UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE INFORMÁTICA
CURSO DE BACHARELADO EM CIÊNCIA DA COMPUTAÇÃO

DANIEL EMILIO BECK

Aprimorando o tratamento de Expressões Multipalavras em um tradutor automático baseado em regras

Trabalho de Conclusão apresentado como
requisito parcial para a obtenção do grau de
Bacharel em Ciência da Computação

Profa. Dra. Aline Villavicencio
Orientador

Porto Alegre, dezembro de 2009

CIP – CATALOGAÇÃO NA PUBLICAÇÃO

Beck, Daniel Emilio

Aprimorando o tratamento de Expressões Multipalavras em um tradutor automático baseado em regras / Daniel Emilio Beck. – Porto Alegre: Graduação em Ciência da Computação da UFRGS, 2009.

45 f.: il.

Trabalho de Conclusão (bacharelado) – Universidade Federal do Rio Grande do Sul. Curso de Bacharelado em Ciência da Computação, Porto Alegre, BR–RS, 2009. Orientador: Aline Villavicencio.

1. Processamento da linguagem natural. 2. Linguística computacional. 3. Expressões Multipalavras. 4. Tradução automática. I. Villavicencio, Aline. II. Título.

“Nada na vida deve ser temido, apenas compreendido”
— MARIE CURIE

AGRADECIMENTOS

Primeiramente, gostaria de agradecer à minha mãe, Maria Teresa pelo amor incondicional e por demonstrar esse amor tanto nas pequenas quanto nas grandes atitudes, aos meus irmãos, Dary e Diego, pelo amor e por me mostrarem que eu devo sempre seguir os meus sonhos, e ao meu pai, Dirceu, que nunca deixou que me faltasse nada durante esses anos.

Agradeço também às minhas tias, tios, primas, primos e todos os parentes que sempre me apoiaram e não deixaram faltar carinho durante a minha vida.

Também quero agradecer a todos meus amigos que me acompanharam e que sempre foram tão fundamentais para que pudesse estar feliz. Todos são importantes para mim, mas alguns deles eu não posso deixar de citar os nomes: Diego, Marcos, Fábio, Kao, Paula e Joana.

Agradeço à UFRGS que, com seus professores e funcionários, me fizeram acreditar que é possível haver educação pública e gratuita de qualidade. Quero agradecer especialmente àqueles professores que são apaixonados pela sua profissão e fizeram questão de trazer essa paixão durante as disciplinas que cursei. Vocês foram fundamentais para que eu escolhesse seguir na área acadêmica, mesmo com todas as adversidades que ela contém.

Desejo também um agradecimento especial à minha orientadora, Profa. Aline Villavicencio, por ter me apresentado a área de Processamento da Linguagem Natural e pela motivação e apoio demonstrado durante minha trajetória, e também ao meu colega e amigo Carlos Eduardo Ramisch, pelas longas conversas e reflexões sobre a vida acadêmica e pela ajuda inestimável dedicada a esse e outros trabalhos.

Finalmente, gostaria de agradecer à comunidade *Open Source* do projeto Apertium, em especial Jimmy O'Regan e Francis Tyers, pela oportunidade de contribuição e pelo auxílio dado durante a execução desse trabalho.

SUMÁRIO

LISTA DE ABREVIATURAS E SIGLAS

ACL	*Association for Computational Linguistics*
BLEU	*Bilingual Evaluation Understudy*
CIDE	*Cambridge International Dictionary of English*
DARPA	*Defense Advanced Research Projects Agency*
EMP	Expressão Multipalavra
FAHQUT	*Fully Automated High Quality Translation of Unrestricted Text*
HTML	*Hypertext Markup Language*
IC	Intervalo de Confiança
NIST	*National Institute of Standards and Technology*
PER	*Position-independent Word Error Rate*
PLN	Processamento da Linguagem Natural
RTF	*Rich Text Format*
TA	Tradução Automática
WER	*Word Error Rate*
XML	*Extensible Markup Language*

LISTA DE FIGURAS

LISTA DE TABELAS

RESUMO

A Tradução Automática (TA) é uma das aplicações mais clássicas do Processamento da Língua Natural (PLN). As pesquisas nessa área geraram uma série de ferramentas e sistemas usados largamente hoje em dia. No entanto, na maior parte das aplicações, as traduções geradas por esses sistemas não são da qualidade exigida, sendo necessário um processo posterior de revisão.

Uma das formas de melhorar os sistemas de TA é utilizar recursos linguísticos, como dicionários e ontologias. A idéia deste trabalho é aprimorar um sistema existente adicionando um dicionário de Expressões Multipalavras (EMPs), um fenômeno linguístico de grande importância para aplicações de PLN.

Neste trabalho foi proposta uma abordagem para o tratamento de EMPs em um sistema de Tradução Automática, o Apertium, através da representação adequada das mesmas. O impacto causado pela adoção desta abordagem foi avaliado no contexto de TA com o par de línguas Inglês-Espanhol.

Palavras-chave: Processamento da linguagem natural, linguística computacional, Expressões Multipalavras, tradução automática.

Improving the Multiword Expression treatment in a rule-based machine translator

ABSTRACT

Machine Translation (MT) is one of the most classic Natural Language Processing (NLP) applications. Research in this area generated a series of tools and systems widely used today. However, in most applications, the translations generated by these systems do not have the demanded quality, requiring a subsequent review process.

One way to improve MT systems is to use linguistic resources, such as dictionaries and ontologies. The idea of this work is to improve an existing system by adding a dictionary of Multiword Expressions (MWEs), a linguistic phenomenon of great importance for NLP applications.

In this work, we propose an approach for the EMPs treatment in a Machine Translation system, Apertium, through appropriate representation of the same. The impact caused by the adoption of this approach was evaluated in the MT context with the language pair English-Spanish.

Keywords: natural language processing, computational linguistics, Multiword Expressions, machine translation.

1 INTRODUÇÃO

O Processamento da Linguagem Natural (PLN) é a área da Ciência da Computação que visa tornar os computadores capazes de compreender e se comunicar utilizando a linguagem humana. Essa difere das linguagens de máquina essencialmente pela sua ambiguidade e pela sua constante evolução. Dessa forma, sistemas de PLN devem levar em conta esses aspectos no momento da sua concepção, utilizando para isso, conceitos e técnicas derivadas tanto da Computação quanto da Linguística. Exemplos desses sistemas incluem tradutores e sumarizadores automáticos, geradores de frases e sistemas de processamento da fala.

As diferenças existentes entre a linguagem humana e a de máquina geram uma série de dificuldades para seu processamento computacional. Para lidar com elas, são necessários recursos linguísticos que reúnam as informações essenciais para o correto tratamento do texto. Esses recursos podem ser estruturados, como dicionários e ontologias, ou não, como os corpora (grandes coleções de textos). Um dos papéis da Computação está justamente em encontrar formas eficientes e eficazes de construir e utilizar esses recursos, buscando para isso, técnicas e conceitos de áreas como Inteligência Artificial e Bancos de Dados.

Sistemas de Tradução Automática (TA) são o foco desse trabalho. A TA é umas das áreas mais clássicas e visadas em PLN, tendo motivado o desenvolvimento de muitas das técnicas e ferramentas existentes. O estado-da-arte em TA inclui softwares como o Apertium, o Google Translator e o Systran. Ainda que eles sejam utilizados comercialmente hoje em dia, na grande maioria dos casos eles não fornecem uma tradução de alta qualidade, dependendo da tarefa de tradução e do contexto de utilização. Dessa forma, na maioria dos casos é necessária uma revisão posterior por parte de um especialista. Sendo assim, há muito espaço para melhorias nos sistemas atuais, de forma a aumentar a utilidade de uma tradução automática e, com isso, minimizar o esforço e o tempo de pós-edição.

O objetivo desse trabalho é o aprimoramento de um sistema de TA. A idéia é melhorá-lo adicionando um novo recurso linguístico destinado ao tratamento de Expressões Multipalavras (EMPs), conjuntos de palavras com composicionalidade limitada. Expressões como *chutar o balde*, *colher de pau* e *telefone ocupado* são exemplos de EMPs. Elas são um fenômeno linguístico de extrema importância para qualquer sistema de PLN, já que estima-se que em torno de 50% do léxico de uma língua seja composto por expressões desse tipo (JAC 97). No entanto, o tratamento delas dentro de um sistema de TA é difícil, principalmente devido à natureza heterogênea e extremamente flexível dessas expressões.

Neste trabalho, o sistema a ser aprimorado será o Apertium (ARM 2007). Por ser um software de código aberto, ele tem sido utilizado em vários trabalhos de pesquisa. Além disso, é um sistema que possui a necessidade de melhorias no tratamento de EMPs, ainda

que consiga traduzir corretamente alguns tipos específicos delas. Uma das dificuldades do Apertium é tratar casos em que uma EMP possui flexões em mais de uma de suas componentes (como por exemplo, a expressão em espanhol *dirección general*, que no plural fica *direcciones generales*). O foco será estender o tratamento de EMPs para que levem em conta esse tipo de expressão.

O conteúdo do trabalho está estruturado da seguinte forma:

- O Capítulo 2 visa a explicar melhor o processo de TA, exibindo suas dificuldades específicas. Também são detalhados a classificação dos sistemas de TA e as métricas utilizadas na avaliação desses sistemas.
- O Capítulo 3 define mais especificamente as EMPs e suas características, além de conter uma classificação mais detalhada de seus variados tipos.
- O Capítulo 4 contém uma explicação detalhada do sistema Apertium, de seu tratamento de EMPs e do aprimoramento realizado.
- O Capítulo 5 consiste em experimentos realizados que visam a investigar e validar a melhoria realizada dentro do Apertium.
- O Capítulo 6 discute sobre as conclusões e as possibilidades de trabalhos futuros.

Este trabalho foi realizado no contexto de uma das idéias propostas para participação no evento *Google Summer of Code 2009*[1] e contou com a colaboração da comunidade responsável pelo projeto Apertium, em especial, os pesquisadores e mantenedores Jimmy O'Regan e Francis Tyers.

[1] code.google.com/gsoc

2 TRADUÇÃO AUTOMÁTICA

O objetivo da Tradução Automática é usar computadores para auxiliar todo ou uma parte do processo de tradução entre duas linguagens naturais (normalmente chamadas de língua fonte e língua alvo) (JUR 2000). Sua história vem desde o séc. XVII, quando filósofos como Descartes e Leibniz estabeleceram propostas para códigos que relacionariam palavras entre linguagens. No entanto, a primeira implementação de um sistema de TA ocorreu somente em 1954, com o experimento de Georgetown, onde 49 sentenças em Russo foram traduzidas automaticamente para o Inglês. Desde então, tem se buscado a chamada "tradução automática de alta qualidade de textos irrestritos" (FAHQUT - *fully automated high quality translation of unrestricted text*). Entretanto, vários obstáculos existem para que se alcance esse objetivo, devido às muitas diferenças que ocorrem nos diversos níveis linguísticos das línguas diferentes:

Léxico: Uma palavra em uma língua fonte pode ter como correspondente, na língua alvo, uma Expressão Multipalavra ou pode não ter correspondente. Por exemplo, a palavra *sinaleira* no Português tem como equivalente *traffic lights* no Inglês.

