INSTITUTO SUPERIOR TÉCNICO
Universidade Técnica de Lisboa

Tradução Automática de Notícias Televisivas

Alexandre José Almeida Gusmão

Dissertação para obtenção do Grau de Mestre em

Engenharia Informática e de Computadores

Júri

Presidente:	Professora Doutora Ana Maria Severino de Almeida e Paiva
Orientador:	Professora Doutora Maria Luísa Torres Ribeiro Marques da Silva Coheur
Vogal 1:	Professor Doutor Diamantino António Caseiro
Vogal 2:	Professor Doutor Pável Pereira Calado

Novembro 2008

Agradecimentos

Gostaria de agradecer às seguintes pessoas pelo seu contributo na realização deste trabalho:

- Um agradecimento muito especial ao Professor Doutor Diamantino António Caseiro que, apesar de se encontrar entre Portugal e Estados Unidos da América, sempre esteve presente nos momentos certos, na hora das minhas dúvidas e questões. A sua ajuda e acompanhamento foram indispensáveis para a realização de uma dissertação desta natureza;
- À Professora Doutora Maria Luísa Torres Ribeiro Marques da Silva Coheur pela sua total disponibilidade e empenho na realização deste meu trabalho. A sua acção foi determinante para a linha orientadora deste projecto.
- Aos meus pais e restante família, que sempre me incentivaram a terminar este projecto e me ajudaram ao longo de todo o meu curso;
- A todos os meus amigos que sempre me animaram e ajudaram.

Lisboa, 19 de Novembro de 2008

Alexandre José Almeida Gusmão

Aos meus pais que me
proporcionaram a realização deste
curso

Resumo

Os sistemas de tradução automática de fala para texto têm como objectivo a compreensão de fala de uma dada língua fonte, por ouvintes não falantes dessa língua. Para além de inúmeras aplicações, um exemplo concreto deste tipo de sistema, é um tradutor automático de noticiários. A existir, este sistema permitiria a compreensão de noticiários por parte de não falantes da língua em que o telejornal é emitido, bem como, por exemplo, auxiliar pessoas com problemas de audição.

Neste trabalho descreve-se um sistema de tradução automático de fala transcrita, de Português para Inglês, no âmbito das notícias televisivas.

Nos primeiros anos da tradução automática, eram os sistemas de tradução automática baseados em regras aqueles que apresentavam melhores resultados. No entanto, com o aparecimento de grandes quantidades de corpora paralelos (Capítulo 3), com o aumento do poder computacional e com o melhoramento/aparecimento de novos modelos de tradução, os sistemas baseados em métodos estatísticos são os que, hoje em dia, apresentam os melhores resultados. Por este motivo, foram estes tipos de sistemas os escolhidos para a tarefa desta dissertação.

Os sistemas baseados em métodos estatísticos têm de ser treinados usando grandes quantidades de corpora paralelos (Português-Inglês). Dado a inexistência deste tipo de corpus no domínio das notícias televisivas, os treinos foram feitos recorrendo aos corpora do Parlamento Europeu. No entanto, para as tarefas de afinação do sistema, foram manualmente construídos corpora paralelos baseados em notícias televisivas, com base nas notícias da *Euronews*. Nesta tese apresentam-se as várias experiências efectuadas tendo em vista uma tradução de qualidade.

Abstract

Automatic speech to text translation systems have for main purpose the comprehension of speech from a source language, by those who do not speak it. Besides a lot of applications, an example of this type of system is an automatic machine translation in the broadcast news domain. These types of systems could allow a better comprehension of broadcast news for foreign listeners, as well as, for example, help people with some kind of hearing problem.

This paper describes an automatic translation system from speech to text, from Portuguese to English, in the broadcast news domain.

In the first years of automatic translation, rule-based systems produced the best results. However, the arrival of huge amounts of parallel corpora, the improvement of computational power and also the appearance of new translation models, resulted in the achievement of better results for translation systems based on statistical methods. For this reason, statistical machine translation methods were chosen to be used in the development of the machine translation system presented in this thesis.

