    -0.5528419
-1.443624   n.m.pl                                    -2.009864
-1.128995   n.m.sg                                    -2.237576
-3.266755   n.m.sp                                    -0.8589744
-2.651043   n.mf.pl                                   -1.217302
-2.725413   n.mf.sg                                   -1.136932
-3.764987   n.mf.sp                                   -0.7465522
…
-3.939476   n.f.sg np.loc                             -0.1391074
-3.055138   n.f.sg np.m.sg                            -0.3259065
-4.113434   n.f.sg np.m.sp
-3.73018    n.f.sg np.mf.sg                           -0.119927
…
-0.4074313  detnt abr.m.sg nc                         -0.4771213
-0.2708992  detnt abr.m.sg sent                       -0.1345246
-0.3246649  n.f.sg abr.m.sg nc                        -0.4771213
-0.1349109  n.m.sg abr.m.sg nc                        -0.1136769
…
-0.3649331  cnjcoo adj.f.pl adj.f.pl pr+det.def.f.sg  -0.4771212
-0.6266205  n.f.pl adj.f.pl adj.f.pl cm
-1.102147   n.f.pl adj.f.pl adj.f.pl cnjcoo
…
-0.2385238  n.f.sg pr+det.def.m.sg nc abr.m.sg nc     -0.4771212
-0.4323423  n.f.sg pr+det.def.m.sg nc abr.m.sg sent   -0.4771212
-0.3518102  n.m.sg pr+det.def.m.sg nc abr.m.sg adj.mf.sg
…
-0.00219789 n.m.sg pr n.f.sg lpar abr.f.sg rpar       -0.4771213
-0.01164672 n.f.sg pr nc+np.loc lpar abr.f.sg nc      -0.4771211
-0.5252308  n.acr.f.sg sent nc sent abr.f.sg rpar
…
-0.2221258  n.f.sg cm n.f.pl cnjcoo n.f.pl adj.f.pl
pr+det.def.f.sg
-0.2802625  n.f.sg cm n.f.sg cnjcoo n.f.pl adj.f.pl lpar
-0.7416462  n.f.sg cm n.f.sg cnjcoo n.f.pl adj.f.pl
pr+det.def.f.sg
```

Após o treinamento dos dois modelos de língua – um de formas superficiais e outro de PoS ou de PoS+informação morfológica – a seguir é apresentado o comando para treinamento do modelo de tradução. Vale lembrar que, nesse caso, o arquivo usado para treinamento deve conter os fatores separados por "|" como exemplos apresentados nas Tabelas 3 e 4; enquanto o arquivo de teste, nas configurações que apenas variam o modelo de língua, é o mesmo usado na tradução automática estatística baseada em frases, ou seja, composto por formas superficiais.

```
% bin/moses-scripts/scripts-YYYYMMDD-HHMM/training/train-factored-phrase-model.perl \
 -scripts-root-dir bin/moses-scripts/scripts-YYYYMMDD-HHMM/ \
 -root-dir work -corpus work/corpus/fatorada.lowercased \
 -f pt -e en \
 -giza-option m1=5,m2=0,mh=5,m3=3,m4=3 \
 -alignment grow-diag-final-and \
 -max-phrase-length 7 \
 -reordering msd-bidirectional-fe \
 -lm 0:3:/work/lm/novo.lm:0 \
 -lm 2:3:/work/lm/novo_pos.lm:0 \
 --translation-factors 0-0,2
```

O parâmetro `--translation-factors 0-0,2` na linha de comando de treinamento acima especifica que o fator de entrada para a tabela de tradução é forma superficial (0) e os fatores de saída são forma superficial (0) e PoS (2). No caso do treinamento com forma superficial (0) e PoS+informação morfológica (3), as últimas duas linhas do comando acima devem ser substituídas por: `--lm 3:7:/work/lm/novo_pos+mor.lm:0` e `--translation-factors 0-0,3`.