Morfológico: Nesse nível, as línguas podem diferir na quantidade de flexões por palavra e/ou no grau de segmentação dessas flexões. Em Inglês, um adjetivo é invariável (não possui flexões, como por exemplo *beautiful*) enquanto em Português eles podem ter gênero, número e grau (como por exemplo *bonito*, que pode ter flexões como *bonita* e *bonitinho*).

Sintático: A ordem como os sujeitos, verbos e objetos são posicionados dentro das sentenças é uma das diferenças sintáticas mais notáveis. Por exemplo, o Português e o Inglês utilizam a ordem *SVO* (sujeito-verbo-objeto), enquanto o Japonês possui uma ordem *SOV* e o Árabe possui ordem *VSO*. Além disso, existem línguas que não possuem uma ordem fixa, como o Húngaro e o Polonês.

Semântico: As palavras de uma língua fonte podem ter mais de um correspondente legítimo na língua alvo, mas com significados diferentes. Um desses casos é o substantivo *step* no Inglês, que no Português pode significar *passo*, *etapa* ou *degrau*.

Cultural: A sentença resultante de uma tradução pode não parecer "adequada" para a cultura da língua alvo. Os problemas nesse nível normalmente são encontrados em traduções de livros ou poemas, onde os tradutores humanos devem atentar para o objetivo do autor (JUR 2000). Textos mais técnicos geralmente não necessitam desse tratamento.

Neste trabalho o foco será em aprimoramentos no nível morfológico em sistemas de TA. Para tanto, a Seção 2.1 trata da classificação dos sistemas de TA enquanto que a Seção 2.2 mostra algumas técnicas de avaliação das traduções geradas por esses sistemas.

2.1 Classificação dos sistemas de TA

O processo de tradução como um todo pode ser dividido em três etapas:

Análise: O primeiro passo é transformar o texto original em uma representação intermediária.

Transferência: Nessa etapa, o texto analisado é transformado no texto correspondente na língua alvo, mas ainda em sua representação intermediária.

Geração: Finalmente, é gerado o texto traduzido a partir da representação devolvida pela etapa de transferência.

Não existe uma definição precisa sobre a forma de executar essas etapas pois podem haver variações no esforço dedicado a cada uma delas. No entanto, existe uma relação entre esses esforços: se pouco esforço for dedicado à análise e à geração, maior esforço terá que ser dedicado à transferência e vice-versa. Essa relação é clarificada no Triângulo de Vauquois (VAU 68), mostrado na figura 2.1. Uma das classificações possíveis para os tradutores automáticos é relativa à posição desses tradutores dentro do triângulo:

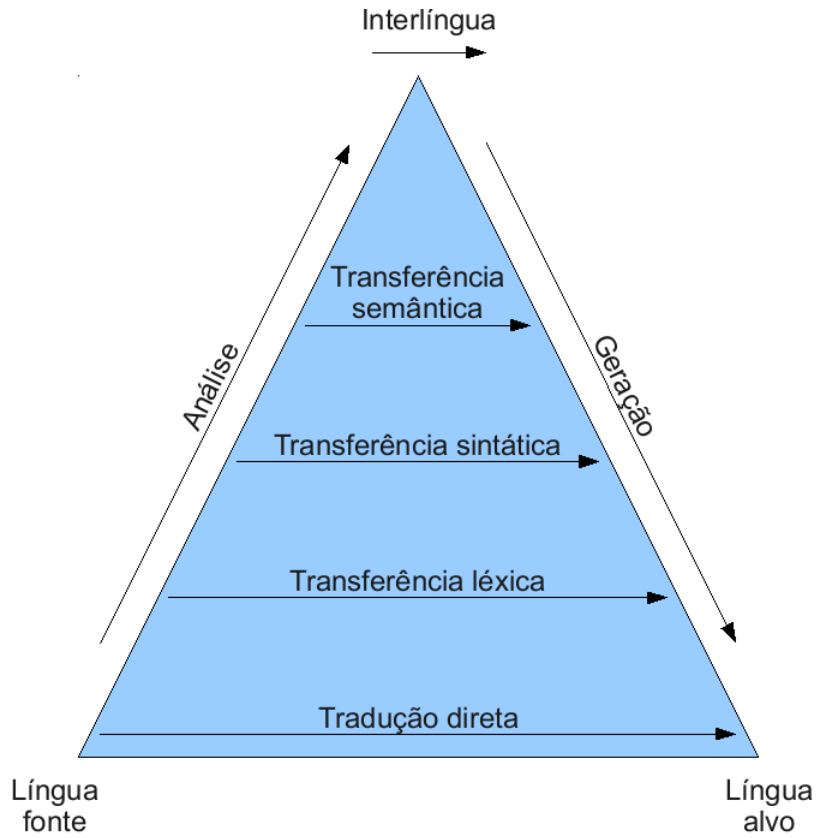

Figura 2.1: Triângulo de Vauquois

TA baseada em interlíngua: Esses modelos utilizam uma linguagem própria de representação de conhecimento (a interlíngua) como estágio de representação intermediário. A ideia da interlíngua é ser independente do par de línguas utilizado na tradução, o que na prática significa retirar a etapa de transferência do processo. A vantagem desses modelos é que o esforço para estender o motor é reduzido pois são necessários apenas O(n) sistemas para traduzir entre cada língua e a interlíngua (os outros modelos são dependentes do par de línguas utilizando, resultando em O(n^2) sistemas). A desvantagem é o fato de interlínguas serem de desenvolvimento extremamente difícil, pois elas devem resolver todas as ambiguidades existentes e ao mesmo tempo representar toda a informação do texto de forma eficiente.

TA baseada em tradução direta: Os modelos de tradução direta fazem o oposto dos modelos de interlíngua: teoricamente, não realizam análise nem geração, preocupando-se somente com a transferência. Essa é normalmente dividida em duas: transferência léxica, onde ocorre a tradução da palavra, e transferência estrutural, onde as palavras traduzidas são eventualmente reorganizadas, com o intuito de colocar o texto na forma correta da língua alvo. Esses modelos têm a vantagem de não estarem sujeitos aos erros de eventuais etiquetadores, *parsers* e analisadores semânticos utilizados em outros modelos.

TA baseada em transferência: Apesar do nome, são um meio termo entre os dois modelos acima. A análise é feita até um nível intermediário, podendo ser morfológico (utilizando dicionários e etiquetadores, por exemplo), sintático (através de *parsers*, ferramentas que detectam as funções sintáticas de cada palavra e/ou grupo de palavras) ou semântico (utilizando representações semânticas dependentes de língua). O texto analisado é então traduzido (na etapa de transferência) para a língua alvo (podendo ser no mesmo nível ou não). Finalmente é feita a geração a partir do texto traduzido, de acordo com o nível em que ele está representado.

As técnicas utilizadas para percorrer o caminho de tradução (as setas dentro do triângulo de Vauquois) também definem uma classificação para os sistemas de TA. Essas técnicas podem ser de dois tipos:

Linguísticas Utilizam recursos estruturados, como dicionários, ontologias e conjuntos de regras, construídos manualmente ou de forma semi-automática por especialistas.

Estatísticas Utilizam recursos não-estruturados, como os corpora paralelos. A ideia é extrair automaticamente desses corpora as informações linguísticas necessárias para o processo de tradução. Para isso, utilizam-se métodos estatísticos e de aprendizagem de máquina.

A combinação dessas duas classificações dá origem a uma série de paradigmas de TA. O Apertium, foco desse trabalho, é um tradutor que utiliza o paradigma *baseado em regras*. Os tradutores que utilizam esse paradigma realizam a análise até o nível morfológico, através de dicionários monolíngues, e utiliza dicionários bilíngues e conjuntos de regras (daí o nome do paradigma) para realizar a transferência. A figura 2.2 contextualiza esse paradigma no Triângulo de Vauquois.

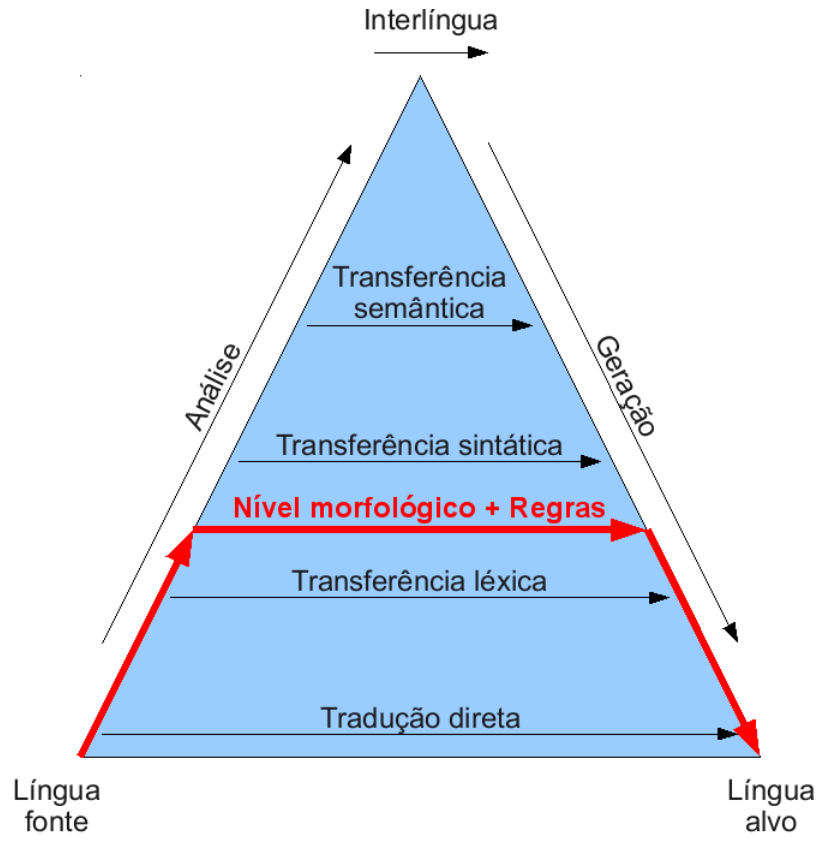

Figura 2.2: Localização do paradigma baseado em regras dentro do Triângulo de Vauquois

2.2 Técnicas de Avaliação

A grande maioria dos textos são complexos o suficiente para gerar mais de uma tradução "adequada", devido à riqueza e variabilidade do léxico, da sintaxe e da semântica de ambas as línguas. Esse fato é visto mesmo quando se faz o processo de forma manual: tradutores humanos diferentes provavelmente proporão traduções diferentes para um mesmo texto. Sendo assim, determinar o que seria uma boa tradução não é uma tarefa trivial.

Esses fatores levaram a diversas pesquisas sobre como desenvolver um bom método de avaliação. Nesta seção serão vistas tanto técnicas manuais quanto técnicas automáticas.

2.2.1 Avaliação manual

Conforme dito anteriormente, mesmo tradutores humanos podem gerar traduções diferentes para um mesmo texto. Assim, é de se esperar que, ao avaliar uma tradução feita por um motor de TA, haja divergências quanto à "qualidade" da tradução. Dessa forma, o que se busca é uma forma sistemática de agrupar várias avaliações diferentes em uma só métrica.