Automatic translation systems based on statistical methods need a lot of parallel corpora (Portuguese - English) to perform the training. Due to the lack of this kind of corpus in broadcast news domain, the training process was executed with European Parliament corpora. However, to perform the tuning of the system, a parallel corpora based on broadcast news from Euronews were manually built. This thesis presents the various experiences done in order to achieve a translation with quality.

Palavras Chave
Keywords

Palavras Chave

Tradução Automática

Reconhecimento de Fala

Tradução de Texto

Modelo de Linguagem

Modelo de Tradução

Keywords

Statistical Machine Translation

Speech Recognition

Text Translation

Language Model

Translation Model

Índice

1 Introdução **1**

1.1 Motivação 1

1.2 Problemática e abordagem 1

1.3 Organização da tese 2

2 Estado da Arte **5**

2.1 Histórico 5

2.2 Tipos de Sistema de tradução 7

2.2.1 Sistemas de tradução baseados em regras 7

2.2.1.1 Sistemas directos 7

2.2.1.2 Sistemas de transferência 7

2.2.1.3 Sistemas interlíngua 7

2.2.2 Sistema de tradução que utilizam métodos estatísticos 9

2.2.2.1 Vantagens e Desvantagens 9

2.2.2.2 Modelação 10

2.3 Modelos de Tradução 11

2.3.1 Modelos de tradução IBM 12

2.3.2 Sistemas baseados em Segmentos 13

2.3.3 Sistemas baseados em sintaxe 16

2.4 Avaliação 17

2.4.1 WER (*Word Error Rate*) . . . 18
2.4.2 PER (*Position Independent Word Error Rate*) . . . 18
2.4.3 BLEU (*Bilingual Evaluation Understudy*) . . . 18
2.4.4 NIST (*National Institute of Standards and Technology*) . . . 19
2.5 Reconhecimento e Tradução da fala . . . 19
2.6 Conclusão . . . 21

3 O sistema de tradução **23**
3.1 Descrição geral do Sistema . . . 23
3.1.1 Modelo de linguagem . . . 23
3.1.2 Normalização do corpus de treino . . . 24
3.1.3 Treino do sistema . . . 25
3.1.4 Filtragem da tabela de segmentos . . . 25
3.1.5 A afinação de parâmetros (Tuning) . . . 26
3.1.6 Avaliação . . . 27
3.2 Ferramentas utilizadas . . . 27
3.2.1 Modelos de linguagem . . . 27
3.2.2 Treino do sistema . . . 28
3.3 Os Corpora utilizados . . . 29

4 Tradução de Texto e Avaliação **33**
4.1 Introdução . . . 33
4.1.1 Parlamento Europeu . . . 33
4.1.2 Parlamento Europeu Filtrado . . . 34
4.2 Baseline . . . 35
4.3 Experiência 1 . . . 36

4.4 Experiência 2 39
4.5 Experiência 3 41
4.6 Experiência 4 43
4.7 Experiência 5 45
4.7.1 *Feature* de contagem do número de palavras 45
4.7.2 *Feature Part-Of-Speech* (POS) 46
4.8 Ponto de Comparação 48
4.9 Resumo 48

5 Conclusões e Trabalho Futuro **51**

Lista de Figuras

2.1 Triângulo de *Vauquois* 8

2.2 Exemplo de um alinhamento possível 12

2.3 Operações de Reordenamento, Introdução e Tradução 17

3.1 Normalização do corpus de teino 24

3.2 Treino do sistema 25

3.3 Filtragem da tabela de segmentos 26

4.1 Interpolação dos Modelos de Linguagem 42

Lista de Tabelas

2.1 Taxas de erro na tradução da fala do projecto TC-STAR 21

2.2 Tipos de entrada para o sistema de tradução 21

3.1 Exemplo de pares de frases retiradas do *website* da *Euronews*. 31

4.1 Sistema de tradução, baseado exclusivamente nas sessões do Parlamento Europeu e respectivos valores de BLEU. 34

4.2 Sistema de tradução, baseado nas sessões do Parlamento Europeu mas já com a tabela de segmentos filtrada apenas com as palavras pertencentes aos corpus de desenvolvimento. 35