Os valores de BLEU e NIST para esses experimentos comparados com os valores obtidos para a tradução fatorada considerando-se apenas o modelo de formas superficiais, ou seja, a tradução automática estatística baseada em frases tradicional (E3 da Tabela 7), são apresentados na Tabela 11.

Tabela 11 – Valores de BLEU e NIST para os experimentos com tradução fatorada pt-en e en-pt com modelo de língua adicional para fatores de *part-of-speech* e traços morfológicos com base no *corpus de treinamento* composto por sentenças formatadas como as da Tabela 3.

	pt-en		**en-pt**	
	BLEU	**NIST**	**BLEU**	**NIST**
E3	0,3903	8,3008	0,3589	7,8312
LMP3	0,3912	8,3840	0,3572	7,7740
LMP7	0,3917	8,3643	0,3542	7,7500
LMPM3	0,3920	8,3916	0,3697	7,9661
LMPM7	**0,3923**	**8,3926**	**0,3703**	**7,9699**

Como é possível notar pelos valores da Tabela 11, o uso do modelo de língua de PoS (LMP3 e LMP7) trouxe uma melhora nos valores de BLEU e NIST para a tradução pt-en, mas não para a tradução en-pt, indicando que nesse último sentido de tradução apenas a

informação de PoS não tem impacto na escolha pela melhor tradução. Já o uso do modelo de língua de PoS com informação morfológica de ordem 7 (LMPM7) trouxe melhora nos valores de BLEU e NIST que se mostrou estatisticamente significante de acordo com *bootstrapping* (99% de confiança e 10000 amostras) apenas para o sentido de tradução en-pt. Tal significância de aumento de performance em termos de BLEU e NIST era esperada uma vez que o uso do modelo de língua adicional oferece mais informações para diferenciar as formas superficiais no momento da tradução para o português (a língua alvo nesse caso), língua que apresenta mais variações morfológicas do que o inglês.

Considerando-se que as sentenças usadas no treinamento com o *corpus* etiquetado com informação sintática (veja exemplos na Tabela 4) diferem um pouco das sentenças sem esses fatores (veja exemplos na Tabela 3), para permitir a comparação, os valores de BLEU e NIST para a tradução fatorada com modelos de língua adicionais para os fatores sintáticos foram calculados separadamente. Assim, além do modelo de língua normalmente gerado e utilizado para formas superficiais, realizou-se experimentos também com a tradução fatorada usando modelos de língua adicionais gerados para:

- *part-of-speech* (fator 1 das sentenças na Tabela 4) – modelos de língua de ordem 3 (até 3-gramas – LMP'3) e 7 (até 7-gramas – LMP'7);
- etiquetas sintáticas (fator 2 das sentenças na Tabela 4) – modelos de língua de ordem 3 (até 3-gramas – LMS3) e 7 (até 7-gramas – LMS7);

Os valores de BLEU e NIST para esses experimentos foram comparados com os valores obtidos para a tradução fatorada considerando-se apenas o modelo de formas superficiais, denominada nesse caso de *baseline* (trata-se do treinamento seguindo a configuração E3, mas desta vez com base no *corpus* "modificado" pelos *parsers*), são apresentados na Tabela 12.

Tabela 12 – Valores de BLEU e NIST para os experimentos com tradução fatorada pt-en e en-pt com modelo de língua adicional para fatores de *part-of-speech* e etiquetas sintáticas com base no *corpus de treinamento* composto por sentenças formatadas como as da Tabela 4.