Nesse intuito, os estudos que mais se destacaram são os da *Defense Advanced Research Projects Agency* (DARPA, antiga ARPA), onde se buscou formalizar várias métricas e seus respectivos valores para serem usados por juízes humanos (WHI 94). Este é um programa que começou em 1991 e continua sendo melhorado até hoje. Atualmente, os sistemas utilizados pela DARPA se baseiam nas métricas de *adequação*, que corresponde à quantidade de informação transferida do texto original ao texto traduzido, e de *fluência*,

que diz o quão correto está o texto naquela língua.

As medidas de adequação e de fluidez levam em conta simultaneamente a correspondência *semântica* entre ambos os textos e a correção *sintática* do texto gerado. O exemplo 2.1 mostra três possibilidades de tradução para o inglês de uma mesma frase em português. A primeira tradução, apesar da baixa fluência devida a erros gramaticais, veicula exatamente a mesma informação que a frase original, portanto a adequação é alta. A segunda tradução privilegia uma alta fluência, com uma frase gramaticalmente correta, mas que em detrimento da adequação não tem o mesmo significado que a frase original. Finalmente, a frase em inglês de número três combina alta adequação e alta fluência em uma tradução que pode ser considerada de alta qualidade.

Exemplo 2.1

Frase original: O menino comeu chocolate enquanto redigia o relatório
Tradução 1: While boy chocolate ate, boy report writing was.
Tradução 2: The man took his breakfast while he was reading the newspaper.
Tradução 3: The boy ate chocolate while he was writing the report.

Ainda que se possa obter uma estimativa de qualidade de um sistema de TA através de avaliações manuais, elas são muito lentas e custosas. Com o intuito de promover o desenvolvimento rápido de motores de TA, buscam-se hoje em dia técnicas automáticas de avaliação, mais rápidas e mais baratas.

2.2.2 WER

A idéia geral das técnicas automáticas de avaliação é comparar o resultado com uma tradução de referência, feita por um tradutor humano. Uma delas é a *Word Error Rate* (WER), uma métrica bastante utilizada no reconhecimento de fala. Ela é baseada da distância de Levenshtein (LEV 66), mas aplicada ao nível de palavras ao invés de letras. Sua fórmula é:

$$WER = \frac{S+D+I}{N}$$

onde:

- S é o total de substituições
- D é o total de deleções
- I é o total de inserções
- N é o total de palavras na referência

Devido ao fato do número de edições ser dividido pelo número de palavras na referência, candidatos mais longos podem apresentar taxas de erro maiores que 1. Isso torna a WER uma medida que favorece candidatos mais curtos (ARN 2008). Além disso, as trocas de ordem são bastante penalizadas (como mostra o Exemplo 2.2), o que pode não ser necessariamente um indício de uma má tradução (especialmente em línguas com ordem livre de palavras).

Exemplo 2.2

Candidato: `O homem correu para dentro de sua casa`
Referência: `O homem correu para dentro da casa dele`
Substituições: 3
Deleções: 0
Inserções: 0
WER = 3/8 = 0.375

2.2.3 PER

Uma variação da WER é a *Position-independent Word Error Rate* (PER) (TIL 97). A diferença entre as duas é que a PER não leva em conta a ordem das palavras. A fórmula é a mesma da WER, mas as substituições são contadas somente quando uma palavra é explicitamente trocada por outra, independentemente da ordem, enquanto inserções e deleções só existem se os tamanhos entre o candidato e a referência são diferentes.

No caso do Exemplo 2.2, ao invés de 3 substituições, haveriam apenas 2 pois o substantivo *casa* não foi modificado. Dessa forma o resultado da PER seria 0.25.

Tanto a WER quanto a PER são medidas que tem relação direta com a distância entre os textos. Portanto quanto menor o resultado dessas medidas, melhor é a tradução sendo analisada.

2.2.4 BLEU

A métrica BLEU (*Bilingual Evaluation Understudy*) é baseada na WER e foi desenvolvida pela IBM, tendo atingido uma grande correlação com avaliações humanas (PAP 2001). A BLEU utiliza a idéia de precisão, verificando quantas vezes um n-grama (sequência de *n* palavras) do candidato aparece na tradução de referência. No entanto, somente considerar a precisão pode trazer resultados espúrios, como mostra o Exemplo 2.3. Na prática o que a BLEU usa é a chamada *precisão modificada*, que também considera se os n-gramas aparecem exatamente na mesma quantidade.

Exemplo 2.3

Candidato: `the the the the the the the`
Referência: `The cat is on the mat`
Precisão de unigramas: 7/7
Precisão modificada de unigramas: 2/7

A fórmula genérica para cada candidato é definida da seguinte forma:

$$P_n = \frac{\sum_{n\text{-}grama \in candidato} Count_{clip}(n\text{-}grama)}{\sum_{n\text{-}grama \in candidato} Count(n\text{-}grama)}$$

onde:

- P_n é a precisão modificada para os n-gramas de tamanho n.
- $Count(n\text{-}grama)$ é o total de ocorrências do n-grama no candidato.
- $Count_{clip}(n\text{-}grama)$ é o total de ocorrências do n-grama na referência.

Para o estabelecimento da métrica, a fórmula é calculada para valores de n que vão de 1 até N. Valores altos para P_1 (unigramas) tendem a satisfazer o critério de *adequação*, enquanto que valores altos para n-gramas maiores tendem a satisfazer o critério de *fluência*. Assim, a métrica reúne todos os valores calculados até N usando a média geométrica entre eles. O valor utilizado por Papineni et al. é $N = 4$, devido a esse valor ser o mais próximo de julgamentos feitos por avaliadores humanos (PAP 2001).

Alguns candidatos podem ser pequenos demais e consequentemente exibir altos valores de precisão modificada, apesar de serem considerados traduções inadequadas, como mostra o Exemplo 2.4. Sendo assim, o resultado da média é multiplicado por um fator de brevidade do candidato. Levar em conta os tamanhos de cada sentença (candidata e de referência) separadamente poderia punir demais o valor da métrica, assim, esse fator é calculado utilizando os tamanhos dos textos inteiros.

Exemplo 2.4

Candidato: `of the`
Referência: `It is the guiding principle which guarantees the military forces always being under the command of the Party.`
Precisão modificada de unigramas: 2/2
Precisão modificada de bigramas: 1/1

O valor da penalidade por brevidade é calculado da seguinte forma:

$$BP = \begin{cases} 1 & \text{se } c > r \\ e^{(1-r/c)} & \text{se } c \leq r \end{cases}$$

onde:

- r = número de palavras no texto de referência.
- c = número de palavras no texto candidato.

Finalmente, a métrica BLEU possui a seguinte fórmula:

$$BLEU = BP \cdot \exp(\sum_{n=1}^{N} w_n log P_n)$$

onde:

- BP = fator de brevidade do texto candidato.
- P_n = precisão modificada dos n-gramas.
- N = limite superior dos n-gramas a serem considerados((PAP 2001) usa 4).
- W_n = peso do n-grama ((PAP 2001) usa $1/N$, o que resulta numa distribuição uniforme dos valores).

2.2.5 NIST

A métrica NIST (DOD 2002) é uma variação da BLEU. A diferença básica é que a NIST considera o quão "informativo" um n-grama é: um peso maior é dado se o n-grama aparece com uma frequência menor no texto. A idéia é valorizar mais os n-gramas que supostamente possuam uma maior carga de informação no texto, determinando pesos diferenciados para esses n-gramas. Esses pesos são calculados como determina a fórmula abaixo:

$$Info(w_1 \dots w_n) = log_2 \left(\frac{\text{ocorrências de } w_1 \dots w_{n-1}}{\text{ocorrências de } w_1 \dots w_n} \right)$$

Além disso, a penalidade por brevidade é calculada de forma a minimizar o seu impacto se a variação de palavras entre a referência e o candidato for relativamente pequena. Com isso, a fórmula completa da NIST é:

$$NIST = \sum_{n=1}^{N} \left\{ \frac{\sum_{\text{todas co-ocorrências de } w_1 \dots w_n} Info(w_1 \dots w_n)}{\sum_{\text{todos } w_1 \dots w_n \text{ no candidato}} (1)} \right\} \cdot \exp \left\{ \beta log_2 \left[min \left(\frac{L_{sys}}{\bar{L}_{ref}}, 1 \right) \right] \right\}$$

onde:

- β = fator de brevidade do texto candidato.
- N = 5.
- $\bar{L}_{ref}$ = número médio de palavras na tradução de referência.
- L_{sys} = número de palavras no texto candidato.

Diferente da WER e da PER, as medidas BLEU e NIST são inversamente proporcionais à diferença entre os textos. Ou seja, quanto maiores os resultados dessas medidas, melhor é a tradução.

Tanto a BLEU quanto a NIST são as medidas automáticas mais utilizadas atualmente na avaliação dos motores de TA, tanto por serem rápidas de calcular quanto por terem uma boa correlação com julgamentos humanos. Por conta disso, utilizaremos essas duas métricas nos experimentos desse trabalho.

3 EXPRESSÕES MULTIPALAVRAS

As Expressões Multipalavras (EMPs) podem ser definidas como entidades compostas por duas ou mais palavras cuja composicionalidade léxica, sintática, semântica, pragmática ou estatística é limitada (SAG 2002). Diz-se que uma expressão é composicional quando é possível determinar suas características a partir das características de seus componentes isolados, e EMPs podem ser não-composicionais em um ou mais níveis. Um exemplo é a expressão *colher de pau*, cujo significado seria "colher de madeira" e não "colher feita de pau(s)". Mesmo que essa diferença seja sutil (já que os significados de "pau" e "madeira" se confundem em muitos contextos), ela deve ser considerada pois tratar a expressão como composicional pode gerar erros no seu tratamento. Exemplos mais extremos de não-composicionalidade são as expressões idiomáticas como *chutar o balde*, que significa "ficar com raiva".

Estima-se que, em qualquer língua, o seu total seja comparável ao total de palavras simples (JAC 97) no vocabulário de um falante. Por outro lado, elas possuem características bastante peculiares, tornando o seu tratamento adequado algo não-trivial.

A dificuldade em tratar as EMPs vem do fato de elas não terem um comportamento uniforme: elas podem ser fixas (*ad hoc*), sofrer flexão morfológica (*kicked the bucket*) ou ainda ter flexibilidade sintática (*don't give it up*). Dessa forma, abordagens restritivas para sua detecção e tratamento, onde por exemplo, as EMPs são consideradas como uma palavra só separadas por espaços (abordagem *palavras-com-espaços*), possuem o problema de não abranger casos mais flexíveis, como EMPs que podem aparecer separadamente no texto (caso de *give up*). Por outro lado, abordagens mais flexíveis acabam aceitando casos inexistentes (p.ex. gerando *carneiro expiatório*, quando se desejava apenas *bode expiatório*).