4.3 Sistema de tradução, com a tabela de segmentos filtrada, corpus de desenvolvimento, teste e modelo de linguagem baseados nas notícias televisivas e corpus de treino baseado nas sessões do Parlamento Europeu. 36

4.4 Frases comparáveis, mas não necessariamente traduções directas 37

4.5 Alteração do corpus de desenvolvimento e teste 38

4.6 Sistema de tradução, com tabela de segmentos filtrada, corpus do modelo de linguagem baseado nas notícias televisivas, corpus de *Tuning* (desenvolvimento) e teste alterados e, corpus de treino baseado nas sessões do Parlamento Europeu. 39

4.7 Perplexidade dos modelos de linguagem das notícias televisivas, para diferentes *cut-off*. 40

4.8 Perplexidade dos modelos de linguagem dos jornais, para diferentes *cut-off*. . . . 40

4.9 Perplexidade do Modelo de Linguagem Interpolado. 41

4.10 Sistema de tradução, com tabela de segmentos filtrada, modelo de linguagem interpolado, corpus de *Tuning* (desenvolvimento) e teste alterados e, corpus de treino baseado nas sessões do Parlamento Europeu. 42

4.11 Traduções obtidas e respectivas frases de entrada e referência 43

4.12 Sistema de tradução com recaser automático. 44

4.13 Contribuição da *feature* de contagem do número de palavras no processo de *rescoring*. 45

4.14 Equivalência de etiquetas de POS entre Português e Inglês 46

4.15 Padrões de penalização para Inglês 47

4.16 Contribuição da *feature* POS. 47

4.17 Contribuição das *features* de diferença do número de palavras e POS. 48

4.18 Sistema de tradução do google e respectivos valores de BLEU obtidos com o mesmo corpus de desenvolvimento. 48

4.19 Traduções obtidas pelo *Google* Vs Sistema de tradução. 49

5.1 Corpora utilizado e respectivas descrições. 53

5.2 Experiências realizadas. 53

1 Introdução

1.1 Motivação

Os sistemas de tradução automática de fala para texto têm como principal objectivo permitir a compreensão de fala de uma dada língua fonte, por ouvintes não falantes dessa língua. Para além de inúmeras aplicações, um exemplo concreto deste tipo de sistema, é um tradutor automático de noticiários. A existir, este sistema permitiria a compreensão de noticiários por parte de não falantes da língua em que o telejornal é emitido, bem como auxiliar pessoas com problemas de audição.

Esta tese apresenta um sistema automático de tradução de fala transcrita para texto, de notícias televisivas.

1.2 Problemática e abordagem

Um sistema de tradução automática de fala para texto pode ser dividido em dois subsistemas em que o primeiro é o responsável pelo reconhecimento de fala e o segundo pela tradução. De modo a limitar a tarefa, foi usado o AUDIMUS (Meinedo 2003), um sistema de reconhecimento de fala contínua para o Português Europeu, na tarefa de reconhecimento de notícias televisivas. Deste modo, a problemática da tese passou a centrar-se apenas na tarefa da tradução automática de fala transcrita.

A tradução automática é uma tarefa complicada, dado que cada língua tem características próprias. Para além disso inúmeras palavras podem ter vários significados e consequentemente diversas traduções, levando a que cada frase possa ter mais que uma tradução.

Adicionalmente, pode não existir uma escolha óbvia para a tradução de uma palavra.

São duas as principais linhas de trabalho dedicadas à problemática da tradução automática, dividindo-se os sistemas envolvidos em: a) sistemas baseados em regras; b) sistemas baseados em métodos estatísticos. Os sistemas baseados em métodos estatísticos são mais facilmente integrados com os sistemas de reconhecimento de fala e, nos últimos anos, são os que têm apresentado os melhores resultados. Dado que hoje em dia existem sistemas baseados em métodos estatísticos de qualidade, com grande possibilidade de parametrização, de distribuição livre e que implementam diversos modelos de tradução, estes foram a base de trabalho desta tese. Há ainda que destacar, de entre os modelos usados nos sistemas baseados em métodos estatísticos, os modelos baseados em segmentos (em inglês *phrase-based*), dado que foram os escolhidos para realizar esta tarefa, sendo portanto estudados com uma maior profundidade.