	pt-en		**en-pt**	
	BLEU	**NIST**	**BLEU**	**NIST**
Baseline	0,3646	7,9542	**0,3722**	**7,9424**
LMP'3	**0,3690**	**7,9898**	0,3658	7,8337
LMP'7	0,3693	7,9637	0,3636	7,7902
LMS3	0,3592	7,8858	0,3569	7,7422
LMS7	0,3609	7,9100	0,3507	7,6667

Como apresentado na Tabela 11, os modelos de língua adicionais com PoS e informação morfológica trouxeram melhora nos valores de BLEU e NIST, porém, o mesmo não pode ser dito a respeito do uso de etiquetas sintáticas como demonstram os valores da Tabela 12. Aqui é possível notar que o modelo de língua de PoS aumenta um pouco os valores das métricas apenas na direção pt-en e o uso de informação sintática denegriu a performance em termos de BLEU e NIST. Vale dizer que as informações sintáticas usadas aqui são apenas superficiais (etiqueta do nó mais interno próximo ao nó-folha que representa a forma superficial em questão) e, portanto, há muito ainda o que ser explorado a esse respeito.

Em seguida, foram projetadas algumas configurações para treinar a geração de modelos de tradução fatorada com vários passos de tradução e geração. Seguindo as configurações apresentadas em (Koehn et al., 2007b), o próximo experimento foi realizado com os mesmos fatores (superficiais e PoS+informação morfológica), porém ao invés de mapear da forma superficial fonte diretamente na forma superficial e PoS+informação morfológica alvo, esse processo foi quebrado em dois passos de mapeamento:

- um passo de tradução que mapeia apenas formas superficiais em formas superficiais (`--translation-factors 0-0`) e
- um passo de geração que gera as categorias de *part-of-speech* + informações morfológicas (3) a partir das formas superficiais (0) no lado alvo (`--generation-factors 0-3`)

Esse treinamento é realizado por meio da linha de comando:

```
% bin/moses-scripts/scripts-YYYYMMDD-HHMM/training/train-factored-phrase-model.perl \
  -scripts-root-dir bin/moses-scripts/scripts-YYYYMMDD-HHMM/ \
  -root-dir work \
  -corpus work/corpus/fatorada.lowercased \
  -f pt -e en \
  -giza-option m1=5,m2=0,mh=5,m3=3,m4=3 \
  -alignment grow-diag-final-and \
  -max-phrase-length 7 \
  -reordering msd-bidirectional-fe \
  -lm 0:3:/work/lm/novo.lm:0 \
  -lm 3:7:/work/lm/novo_pos+mor.lm:0 \
  --translation-factors 0-0 \
  --generation-factors 0-3 \
  --decoding-steps t0,g0
```

Veja que, no comando acima, é necessário especificar em qual ordem os passos de mapeamento serão aplicados com (`--decoding-steps t0,g0`). Agora, além de uma tabela de frases é construída, também, uma tabela de geração contendo, por exemplo, entradas como as apresentadas a seguir:

Trecho da tabela de geração do fator 0 (formas superficiais) para o 3 (PoS+informação morfológica)

```
…
desaparece vblex.pri.p3.sg 1.0000000  0.0005908
gaviões n.m.pl 1.0000000  0.0000548
esclarecedor adj.m.sg 1.0000000  0.0001342
recobre vblex.prs.p3.sg 1.0000000  0.0043764
diagnosticada v.pp 1.0000000  0.0015949
sexta adj.f.sg 1.0000000  0.0002951
…
busca vblex.pri.p3.sg 0.1956522  0.0017724
busca n.f.sg 0.8043478  0.0019027
```

Entre as entrada tem-se, por exemplo, palavras que só ocorrem com uma combinação de PoS e traços morfológicos em todo o *corpus* de treinamento (as 6 primeiras entradas para as quais a probabilidade é 1.0). Outras, como “busca”, por exemplo, são mais frequentes com uma combinação de PoS e traços morfológicos (neste caso, “busca” como substantivo, feminino, singular) do que com outras (por exemplo, “busca” como verbo).

A Tabela 13 apresenta os valores de BLEU e NIST para a tradução automática após treinamento com a configuração apresentada acima (denominada aqui como LMPM7GT). Como se pode notar pelos valores dessa tabela, não houve melhora nos valores das medidas usando a nova configuração (LMPM7GT) que separa os passos de tradução e geração quando comparada à configuração na qual essa separação não ocorre (LMPM7).