Segundo (SAG 2002), as EMPs podem ser classificadas como:

Expressões fixas São aquelas que não apresentam flexões morfológicas e/ou aquelas que não podem aparecer separadas no texto, ou seja, sua forma é fixa. Exemplos desse tipo incluem expressões como *ad hoc* e *Porto Alegre*. Abordagens do tipo palavras-com-espaços são as mais adequadas nesse caso pois tratá-las de forma composicional implicaria em por exemplo, criar entradas léxicas individuais para *ad* e *hoc*, no caso de *ad hoc*.

Expressões semi-fixas Expressões que admitem eventuais flexões morfológicas. Incluem-se nessa categoria substantivos compostos como *colher de pau* e expressões idiomáticas fixas como *bater as botas*.

Expressões sintaticamente flexíveis São as expressões que permitem variações sintáticas, como aparecerem com suas componentes em ordem variada. As expres-

sões idiomáticas decomponíveis como *tirar o cavalinho da chuva* aparecem nesse grupo. Ela é dita decomponível porque podemos separar seu significado (“desistir da idéia”) nos significados separados de suas componentes *tirar* (“desistir de”) e *o cavalinho da chuva* (“a idéia”). Note que essa análise não pode ser feita para *bater as botas* (“morrer”), por exemplo. No Inglês, as construções verbo-partícula (*give up, track down*) também são exemplos de expressões desse grupo.

Expressões institucionalizadas Apesar de serem completamente composicionais, essas expressões são consideradas EMPs devido a sua alta ocorrência estatística, como a expressão *sal e pimenta*. Ainda que “pimenta e sal” seja uma construção correta e compreensível, não se usa esta expressão nessa forma. As expressões institucionalizadas podem variar morfologica e sintaticamente. Um exemplo são as colocações, ou seja, associações usuais entre duas palavras motivadas pelo seu uso e não por restrições sintáticas ou semânticas.

3.1 EMPs em recursos linguísticos

Uma das formas de adicionar informações relativas às EMPs em um sistema de PLN é através de recursos linguísticos, como dicionários, que contenham também eventuais informações específicas ao sistema.

A construção desses recursos exige a aquisição de EMPs. Ela pode ser feita manualmente ou de forma automática:

Aquisição manual Nesse caso, especialistas em lexicografia são responsáveis pela construção manual de tal recurso. Os recursos gerados de forma manual normalmente não são restritos a reunir apenas EMPs, sendo mais gerais. Exemplos desse tipo de recurso são o WordNet (MIL 90) e o CIDE (PRO 95), respectivamente um tesauro e um dicionário feitos para a língua inglesa. Ainda que ambos possuam entradas multipalavras, seu propósito não é especificamente a catalogação de EMPs. Outros recursos desse tipo são as terminologias e os glossários, ou seja, coleções de termos podendo ser EMPs ou não. Um exemplo desse tipo é o Glossário de Gestão Ambiental (KRI 2007).

Aquisição automática Devido ao alto custo envolvido na construção manual, a automatização desse processo tem sido objeto de pesquisa dentro da área de EMPs. As técnicas mais usadas nesse caso envolvem a extração de informações de frequência de n-gramas (sequências de n palavras) a partir de corpora, a utilização dessas frequências em métodos estatísticos adequados e a filtragem desses n-gramas de acordo com o resultado desses métodos. No entanto as EMPs não possuem um comportamento homogêneo e isso se manifesta durante o processo de aquisição automática. Assim, os trabalhos existentes costumam se focar em tipos específicos de EMPs como nomes compostos (COP 2005) e construções verbo-partícula (BAL 2002).

4 MELHORANDO O TRATAMENTO DE EMPS NO APERTIUM

Neste capítulo, será proposto um aprimoramento para o tratamento das EMPs em um sistema de tradução automática, o Apertium. Na Seção 4.1, a arquitetura do Apertium será descrita mais detalhadamente, mostrando todas as etapas do processo de tradução. O formato dos dicionários utilizados é descrito na Seção 4.2. Nesta mesma seção, também é explicado como é feito o tratamento das EMPs no Apertium e quais as limitações existentes. Finalmente, na Seção 4.3 é discutida uma proposta de solução para algumas dessas limitações: um novo módulo chamado *mweprocessor*. Também são definidos o formato dos dicionários a serem utilizados e o algoritmo implementado pelo módulo.

4.1 Apertium

O Apertium é um sistema de TA baseado em regras, de código aberto, desenvolvido pelo Laboratório de Estudos Linguísticos da Universidade de Alicante (ARM 2007). Foi inicialmente desenvolvido com o intuito de ser utilizado em pares de línguas similares, como o Espanhol e o Catalão. Hoje em dia, no entanto, já são disponibilizados pares mais distantes como o Espanhol-Inglês. Dentro da classificação descrita na Seção 2.1, ele é considerado um meio termo entre um motor de tradução direta e um motor de tradução por transferência.

A implementação do Apertium subdivide as três etapas padrão de um tradutor automático em subetapas. Essas subetapas são implementadas através de módulos independentes, numa arquitetura semelhante a um pipeline, conforme mostrada na Figura 4.1. A vantagem dessa implementação é que ela facilita a inclusão de novos módulos no processo.

As próximas subseções explicam mais detalhadamente a função de cada subetapa, além de incluir exemplos.

4.1.1 Deformatação

O módulo de deformatação é responsável pela retirada de tags de *markup* em textos do tipo HTML ou RTF, por exemplo. Conforme mostrado no Exemplo 4.1, as tags são envolvidas em colchetes, que são ignorados pelos módulos subsequentes do Apertium.

Exemplo 4.1

Entrada: `<h1> a casa </h1>`
Saída: `[<h1>] a casa [</h1>]`

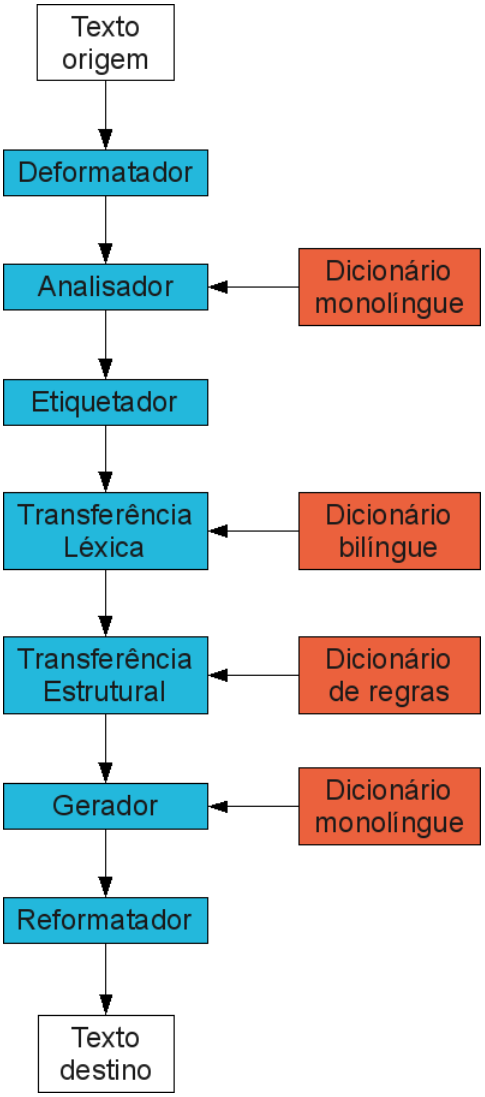

Figura 4.1: Arquitetura simplificada do Apertium

4.1.2 Análise morfológica

Nessa etapa, cada palavra na *forma superficial* é analisada e suas possíveis *formas léxicas* são geradas.

A forma superficial de uma palavra é aquela em que ela aparece no texto enquanto a forma léxica é composta pelo lema e pela classificação morfológica da palavra, como mostra o Exemplo 4.2. A classificação morfológica diz se a palavra é um substantivo, verbo, etc., e também a sua respectiva flexão (gênero, número, tempo verbal, etc.). Para cada forma superficial, é gerado um conjunto dessas formas léxicas, já que pode haver ambiguidades na morfologia da palavra.

Exemplo 4.2 (Forma superficial e formas léxicas da palavra *casas*)

Forma superficial: `casas`
Formas léxicas: `casa<n><f><pl>` *e* `casar<vblex><pri><p2><sg>`

O Apertium utiliza os caracteres ^ e *$* para delimitar o início e o fim de cada palavra. Cada palavra inicia com a sua forma superficial e é seguida pelas formas léxicas, separadas com barras invertidas. Cada forma léxica é determinada de acordo com o dicionário

monolíngue utilizado, gerando um lema e uma sequência de etiquetas (delimitadas pelos caracteres < e >) que mostram a classificação morfológica daquela palavra.

No Exemplo 4.3, a palavra *a* gera três formas léxicas: a preposição *a* (<*pr*>), o artigo definido feminino singular *a* (<*det*> <*def*> <*f*> <*sg*>) e o pronome pessoal oblíquo feminino da terceira pessoa do singular *a* (<*prn*> <*pro*> <*p3*> <*f*> <*sg*>). Já a palavra *casa* gera outras três formas léxicas: o substantivo feminino singular *casa* (<*n*> <*f*> <*sg*>), o verbo *casar* na segunda pessoa do singular do modo imperativo (<*vblex*> <*imp*> <*p2*> <*sg*>) e o verbo *casar* na terceira pessoa do singular do presente do modo indicativo (<*vblex*> <*pri*> <*p3*> <*sg*>).

Exemplo 4.3

Entrada: `[<h1>] a casa [</h1>]`
Saída: `[<h1>] ^a/a<pr>/o<det><def><f><sg>/ o<prn><pro><p3><f><sg>$ ^casa/casa<n><f><sg>/ casar<vblex><imp><p2><sg>/ casar<vblex><pri><p3><sg>$ [</h1>]`

4.1.3 Etiquetagem

Conforme dito anteriormente, devido às ambiguidades morfológicas, muitas vezes uma forma superficial gera mais de uma forma léxica. A etiquetagem é responsável pela escolha da forma léxica mais adequada, como mostra o Exemplo 4.4.

Exemplo 4.4

Entrada: `[<h1>] ^a/a<pr>/o<det><def><f><sg>/ o<prn><pro><p3><f><sg>$ ^casa/casa<n><f><sg>/ casar<vblex><imp><p2><sg>/ casar<vblex><pri><p3><sg>$ [</h1>]`
Saída: `[<h1>] ^o<det><def><f><sg>$ ^casa<n><f><sg>$ [</h1>]`

Ao contrário dos outros módulos do Apertium, o módulo responsável pela etiquetagem utiliza dados estatísticos, mais especificamente, modelos de Markov (CUT 92) criados a partir de corpora.

4.1.4 Transferência léxica

A transferência léxica é responsável pela tradução em si: cada palavra na língua fonte é traduzida para a língua alvo de acordo com a sua forma léxica, utilizando para isso um dicionário bilíngue. O Exemplo 4.5 considera uma tradução do Português para o Espanhol.

A tradução é feita buscando o lema no dicionário bilíngue e substituindo-o pelo lema correspondente na língua alvo. As etiquetas são mantidas as mesmas da forma lexical na língua fonte, a menos que a entrada correspondente ao lema no dicionário indique explicitamente uma substituição de etiquetas (por exemplo, trocar o gênero da palavra).

Exemplo 4.5

Entrada: `[<h1>] ^o<det><def><f><sg>$ ^casa<n><f><sg>$ [</h1>]`
Saída: `[<h1>] ^el<det><def><f><sg>$ ^casa<n><f><sg>$ [</h1>]`

4.1.5 Transferência estrutural

Apenas traduzir as palavras em si pode não ser suficiente para a correta tradução do texto como um todo, devido às idiossincrasias estruturais de cada língua. Deve-se levar em conta também a estrutura sintática das frases na língua alvo. A transferência estrutural é responsável por essa transformação, utilizando para isso um dicionário de regras.

No caso do Exemplo 4.5 o módulo verifica a concordância do artigo com o substantivo. Como essa concordância já está correta, nenhuma modificação é feita.

4.1.6 Geração morfológica

Após a transferência, o texto obtido está na língua alvo, mas ainda possui palavras nas suas respectivas formas léxicas. O módulo de geração é responsável pela transformação das formas léxicas nas respectivas formas superficiais, utilizando também um dicionário monolíngue, dessa vez da língua alvo. Este processo pode ser visto no Exemplo 4.6.

Exemplo 4.6

Entrada: `[<h1>] ^el<det><def><f><sg>$ ^casa<n><f><sg>$ [</h1>]`
Saída: `[<h1>] la casa [</h1>]`

4.1.7 Reformatação

A última etapa recoloca as tags de *markup* retiradas pela primeira etapa, como mostrado no Exemplo 4.7. Esse processo é necessário para que o texto resultante possua o mesmo formato (HTML, por exemplo) do texto original.

Exemplo 4.7

Entrada: `[<h1>] la casa [</h1>]`
Saída: `<h1> la casa </h1>`

4.2 Dicionários

Além do motor em si, cada par de línguas do Apertium é composto por quatro dicionários, isto é, cinco arquivos em formato XML que reúnem as informações linguísticas necessárias daquele par:

Dicionários monolíngues São dois, um para cada língua do par. Utilizados nas etapas de análise e de geração, possuem as informações morfológicas das palavras. Esses dicionários possuem duas seções, uma contendo os lemas das palavras e outra contendo paradigmas de flexão. Dessa forma, ao invés de serem listadas exaustivamente todas as flexões de uma palavra, o dicionário contém somente uma entrada, correspondente ao lema dessa palavra e o seu respectivo paradigma de flexão. A idéia dos paradigmas é dizer, por exemplo, que *casa* flexiona da mesma forma que *abadia* e que *falar* flexiona da mesma forma que *amar*.

Um trecho de cada seção é mostrado na Figura 4.2. A entrada (<*e*>) correspondente à palavra "casa" diz que seu lema é *casa* e que sua flexão segue o paradigma *abadia__n*. Esse paradigma por sua vez diz que se a palavra terminar em "s" (<*l*>*s*</*l*>) ela é um substantivo (<*s n="n"*/>) feminino (<*s n="f"*>) plural (<*s*

n="pl">). O mesmo paradigma diz que se a palavra for igual ao lema (<*l/*>) ela é um substantivo feminino singular (<*s n="s"*>). A ordem em que as informações morfológicas aparecem no paradigma corresponde à ordem em que as etiquetas serão colocadas na forma léxica da palavra.

```
...
<pardef n="abadia__n">
<e>
  <p>
    <l>s</l>
    <r><s n="n"/><s n="f"/><s n="pl"/></r>
  </p>
</e>
<e>
  <p>
    <l/>
    <r><s n="n"/><s n="f"/><s n="sg"/></r>
  </p>
</e>
</pardef>
...
<e lm="casa">
  <i>casa</i>
  <par n="abadia__n"/>
</e>
...
```

Figura 4.2: Trecho extraído do dicionário monolíngue de Português

Dentro do Apertium esses dicionários não são utilizados diretamente. Eles são compilados na forma de transdutores de estados finitos (ROJ 2005). A idéia da compilação prévia é melhorar o desempenho do acesso ao dicionário.

Dicionário bilíngue É utilizado na etapa de transferência léxica. Cada entrada desse dicionário possui o lema na língua fonte e seu correspondente na língua alvo, conforme mostra a Figura 4.3. Esse dicionário também é compilado previamente na forma de transdutores, por questões de desempenho.

Dicionário de regras Usados na transferência estrutural, existe um para cada direção que se deseja traduzir (por exemplo, existe um para tradução do Português para o Espanhol e outro para tradução do Espanhol para o Português). Possuem um conjunto de regras que definem eventuais modificações nas estruturas de cada sentença, como por exemplo a ordem das palavras. Essas regras podem ser definidas por um conjunto bastante rico de instruções, incluindo expressões regulares e instruções de controle de fluxo (*"ifs"*). Um exemplo de regra é mostrado na Figura 4.4. Essa regra específica diz como um artigo (<*pattern-item n="det"/*>) deve se comportar no momento da transferência. Para verificar e eventualmente adaptar a concordância é chamada uma macro (<*call-macro n="f_concord1"*>). Essas macros são descritas dentro desses mesmos dicionários e são criadas manualmente por especialistas.

```
...
<e>
  <p>
    <l>hablar<s n="vblex"/></l>
    <r>falar<s n="vblex"/></r>
  </p>
</e>
<e>
  <i>aumentar<s n="vblex"/></i>
</e>
<e>
  <p>
    <l>posición<s n="n"/></l>
    <r>posição<s n="n"/></r>
  </p>
</e>
...
```

Figura 4.3: Trecho extraído do dicionário bilíngue Português-Espanhol

```
...
<rule> <!- REGLA 1:  DETERMINANT ->
  <pattern>
    <pattern-item n="det"/>
  </pattern>
  <action>
    <call-macro n="f_concord1">
    </call-macro>
  </action>
</rule>
...
```

Figura 4.4: Trecho (simplificado) do dicionário de regras Português-Espanhol

4.2.1 Tratamento de EMPs no Apertium

O Apertium é capaz de processar somente alguns tipos de EMPs:

Expressões fixas: nesse caso, é utilizada a abordagem palavras-com-espaços, o que significa que existe uma entrada específica nos dicionários.

Expressões idiomáticas não-decomponíveis: a mesma abordagem é utilizada, mas um módulo adicional é responsável por assinalar qual dos componentes da expressão flexiona.

Não há tratamento específico para as expressões flexíveis. Já no caso de nomes compostos, os que possuem flexão em apenas uma das componentes (tais como *colher de pau*) são processados da mesma forma que as expressões idiomáticas não-decomponíveis. No entanto, quando mais de uma componente pode apresentar flexão, o seu tratamento torna-se custoso. O problema é o formato dos dicionários monolíngues do Apertium: os paradigmas foram desenvolvidos para comportarem a flexão de apenas uma palavra. Portanto,

não é possível para uma EMP indicar a flexão para mais de uma de suas componentes. Para alguns casos, existem paliativos: no caso do Espanhol, há uma entrada para cada flexão. A figura 4.5 mostra um exemplo, para o caso da EMP *dirección general*, onde há uma entrada para o singular e outra para o plural (*direcciones generales*)

```
<e lm="dirección general" a="mginesti">
  <p>
    <l>dirección<b/>general</l>
    <r>dirección<b/>general<s n="n"/><s n="f"/><s
n="sg"/></r>
  </p>
</e>
<e lm="dirección general" a="mginesti">
  <p>
    <l>direcciones<b/>generales</l>
    <r>dirección<b/>general<s n="n"/><s n="f"/><s
n="pl"/></r>
  </p>
</e>
```

Figura 4.5: Trecho de uma EMP no dicionário monolíngue de Espanhol

Entretanto, esse tipo de tratamento é deselegante e mais propenso a erros (lembrando que os dicionários em sua maioria são construídos manualmente por especialistas). Além disso, esse tipo de abordagem somente é possível quando a EMP possui poucas flexões. Em línguas onde há flexões de caso, como o Polonês, o total de flexões para uma dada palavra é muito grande, tornando essa abordagem totalmente impraticável. Um exemplo é o nome composto *Wielki Tydzień* ("Semana Santa" em Polonês), que possui além das duas flexões de número (singular e plural), 7 flexões de caso, gerando 14 flexões no total.

4.3 O módulo mweprocessor

Como os dicionários monolíngues do Apertium são inadequados para o tratamento das expressões onde há mais de uma componente que possua flexões, um novo formato de dicionário é necessário.

No entanto, alterar os dicionários existentes para que sigam esse novo formato seria dispendioso demais, além de serem necessárias alterações no módulo de análise e de geração morfológica. Dessa forma, os dicionários existentes foram mantidos e, paralelamente, foram criados novos dicionários (um para cada língua), específicos para as EMPs, seguindo esse novo formato. O processamento desses novos dicionários é feito por um novo módulo, chamado *mweprocessor*.

Esse módulo se insere no esquema de funcionamento do Apertium em duas etapas do pipeline, a primeira entre o analisador e o etiquetador e a segunda antes do gerador, conforme a Figura 4.6. Na primeira etapa, o módulo é utilizado no modo de *análise das EMPs*, onde a ideia é que uma sequência de palavras do texto seja analisada e comparada com as entradas do dicionário. Caso essa sequência esteja presente no dicionário, a mesma é considerado uma EMP e é transformada em uma única entidade léxica. Na segunda etapa, onde o módulo é utilizado no modo de *geração das componentes*, cada palavra do texto de entrada é procurada no dicionário. Caso esteja presente (caracteri-

zando uma EMP), ela é “quebrada”, imprimindo suas componentes separadamente no texto de saída.