Tabela 13 – Valores de BLEU e NIST para os experimentos com tradução fatorada pt-en e en-pt com modelo de língua adicional para fatores de *part-of-speech* e traços morfológicos com base no *corpus de treinamento* composto por sentenças formatadas como as da Tabela 3. Nesse experimento foram usados passos de tradução e geração.

	pt-en		**en-pt**	
	BLEU	**NIST**	**BLEU**	**NIST**
E3	0,3903	8,3008	0,3589	7,8312
LMPM7	**0,3923**	**8,3926**	**0,3703**	**7,9699**
LMPM7GT	0,3856	8,3514	0,3665	7,9418

Como relatado por (Koehn et al., 2007b), não faz muito sentido traduzir de modo diferente formas superficiais com mesmo lema, como “cantado” e “cantei”, sendo ainda pior

quando apenas uma palavra é encontrada no *corpus* de treinamento não sendo possível traduzir a outra. Os modelos de tradução fatorada permitem que sejam criados modelos que fazem análise morfológica e a decomposição durante o processo de tradução. Para tanto, esses autores propõe o seguinte treinamento:

```
% bin/moses-scripts/scripts-YYYYMMDD-HHMM/training/train-factored-
phrase-model.perl \
 -scripts-root-dir bin/moses-scripts/scripts-YYYYMMDD-HHMM/ \
 -root-dir work \
 -corpus work/corpus/fatorada.lowercased \
 -f pt -e en \
 -giza-option m1=5,m2=0,mh=5,m3=3,m4=3 \
 -alignment grow-diag-final-and \
 -max-phrase-length 7 \
 -reordering msd-bidirectional-fe \
 -lm 0:3:/work/lm/novo.lm:0 \
 -lm 3:7:/work/lm/novo_pos+mor.lm:0 \
 --translation-factors 1-1+3-3 \
 --generation-factors 1-3+1,3-0 \
 --decoding-steps t0,g0,t1,g1
```

Neste treinamento, estão especificados quatro passos de mapeamento, dois de tradução e dois de geração conforme os parâmetros:

- `--translation-factors 1-1+3-3` especifica
  - um passo de tradução entre lemas (1-1)
  - um passo de tradução entre Pos + informações morfológicas (3-3)
- `--generation-factors 1-3+1,3-0` diz que haverá
  - um passo de geração que atribui possíveis PoS + informação morfológica a lemas (1-3)
  - um passo de geração que mapeia PoS + informação morfológica e lema em formas superficiais (1,3-0)

A Tabela 14 apresenta os valores de BLEU e NIST para a tradução automática após treinamento com a configuração apresentada acima (denominada aqui como LMPM7MA). Como se pode notar pelos valores dessa tabela, não houve melhora nos valores das medidas

usando a nova configuração (LMPM7MA) que faz a análise morfológica durante a tradução, quando comparada à configuração na qual essa análise não ocorre (LMPM7).

Tabela 14 – Valores de BLEU e NIST para os experimentos com tradução fatorada pt-en e en-pt com modelo de língua adicional para fatores de *part-of-speech* e traços morfológicos com base no *corpus de treinamento* composto por sentenças formatadas como as da Tabela 3. Nesse experimento realiza-se a análise morfológica durante a tradução.