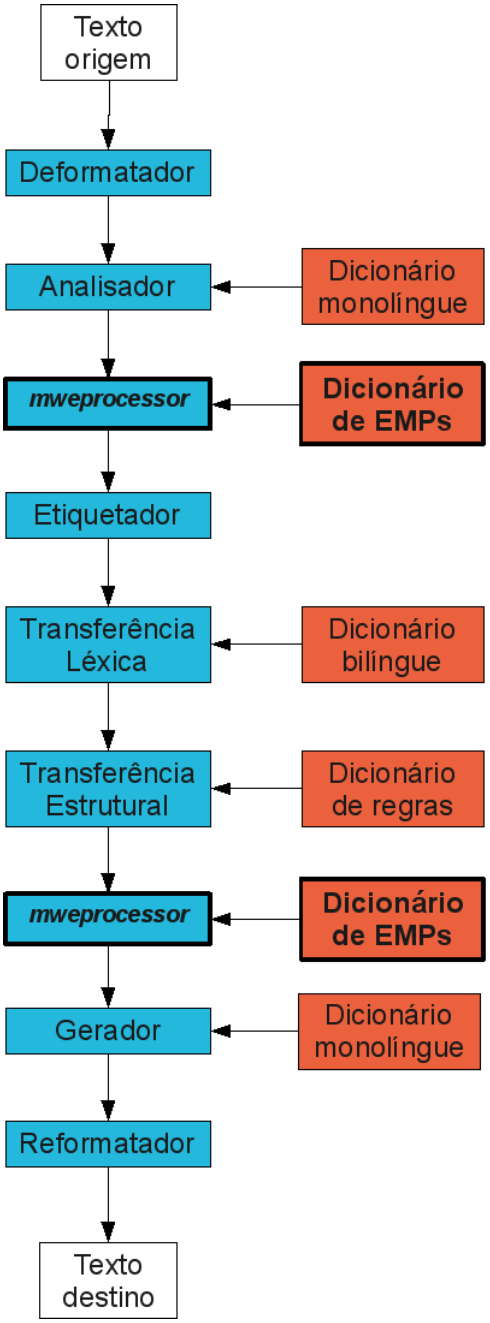

Figura 4.6: Nova arquitetura proposta do Apertium

4.3.1 Dicionário de multipalavras

Assim, como os outros dicionários utilizados no Apertium, os dicionários de EMPs serão descritos utilizando a linguagem XML. Esses dicionários contém uma lista de paradigmas e uma lista de EMPs. Cada EMP é relacionada a um paradigma, que diz como aquela EMP deve flexionar, de acordo com as classificações morfológicas de suas componentes. Um trecho de um dicionário, contendo a descrição de um paradigma e uma EMP que utiliza esse paradigma descrito, é mostrado na Figura 4.7.

```
<mwedictionary>
<pardefs>
...
  <pardef n="dg_par">
    <tagset input1="n.f.s" input2="adj.f.s"
output="n.f.s"/>
    <tagset input1="n.f.pl" input2="adj.f.pl"
output="n.f.pl"/>
    <tagset input1="n.f.s" input2="adj.mf.s"
output="n.f.s"/>
    <tagset input1="n.f.pl" input2="adj.mf.pl"
output="n.f.pl"/>
  </pardef>
...
</pardefs>
<mwes>
...
  <mwe n="direccion general">
    <lemmas lm1="direccion" lm2="general"/>
    <par n="dg_par"/>
  </mwe>
...
</mwes>
</mwedictionary>
```

Figura 4.7: Trecho de um possível dicionário de EMPs em Espanhol

4.3.2 Modo de análise das EMPs

O modo de análise das EMPs é responsável pela detecção das EMPs, para que possa ser feito o seu tratamento correto pelos módulos subsequentes. Para isso, utiliza-se o dicionário de EMPs da língua fonte. Nesse modo, o texto de entrada é processado seguindo o algoritmo descrito abaixo:

1) As primeiras *n* palavras do texto são lidas, formando um n-grama.

2) Os lemas e etiquetas das formas léxicas de cada componente do n-grama são combinados entre si, gerando candidatos a EMP.

3) Cada candidato é procurado no dicionário. Para que ele seja considerado uma EMP, deve haver uma entrada contendo o lema e as etiquetas iguais às do candidato.

- **3a)** Se o candidato existir no dicionário, ele é considerado uma EMP, sendo impresso no texto de saída com suas respectivas forma superficial e formas léxicas.

4) Verifica-se se algum candidato foi considerado EMP durante a etapa 3:

- **4a)** Caso positivo, o processo se repete com as *n* palavras seguintes à EMP.
- **4b)** Caso negativo, a primeira palavra é impressa no texto de saída sem modificações, e o processo se repete com as próximas *n* palavras.

Esse processo é repetido até que se encontre o fim do texto. Neste trabalho *n* está restrito a 2, o que significa que o *mweprocessor* pode somente capturar EMPs formadas por 2 componentes. No entanto, o algoritmo pode ser estendido para capturar EMPs maiores. O Exemplo 4.8 mostra uma possível execução do algoritmo implementado, considerando *n*=2 e a existência da EMP *foreign policy* no dicionário.

Exemplo 4.8 (Possível execução do modo de análise das EMPs)

Texto de entrada: `^its/its<det><pos><sp>$ ^foreign/foreign<adj>$ ^politics/policy<n><sg>/politics<n><pl>$`

Etapa 1: `^its/its<det><pos><sp>$` *e* `^foreign/foreign<adj>$` *são lidas.*

Etapa 2: *Geração dos candidatos. Como cada palavra tem somente uma forma léxica, somente um candidato é gerado. Esse candidato é formado pelas formas léxicas* `its<det><pos><sp>` *e* `foreign<adj>`.

Etapa 3: *O candidato é procurado no dicionário, mas não é encontrado ("its foreign" não é uma EMP).*

Etapa 4: *Como não foi impresso nenhum candidato, passa-se para a Etapa 4b.*

Etapa 4b: *Imprime-se a primeira palavra (*`^its/its<det><pos><sp>$`*) e o processo recomeça na Etapa 1 com as próximas 2 palavras.*

Etapa 1: `^foreign/foreign<adj>$` *e* `^politics/politics<n><sg>/policy<n><pl>$` *são lidas.*

Etapa 2: *Geração dos candidatos. Novamente, somente um é gerado, contendo as formas léxicas* `foreign/foreign<adj>` *e* `policy<n><sg>`.

Etapa 3: *O candidato é procurado no dicionário e é encontrado.*

Etapa 3a: *O candidato encontrado é considerado uma EMP e é impresso na saída, com sua forma superficial e sua forma léxica (*`^foreign policy/foreign policy<n><sg>$`.

Etapa 4: *Como houve a impressão de pelo menos um candidato, passa-se para a Etapa 4a.*

Etapa 4a: *Repete-se o processo com as 2 palavras seguintes à EMP impressa. No entanto, como é encontrado o fim do texto, o processo termina.*

Texto de saída: `^its/its<det><pos><sp>$ ^foreign politics/foreign policy<n><sg>$`

4.3.3 Modo de geração das componentes

As informações correspondentes às EMPs não estão presentes nos dicionários monolíngues. Esses dicionários possuem apenas informações relativas às *componentes* delas. Por causa disso, um texto contendo as EMPs em suas formas léxicas não pode ser processado corretamente pelo módulo de geração. Dessa forma, antes da geração, o *mweprocessor* é novamente utilizado, mas no modo de *geração das componentes*. Esse modo faz o caminho inverso do modo de análise das EMPs, fazendo a sua fragmentação e gerando suas respectivas componentes. Para isso, ele utiliza o dicionário de EMPs da língua alvo.

O algoritmo para esse modo é descrito abaixo:

1) A primeira palavra do texto é lida.

2) Essa palavra é procurada no dicionário.

2a) Se não for encontrada, ela é impressa no texto de saída de forma inalterada.

2b) Caso contrário, ela é considerada uma EMP e as formas léxicas das suas componentes são impressas na saída.

3) Repete-se o processo com a próxima palavra.

O processo se repete até o fim do texto. O Exemplo 4.9 mostra uma possível execução, que considera a existência da EMP *política exterior* no dicionário.

Exemplo 4.9 (Possível execução do modo de geração das componentes)
Texto de entrada: `^suyo<prn><tn><pos><f><sg>$ ^política exterior <n><f><sg>$`
Etapa 1: `^suyo<prn><tn><pos><f><sg>$` *é lida.*
Etapa 2: *A palavra é procurada no dicionário, mas não é encontrada ("suyo" não é uma EMP).*
Etapa 2a: *Imprime-se a palavra sem modificações.*
Etapa 3: *Repete-se o processo com a próxima palavra.*
Etapa 1: `^política exterior<n><f><sg>$` *é lida.*
Etapa 2: *A palavra é procurada e dessa vez ela é encontrada.*
Etapa 2b: *Sendo considerada uma EMP, ela é fragmentada em suas componentes (*`política<n><f><sg>` *e* `exterior<adj><mf><sg>`*), que são impressas no texto de saída.*
Etapa 3: *Repete-se o processo com a próxima palavra. No entanto, como se chegou a fim do texto, o algoritmo termina.*
Texto de saída: `^suyo<prn><tn><pos><f><sg>$ ^política<n><f><sg>$ ^exterior<adj><mf><sg>$`

5 AVALIAÇÃO

Para avaliar a diferença causada pelo novo módulo nas traduções, uma série de experimentos foram realizados e avaliados de acordo com as métricas descritas na Seção 2.2. A Seção 5.1 descreve os dados utilizados na avaliação, enquanto que a Seção 5.2 explica mais detalhadamente a metodologia utilizada nos experimentos. Os resultados são mostrados na Seção 5.3, juntamente com uma avaliação qualitativa.

5.1 Dados

Para realizar os experimentos, foi utilizado o corpus *News Commentary*[1], disponibilizado durante o terceiro *ACL Workshop On Statistical Machine Translation (WMT2008)*. Fizeram parte desse evento a submissão de propostas para várias tarefas relacionadas à TA, incluindo tanto a melhoria de sistemas quanto a definição de novas métricas automáticas de avaliação.

O News Commentary (que será referido a partir daqui como *NC*) foi um dos corpora utilizados para avaliação e julgamento das soluções propostas durante o *WMT2008*. Ele é composto por mais de 60.000 frases retiradas de notícias de jornais, originalmente em Inglês. Seu formato é de texto puro, contendo uma frase por linha. O corpus possui versões traduzidas manualmente em várias línguas, incluindo Português, Espanhol e Húngaro. Como a avaliação automática necessita de traduções manuais para serem usadas como referência, esse corpus é adequado para a realização dos experimentos.

Para realizá-los, foi escolhido o par de línguas Inglês-Espanhol, por ser um dos pares que contém mais casos das EMPs abordadas e também para facilitar uma avaliação qualitativa (o par Inglês-Português não é suportado pelo Apertium). Sendo assim, as versões do NC utilizadas foram as dessas línguas.

A correspondência entre as frases nas múltiplas versões do NC é feita pela numeração das linhas. Ou seja, a linha 10 da versão em Espanhol contém uma frase que é a tradução da frase contida na linha 10 da versão em Inglês. No entanto, o corpus possui ruído: algumas frases não possuem correspondentes (no lugar delas haviam linhas em branco), outras frases são formadas por palavras sem sentido (como por exemplo, *Ìîñêâó*). Sendo assim, foi feita um pré-processamento do corpus, retirando as frases que não haviam correspondentes e as continham ruído. Essa limpeza foi feita de forma paralela, cada linha excluída no corpus em Inglês era excluída no corpus em Espanhol (e vice-versa) para que a correspondência entre as frases fosse mantida.

A Tabela 5.1 mostra alguns dados do corpus, antes e após o pré-processamento. Já as Figuras 5.1 e 5.2 exibem mais detalhadamente a frequência das palavras nos corpora (após

[1] Disponível em http://www.statmt.org/wmt08/shared-task.html

o pré-processamento) em Inglês (referenciado como $NC_{Inglês}$) e Espanhol (referenciado como $NC_{Espanhol}$), respectivamente.