	pt-en		**en-pt**	
	BLEU	**NIST**	**BLEU**	**NIST**
E3	0,3903	8,3008	0,3589	7,8312
LMPM7	**0,3923**	**8,3926**	**0,3703**	**7,9699**
LMPM7MA	0,3790	8,4106	0,3444	7,7256

A decomposição do processo de tradução em análise morfológica e geração, segundo Koehn et alli (2007b), torna o processo mais robusto. Porém, também relatam a importância de se considerar o fato de uma frase de formas superficiais ter sido vista anteriormente. Afirmam, ainda, que o modelo anterior tem um desempenho precário pois considera apenas formas superficiais confiando em suas propriedades. Na prática, consideram que é melhor utilizar passos de decodificação: um para a análise morfológica e a geração do modelo anterior e outro para o mapeamento direto de formas superficiais (e PoS + informação morfológica do modelo de língua aplicado até então). Esse treinamento é realizado por meio do comando:

```
% bin/moses-scripts/scripts-YYYYMMDD-HHMM/training/train-factored-phrase-model.perl \
	-scripts-root-dir bin/moses-scripts/scripts-YYYYMMDD-HHMM/ \
	-root-dir work \
	-corpus work/corpus/fatorada.lowercased \
	-f pt -e en \
	-giza-option m1=5,m2=0,mh=5,m3=3,m4=3 \
	-alignment grow-diag-final-and \
	-max-phrase-length 7 \
	-reordering msd-bidirectional-fe \
	-lm 0:3:/work/lm/novo.lm:0 \
	-lm 3:7:/work/lm/novo_pos+mor.lm:0 \
	--translation-factors 1-1+3-3+0-0,3 \
	--generation-factors 1-3+1,3-0 \
	--decoding-steps t0,g0,t1,g1:t2
```

Neste treinamento, estão especificados quatro passos de mapeamento, dois de tradução e dois de geração conforme os parâmetros:

- `--translation-factors 1-1+3-3+0-0,3` especifica
  - um passo de tradução entre lemas (1-1)
  - um passo de tradução entre Pos + informações morfológicas (3-3)
  - um passo de tradução de formas superficiais (0-0,3) com mapeamento direto de formas superficiais em formas superficiais e Pos + informações morfológicas
- `--generation-factors 1-3+1,3-0`, assim como no experimento anterior, diz que haverá
  - um passo de geração que atribui possíveis PoS + informação morfológica a lemas (1-3)
  - um passo de geração que mapeia PoS + informação morfológica e lema em formas superficiais (1,3-0)

Agora, uma tabela de tradução adicional é criada (0-0,3) e usada como um passo de decodificação diferente (:t2). A Tabela 15 apresenta os valores de BLEU e NIST para a tradução automática após treinamento com essa nova configuração com passo extra de decodificação (denominada aqui como LMPM7DE). O uso dessa nova configuração (LMPM7DE) trouxe melhora nos valores de BLEU e NIST, quando comparada à melhor configuração até então (LMPM7), porém tal melhora melhora não se mostrou estatisticamente significante de acordo com o *bootstrapping* (com 99% de confiança e 10000 amostras).

Tabela 15 – Valores de BLEU e NIST para os experimentos com tradução fatorada pt-en e en-pt com modelo de língua adicional para fatores de *part-of-speech* e traços morfológicos com base no *corpus de treinamento* composto por sentenças formatadas como as da Tabela 3. Nesse experimento realiza-se a análise morfológica durante a tradução com um passo extra de decodificação baseado em formas superficiais.

	pt-en		**en-pt**	
	BLEU	**NIST**	**BLEU**	**NIST**
E3	0,3903	8,3008	0,3589	7,8312
LMPM7	0,3923	8,3926	0,3703	7,9699
LMPM7DE	**0,3932**	**8,4421**	**0,3713**	**7,9813**

A partir do que foi apresentado nessa seção é possível notar que a tradução fatorada tem potencial para melhorar o desempenho da tradução baseada em frases tradicional e novos

experimentos serão projetados para verificar como explorar a variedade de configurações ainda não testadas.

4 Conclusão

Esse relatório apresentou o processo de instalação e utilização do *toolkit* de código-aberto Moses para a tradução automática estatística, bem como os resultados de vários experimentos realizados para a tradução de textos em português e em inglês. Os experimentos realizados e apresentados neste documento referem-se à tradução automática estatística baseada em frases tradicional, considerada o estado da arte atualmente, e a tradução estatística fatorada, uma extensão da tradução baseada em frases na qual uma palavra não é representada apenas por sua forma superficial, mas sim por um conjunto de fatores.