	Corpora originais		Corpora pré-processados	
	$NC_{Inglês}$	$NC_{Espanhol}$	$NC_{Inglês}$	$NC_{Espanhol}$
Total de frases	64308	64308	63725	63725
Total de palavras	1357811	1588494	1340934	1572112
Total de palavras diferentes	47289	58042	41565	52819

Tabela 5.1: Dados do corpus News Commentary

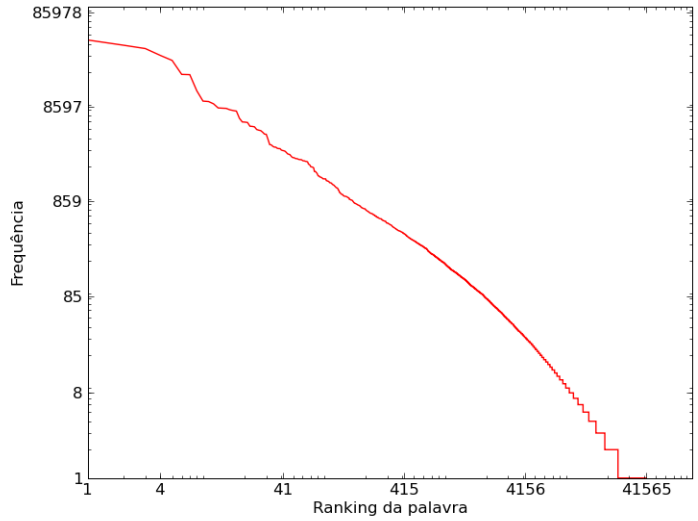

Figura 5.1: Frequência das palavras no *NC* em Inglês (escala logarítmica)

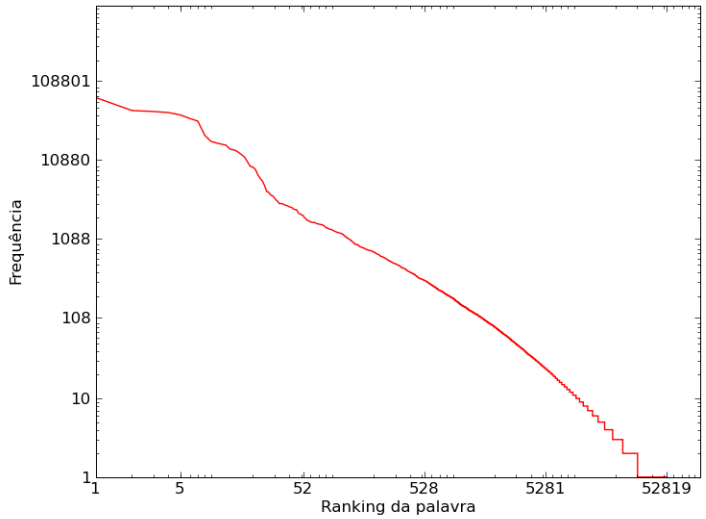

Figura 5.2: Frequência das palavras no *NC* em Espanhol (escala logarítmica)

5.2 Experimentos

O primeiro experimento foi traduzir o corpus em Inglês para o Espanhol utilizando o Apertium na sua arquitetura original, sem o módulo *mweprocessor*. Utilizando o corpus em Espanhol como tradução de referência, foi possível obter os resultados das medidas BLEU e NIST.

Para o segundo experimento, inicialmente foi extraída, de forma manual, uma lista de oito EMPs do corpus em Inglês, com sua respectiva tradução em Espanhol. A extração foi feita por um dos especialistas responsável pelo Apertium. Ainda que pequena, essas amostras extraídas exemplificam adequadamente os casos a serem tratados pelo novo módulo.

A lista de EMPs extraídas, junto com suas frequências, é mostrada na Tabela 5.2 e na Figura 5.3. A Tabela 5.2 também mostra as traduções corretas (segundo o especialista responsável pela extração) dessas EMPs. Com essa lista, foram criados os dois dicionários de EMPs, um para o Inglês (exemplificado na Figura 5.4) e outro para o Espanhol (exemplificado na Figura 5.5), a serem utilizados pelo módulo *mweprocessor*. Além disso, entradas correspondentes a essas EMPs foram adicionadas ao dicionário bilíngue Inglês-Espanhol (um exemplo de entrada adicionada é mostrado na Figura 5.6). Com os dicionários prontos, foi possível então traduzir novamente o corpus, dessa vez usando o Apertium incluindo o *mweprocessor*, e obter as medidas BLEU e NIST para a nova tradução gerada.

EMP em Inglês	Tradução em Espanhol	Frequência da EMP no *NC*
banking crisis	**crisis bancaria**	6
chronic disease	**enfermedad crónica**	4
communist country	**país comunista**	6
ex-communist country	**país ex comunista**	6
foreign policy	**política exterior**	262
political power	**fuerza política**	32
real achievement	**logro real**	4
US citizen	**ciudadano estadounidense**	8

Tabela 5.2: Lista das EMPs extraídas do News Commentary

Com o objetivo de fazer uma análise qualitativa e também de verificar mais precisamente a melhoria na tradução, foi feito um terceiro experimento. Os dois corpora traduzidos foram comparados e as frases onde houve modificações na tradução para o Espanhol foram extraídas. Com isso gerou-se dois fragmentos, um contendo as frases traduzidas pelo Apertium original e outro contendo as mesmas frases, mas traduzidas pelo Apertium modificado (contendo o *mweprocessor*). Esses corpora serão referidos como NC_{frag}.

Para que fosse possível calcular as medidas BLEU e NIST para o NC_{frag}, as frases correspondentes foram extraídas também do corpus original (em Inglês) e do corpus de referência (em Espanhol). Dessa forma, foram criados quatro fragmentos no total, contendo 273 frases cada um. O fato do total de frases modificadas (273) ser menor que o total de ocorrências das EMPs consideradas (328) é porque algumas das EMPs foram corretamente traduzidas pelo Apertium original em algumas das frases, fazendo com que não houvesse diferença em relação à tradução gerada pelo Apertium modificado.

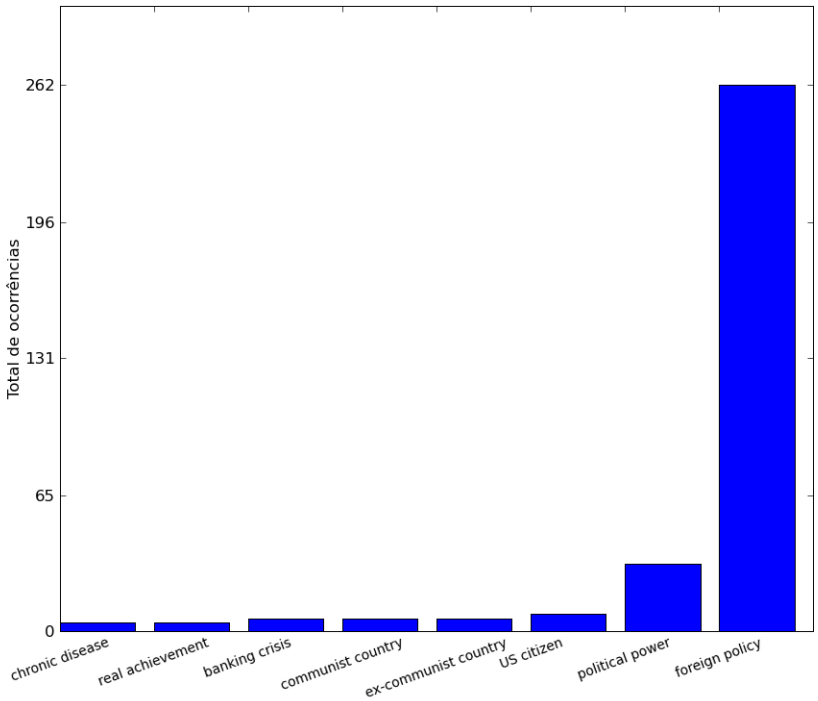

Figura 5.3: Frequência das EMPs no NC

```
...
<pardefs>
...
  <pardef n="real_achievement_par">
    <tagset input1="adj" input2="n.sg" output="n.sg"/>
    <tagset input1="adj" input2="n.pl" output="n.pl"/>
  </pardef>
...
</pardefs>
<mwes>
...
  <mwe n="foreign policy">
    <lemmas lm1="foreign" lm2="policy"/>
    <par n="real_achievement_par"/>
  </mwe>
...
</mwes>
```

Figura 5.4: Trecho extraído do dicionário de EMPs do Inglês

5.3 Resultados

Com o objetivo de comparar os resultados obtidos nos experimentos e verificar se são estatisticamente significantes, foi utilizada a técnica de *bootstrapping* (ZHA 2004). Essa técnica consiste em utilizar o conjunto de teste original para estimar o intervalo de confiança das medidas. Para isso, são extraídas N amostras (com reposição) do conjunto de

```
...
<pardefs>
...
  <pardef n="politica_exterior_par">
    <tagset input1="n.f.sg" input2="adj.mf.sg"
output="n.f.sg"/>
    <tagset input1="n.f.pl" input2="adj.mf.pl"
output="n.f.pl"/>
  </pardef>
...
</pardefs>
<mwes>
...
  <mwe n="politica exterior">
    <lemmas lm1="politica" lm2="exterior"/>
    <par n="politica_exterior_par"/>
  </mwe>
...
</mwes>
```

Figura 5.5: Trecho extraído do dicionário de EMPs do Espanhol

```
...
<e>
  <p>
    <l>foreign<b/>policy<s n="n"/></l>
    <r>política<b/>exterior<s n="n"/><s n="f"/></r>
  </p>
</e>
....
```

Figura 5.6: Exemplo de EMP adicionada ao dicionário bilíngue Inglês-Espanhol

teste e as medidas são calculadas para esses conjuntos. Essas medidas geram valores que seguem uma distribuição normal. Os valores dessas medidas são então ordenados numa lista e, de acordo com o nível de significância[2] desejado, descartam-se dessa lista os valores mais extremos. O intervalo de confiança é então $[lista_{min}, lista_{max}]$, onde $lista_{min}$ e $lista_{max}$ são respectivamente o mínimo e o máximo resultante. Por exemplo, se N=40 (onde N é o total de amostras) e desejamos 95% de nível de significância, o intervalo de confiança estimado é $[3^{a}medida, 38^{a}medida]$, pois para obter os 95% centrais da distribuição, descartamos as duas primeiras e as duas últimas medidas.

Para que seja feita a comparação estatística entre duas medidas, os seus intervalos de confiança são subtraídos um do outro. Se o intervalo resultante não contiver o valor zero, então pode-se dizer que, de acordo com a técnica de *bootstrapping*, eles são estatísticamente diferentes.

A Tabela 5.3 mostra os resultados das medidas BLEU e NIST das duas traduções, tanto para o NC quanto para o NC_{frag}, bem como o intervalo de confiança (IC) resultante

[2] o valor do nível de significância depende do experimento e da confiança desejada pelo autor do experimento

do *bootstrapping*. O total de amostras utilizado no *bootstrapping* foi de 2000 e o nível de significância considerado foi de 95% (valor usado por Zhang et al. (ZHA 2004)).

	NC		NC_{frag}	
	BLEU	NIST	BLEU	NIST
Apertium	0.1995	6.9873	0.1966	5.7519
Apertium + *mweprocessor*	0.1997	6.9909	0.2350	6.1366
IC estimado da diferença	[0,0.007]	[0,0.049]	[0.039,0.046]	[0.400,0.527]

Tabela 5.3: Resultados dos experimentos

EMP	Trecho original	Trecho traduzido pelo Apertium	Trecho traduzido pelo Apertium + *mweprocessor*
banking crisis	My moderate optimism comes from my belief that the US will avoid both a banking crisis and a balance of payments crisis.	Mi moderar el optimismo viene de mi creencia que los EE.UU. evitarán ambos una crisis de banca y un equilibrio de crisis de pagos.	Mi moderar el optimismo viene de mi creencia que los EE.UU. evitarán ambos una crisis bancaria y un equilibrio de crisis de pagos.
ex-communist country	Germany primarily reacted to low-wage competition from ex-communist countries.	Alemania principalmente reaccionada a abajo-competición de sueldo de ex-países comunistas.	Alemania principalmente reaccionada a abajo-competición de sueldo de países ex comunistas.
foreign policy	They will want to protect all social spending, regardless of the consequences for foreign policy.	Querrán proteger todo social gastando, a toda costa de las consecuencias para política extranjera.	Querrán proteger todo social gastando, a toda costa de las consecuencias para política exterior.
real achievement	But "managed democracy"now threatens to unravel all three of his real achievements.	Pero "democracia dirigida"ahora threatens a unravel todo tres de sus consecuciones reales.	Pero "democracia dirigida"ahora threatens a unravel todo tres de sus logros reales.
US citizen	With the exception of 12 US citizens, all were black civilians, while more than 4,000 Kenyans and Tanzanians were injured.	Con la excepción de 12 ciudadanos de EE.UU., todo era civiles negros, mientras más que 4,000 kenianos y Tanzanians fue herido.	Con la excepción de 12 ciudadanos estadounidenses, todo era civiles negros, mientras más que 4,000 kenianos y Tanzanians fue herido.

Tabela 5.4: Exemplos de tradução das EMPs

No corpus *NC* não foi notada diferença estatística significante tanto nos valores da BLEU quanto da NIST. Isso pode ser explicado pelo tamanho pequeno da lista de EMPs utilizada quando comparada com o tamanho do corpus, que faz com que haja uma grande quantidade de frases em que a tradução não é alterada. Já no fragmento NC_{frag} foi observada diferença estatística significante (para mais) nos dois valores, o que mostra que, em frases onde há a ocorrência desse tipo de EMPs, o módulo conseguiu melhorar significativamente a tradução. Esse resultado sugere que, fazendo novos testes com o corpus *NC* utilizando uma lista maior de EMPs, pode-se chegar a uma diferença estatística significante para os valores de BLEU e NIST.

Considerando a melhora obtida no NC_{frag}, foi também feita uma análise qualitativa das frases traduzidas, para que se pudesse averiguar melhor porque essa melhora ocorreu. A tabela 5.4 mostra alguns exemplos de trechos onde houve ocorrência das EMPs do dicionário utilizado. Nesses exemplos, pode-se perceber que a melhora ocorreu basicamente no nível léxico, já que as EMPs passaram a ser corretamente detectadas, o que não acontecia no Apertium em sua arquitetura original. Sendo assim, espera-se que ainda melhores resultados possam ser obtidos se os dicionários de EMPs forem ampliados.

6 CONCLUSÃO E TRABALHOS FUTUROS

As EMPs são um fenômeno linguístico bastante frequente em qualquer língua. Sistemas de TA devem levar em conta o tratamento adequado dessas expressões para que possam gerar traduções de qualidade. Nesse trabalho foi apresentada uma proposta de aprimoramento para a arquitetura do sistema de TA Apertium, que consiste na adição de um módulo, o *mweprocessor*, e um novo formato de dicionário, destinado às EMPs. O novo módulo, juntamente com os novos dicionários, tem o objetivo de tratar EMPs que apresentem flexão em mais de uma de suas componentes, algo que o Apertium não consegue tratar de forma eficiente. O trabalho foi feito em conjunto com a comunidade do projeto Apertium, em especial os mantenedores e pesquisadores Jimmy O'Regan e Francis Tyers, que auxiliaram com o fornecimento das informações linguísticas necessárias, e está disponível para download (em código aberto) no site do Projeto Apertium[1].

Para avaliar a proposta, foram realizados experimentos com o corpus News Commentary, onde ele foi traduzido usando o Apertium em sua arquitetura original e o Apertium na arquitetura modificada (contendo o módulo *mweprocessor*). Com as traduções obtidas, foi feita uma análise quantitativa, através das medidas BLEU e NIST, e qualitativa, através da inspeção das frases onde houve diferença na tradução gerada. As análises mostraram que a nova arquitetura proposta obteve melhorias significativas na qualidade das traduções geradas onde estas EMPs ocorreram, mesmo com a utilização de um número pequeno de EMPs na avaliação.

Esses resultados também mostraram que a melhora na tradução tem uma relação forte com o tamanho dos dicionários utilizados. Assim, uma das possibilidades de trabalhos futuros é a investigação dessa melhora em corpora maiores, utilizando para isso, dicionários com um número maior de entradas. Outras possibilidades incluem:

- Otimização e teste do módulo *mweprocessor*, com o objetivo de torná-lo parte da arquitetura padrão do Apertium. O Apertium já é utilizado comercialmente em várias aplicações de tradução, onde melhorias seriam bem-vindas. No entanto, para que o novo módulo possa ser incluído, é necessário que ele se seja mais otimizado e robusto.

- Extensão do módulo *mweprocessor* para detectar EMPs contendo mais que 2 palavras. Ainda que, mesmo com essa restrição, seja possível cobrir boa parte das EMPs problemáticas, há muitos casos em que elas apresentam 3 ou mais componentes. A extensão do módulo significaria poder adicionar mais entradas nos dicionários de EMPs.

[1] http://apertium.svn.sourceforge.net/viewvc/apertium/branches/gsoc2009/debeck/

- Criação de uma interface intuitiva para edição dos dicionários de EMPs, facilitando sua criação e edição pelos especialistas. A maior parte dos dicionários do Apertium é feita manualmente e já existem propostas de interfaces para os formatos dos dicionários já existentes. Essas interfaces poderiam ser adaptadas para os dicionários de EMPs.

- Integração com um sistema de aquisição semi-automática de EMPs, visando a acelerar a criação dos dicionários. A idéia é utilizar os processos de aquisição semi-automática para que possam capturar não só a EMP em si, mas também as suas flexões e suas traduções. Um sistema desse tipo auxiliaria enormemente a investigação da tradução em corpora maiores, já que facilitaria a construção de dicionários com um número grande de entradas.

- Investigar a melhoria da tradução em textos de domínios específicos, adaptando terminologias e glossários para o formato dos dicionários do Apertium. O corpus utilizado nos experimentos é de domínio geral. Considerando que boa parte das terminologias são EMPs, observar a tradução em um corpus de domínio específico pode trazer resultados interessantes.

REFERÊNCIAS

[ARM 2007] ARMENTANO-OLLER, C. et al. Apertium, una plataforma de código abierto para el desarrollo de sistemas de traducción automática. In: FLOSS INTERNATIONAL CONFERENCE 2007, 2007. **Proceedings...** Servicio de Publicaciones de la Universidad de Cadiz, 2007. p.5–20.

[ARN 2008] ARNE MAUSER, S. H.; NEY, H. Automatic evaluation measures for statistical machine translation system optimization. In: SIXTH INTERNATIONAL LANGUAGE RESOURCES AND EVALUATION (LREC'08), 2008, Marrakech, Morocco. **Proceedings...** European Language Resources Association (ELRA), 2008. http://www.lrec-conf.org/proceedings/lrec2008/.

[BAL 2002] BALDWIN, T.; VILLAVICENCIO, A. Extracting the unextractable: a case study on verb-particles. In: PROC. OF THE 6TH CONFERENCE ON NATURAL LANGUAGE LEARNING (CONLL-2002, 2002. **Anais...** [S.l.: s.n.], 2002. p.98–104.

[COP 2005] COPESTAKE, A.; BRISCOE, T. Noun compounds revisited. In: TAIT, J. I. (Ed.). **Charting a new course**: natural language processing and information retrieval. Berlin: Springer, 2005.

[CUT 92] CUTTING, D. et al. A practical part-of-speech tagger. In: APPLIED NATURAL LANGUAGE PROCESSING, 1992, Morristown, NJ, USA. **Proceedings...** Association for Computational Linguistics, 1992. p.133–140.

[DOD 2002] DODDINGTON, G. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In: HUMAN LANGUAGE TECHNOLOGY RESEARCH, 2002, San Francisco, CA, USA. **Proceedings...** Morgan Kaufmann Publishers Inc., 2002. p.138–145.

[JAC 97] JACKENDOFF, R. Twistin' the night away. **Language**, v.73, p.534–59, 1997.

[JUR 2000] JURAFSKY, D.; MARTIN, J. H. **Speech and language processing**: an introduction to natural language processing, computational linguistics and speech recognition (prentice hall series in artificial intelligence). 1.ed. [S.l.]: Prentice Hall, 2000.

[KRI 2007] KRIEGER, M. et al. **Glossário de gestão ambiental**. [S.l.]: Editora Disal, 2007.

[LEV 66] LEVENSHTEIN, V. I. **Binary codes capable of correcting deletions, insertions, and reversals**. [S.l.: s.n.], 1966. (8).

[MIL 90] MILLER, G. A. et al. Introduction to wordnet: an on-line lexical database*. **Int J Lexicography**, v.3, n.4, p.235–244, January 1990.

[PAP 2001] PAPINENI, K. et al. **Bleu**: a method for automatic evaluation of machine translation. 2001.

[PRO 95] PROCTER, P. (Ed.). **Cambridge international dictionary of english**. Cambridge, UK: [s.n.], 1995.

[ROJ 2005] ROJAS, S. O.; FORCADA, M. L.; SÁNCHEZ, G. R. Construcción y minimización eficiente de transductores de letras a partir de diccionarios con paradigmas. **Procesamiento del Lenguaje Natural**, n.35, p.51–57, 2005.

[SAG 2002] SAG, I. A. et al. Multiword expressions: a pain in the neck for nlp. In: IN COMPUTATIONAL LINGUISTICS AND INTELLIGENT TEXT PROCESSING: THIRD INTERNATIONAL CONFERENCE (CICLing 2002), 2002, Berlin/Heidelberg. **Anais...** Springer, 2002.

[TIL 97] TILLMANN, C. et al. Accelerated dp based search for statistical translation. In: IN EUROPEAN CONF. ON SPEECH COMMUNICATION AND TECHNOLOGY, 1997. **Anais...** [S.l.: s.n.], 1997. p.2667–2670.

[VAU 68] VAUQUOIS, B. A survey of formal grammars and algorithms for recognition and transformation in mechanical translation. In: IFIP CONGRESS (2), 1968. **Anais...** [S.l.: s.n.], 1968. p.1114–1122.

[WHI 94] WHITE, J. S. The arpa mt evaluation methodologies: evolution, lessons, and further approaches. In: CONFERENCE OF THE ASSOCIATION FOR MACHINE TRANSLATION IN THE AMERICAS, 1994., 1994. **Proceedings...** [S.l.: s.n.], 1994. p.193–205.

[ZHA 2004] ZHANG, Y.; VOGEL, S. Measuring confidence intervals for the machine translation evaluation metrics. In: INTERNATIONAL CONFERENCE ON THEORETICAL AND METHODOLOGICAL ISSUES IN MACHINE TRANSLATION (TMI 2004), BALTIMORE, MD USA, OCTOBER 4-6, 2004, 2004. **Anais...** [S.l.: s.n.], 2004.