Os experimentos resultaram em uma configuração de tradução automática estatística baseada em frases considerada como *baseline* para experimentos com modelos de tradução mais complexos. Esse *baseline* foi obtido considerando-se o pré-processamento dos textos de treinamento e teste com a conversão de todas as letras para minúsculas e a conservação dos caracteres de pontuação originais, além de parâmetros de treinamento referentes ao alinhamento lexical, reordenamento de frases entre outros.

Os experimentos com a tradução fatorada, projetados com o intuito de verificar se a mesma traria um ganho no desempenho da tradução em termos de BLEU e NIST mostrou que, para pt-en, enquanto a versão tradicional baseada em frases alcança valores de 0,3903 (BLEU) e 8,3008 (NIST), a fatorada (última configuração apresentada) alcança 0,3932 (BLEU) e 8,4421 (NIST) uma melhora que se não mostrou estatisticamente significante de acordo com *bootstrapping* (Zhang et al., 2004), com 99% de confiança e 10000 amostras. No sentido de tradução en-pt, essa diferença fica ainda mais clara uma vez que a tradução baseada em frases tradicional alcança 0,3589 (BLEU) e 7,8312 (NIST) enquanto a versão fatorada (última configuração apresentada) alcança 0,3713 (BLEU) e 7,9813 (NIST).

Novos experimentos deverão ser projetados e levados a cabo para verificar ainda mais o impacto do uso de fatores adicionais na tradução fatorada, principalmente no que diz respeito à combinação de fatores sintáticos com outros como os traços morfológicos.

Agradecimentos

Agradecemos ao Programa Integrado de Apoio ao Docente Recém-Doutor (PIADRD) da UFSCar pelo apoio financeiro sem o qual não seria possível realizar os experimentos aqui apresentados. Agradecemos também ao mestrando Bruno Akio por seus comentários a respeito do processo de instalação do Moses.

5 Referências Bibliográficas

Alshawi, H.; Bangalore, S.; Douglas, S. (1998). Automatic acquisition of hierarchical transduction models for machine translation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics.*

Aziz, W.F.; Pardo, T.A.S.; Paraboni, I. (2008). An Experiment in Spanish-Portuguese Statistical Machine Translation. In t*he Proceedings of the 19th Brazilian Symposium on Artificial Intelligence - SBIA (Lecture Notes in Computer Science 5249)*, pp. 248-257. Salvador-BA, Brazil. October, 26-30.

Bick, E. (2000). *The Parsing System "Palavras": Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework*. PhD thesis – Aarhus University. Aarhus, Denmark: Aarhus University Press, November 2000.

Brown, P. F.; Della Pietra, V. J.; Della Pietra, S. A.; Mercer , R. L. (1993). The Mathematics of Statistical Machine Translation: Parameter Estimation. *Computational Linguistics*, 19:263–311.

Caseli, H. M. (2007). *Indução de léxicos bilíngües e regras para a tradução automática*. 158 p. Tese (Doutorado) – ICMC-USP, Abril 2007.

Caseli, H.M.; Nunes, I.A. (2009). Statistical Machine Translation: little changes big impacts. In *Proceedings of the 7th Brazilian Symposium in Information and Human Language Technology.*, pp. 1-9.

Collins, M. *Head-driven statistical models for natural language parsing*. PhD thesis – University of Pennsylvania, 1999.

Doddington, G. (2002). Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proceedings of ARPA Workshop on Human Language Technology*. San Diego. pp. 128-132.

Galley, M.; Graehl, J.; Knight, K.; Marcu, D.; Deneefe, S.; Wang, W.; Thayer, I. (2006). Scalable inference and training of context-rich syntactic translation models. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pp. 961-968, Sydney, Australia.

Koehn, P.; Och, F. J.; Marcu, D. (2003). Statistical phrase-based translation. In *Proceedings of the Human Language Technology (HLT/NAACL 2003)*, pp. 127–133.

Koehn, P.; Hoang, H. (2007). Factored translation models. In *Proceedings of EMNLP-2007*, pp. 868–876, Prague.

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. (2007a). Moses: Open Source Toolkit for Statistical Machine Translation. In *Proceedings of ACL 2007*, pp. 177–180, Prague, Czech Republic.

Koehn, P.; Federico, M.; Shen, W.; Bertoldi, N.; Bojar, O.; Callison-Burch, C.; Cowan, B.; Dyer, C.; Hoang, H.; Zens, R.; Constantin, A.; Moran, C. C.; Herbst, E. (2007b). *Open Source Toolkit for Statistical Machine Translation: Factored Translation Models and Confusion Network Decoding*. Technical Report – Johns Hopkins University – Center for Speech and Language Processing, September 2007. 102 p.

Koerner, E. F. K.; Asher, R. E. (1995). *Concise history of the language sciences: from the Sumerians to the cognitivists*, pp. 431–445. Pergamon Press, Oxford.

Lee, Y. S. (2004). Morphological analysis for statistical machine translation. In *Proceedings of HLT/NAACL-2004.*

Lopez, A. (2008). Statistical machine translation. *ACM Comput. Surveys.* Vol. 40, No. 3, p. 1-49. August 2008. DOI= http://doi.acm.org/10.1145/1380584.1380586

Melamed, I. D. (2004). Statistical machine translation by parsing. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, Main Volume. Barcelona, Spain, 2004. pp. 653-660.

Nunes, I.A.; Caseli, H. M. (2009). Primeiros Experimentos na Investigação e Avaliacão da Tradução Automática Estatística Inglês-Português. In *Anais do I Workshop de Iniciação Científica em Tecnologia da Informação e da Linguagem Humana (TILic).* pp. 1-4.

Och, F. J. (2003). Minimum error rate training for statistical machine translation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.

Och, F. J.; Ney, H. (2000). Improved statistical alignment models. In *Proceedings of ACL-2000*. Hong Kong, China, 2000. pp. 440-447.

Och, F. J.; Ney, H. (2004). The Alignment Template Approach to Statistical Machine Translation. *Computational Linguistics*, 30(4):417–449.

Och, F. J.; Gildea, D.; Khudanpur, S.; Sarkar, A.; Yamada, K.; Fraser, A.; Kumar, S.; Shen L.; Smith, D.; Eng, K.; Jain, V.; Jin, Z.; Radev, D. (2004). A smorgasbord of features for statistical machine translation. In *Proceedings of HLT/NAACL-2004.*

Papineni, K.; Roukos, S.; Ward, T.; Zhu, W. (2002). BLEU: a method for automatic evaluation of machine translation. In *Proceedings of ACL-2002*. Philadelphia, PA. pp. 311-318.

Sadat, F.; Habash, N. (2006). Combination of arabic pre-processing schemes for statistical machine translation. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*. Sydney, Australia, 2006. pp. 1-8.

Vogel, S.; Ney, H.; Tillmann, C. (1996). HMM-based word alignment statistical translation. In *Proceedings of COLING 1996*, Copenhagen, pp. 836–841.

Wu, D. (1997). Stochastic inversion transduction grammars and bilingual parsing of parallel corpora. *Computational Linguistics*, v. 23, n. 3, 1997.

Yamada, K.; Knight, K. (2001). A syntax-based statistical translation model. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL-2001).* Toulouse, France, 2001. pp. 1–8.

Zhang, Y.; Vogel, S.; Waibel, A. (2004). Interpreting Bleu/NIST scores: How much improvement do we need to have a better system?. In *Proceedings of LREC 2004.* Lisbon, Portugal, May 2004.