No entanto, os sistemas baseados em métodos estatísticos, necessitam de grandes quantidades de textos paralelos (isto é, conjunto de frases escritas numa linguagem associadas com a respectiva tradução), do domínio pretendido. Assim, um dos principais problemas a resolver foi a falta destes textos para o domínio das notícias televisivas. Este problema pôde ser parcialmente solucionado treinando o sistema com as actas do Parlamento Europeu, uma vez que neste domínio existem mais de 30 milhões de palavras transcritas e traduzidas para a grande parte das línguas europeias. No entanto, dado que este domínio não é exactamente o das notícias televisivas, foram feitas várias tentativas para melhorar a qualidade da tradução.

Nesta dissertação serão então apresentadas as várias tentativas de construir e melhorar um sistema de tradução de fala transcrita em Português para Inglês no âmbito das notícias televisivas, os problemas encontrados e a forma como foram solucionados. Serão igualmente apresentados os resultados das diversas experiências efectuadas.

1.3 Organização da tese

No Capítulo 2 apresenta-se um breve estado da arte de tradução automática, focando essencialmente nos sistemas de tradução automática baseados em métodos estatísticos; no Capítulo 3

são apresentadas as diversas ferramentas utilizadas na construção de um sistema de tradução; no Capítulo 4 são descritas todas as experiências realizadas, bem como os resultados obtidos em cada uma delas. Finalmente, no Capítulo 5 são apresentadas as conclusões e algumas propostas de trabalho futuro.

2 Estado da Arte

2.1 Histórico

O aparecimento dos primeiros computadores e a sua evolução nos anos 40 e 50, estimulou o trabalho na área de tradução automática. Em 1954, foi realizada uma apresentação pública acerca da fiabilidade dos sistemas automáticos, com a contribuição da IBM e Universidade Georgetown. Apesar da gramática e vocabulários utilizados na altura serem muitos restritos, foi o suficiente para impressionar e cativar não só as atenções para a área da tradução automática nos Estados Unidos, como também estimular a introdução de projectos de tradução automática por todo o mundo.

Foi já na segunda metade do século XX que se deu a grande transformação da tradução automática, transitando de pesquisas a nível universitário para a produção em massa a nível de mercado. Obtiveram-se os primeiros resultados através da utilização de dicionários bilingue, sendo que a informação sintáctica e semântica foram introduzidas por regras escritas por especialistas.

Desde há muito tempo que têm sido feitas diversas abordagens à tradução automática, em que inicialmente as que melhores resultados apresentaram foram as baseadas em regras. Como exemplo de um sistema baseado em regras existe o *Systran*, o motor que se encontra por detrás tanto do *Google translate* como também do *Babelfish*.

No entanto, foi em 1993, com o trabalho desenvolvido por Brown (Brown 2003) que o interesse pelos métodos estatísticos renasceu. Esta aproximação revelou-se uma mais valia para a área da tradução automática, tendo sido produzidos melhores resultados em relação aos sistemas baseados em regras (Ney 2005). Desde essa altura que têm surgido alguns melhoramentos significativos na área de tradução automática, muito devido ao facto do constante

aumento do poder computacional das máquinas, aos melhoramentos feitos nos modelos que se encontram na base dos módulos de tradução bem como os algoritmos utilizados e, acima de tudo, devido ao aumento de corpora multilingue, essencial para obter bons resultados através de abordagens estatísticas. A destacar também o aparecimento de métricas de avaliação commumente aceites, que permitem testar e comparar os sistemas (Nicola Ueffing 2004) (Papinemi 2001) (Doddington 2002).

Dentro das abordagens estatísticas, destacam-se os sistemas baseados em segmentos (em inglês *phrase based*), isto é, baseados em pares de segmentos de frases escritos em línguas diferentes, considerados equivalências de tradução. Dado que este será o modelo utilizado no sistema de tradução a desenvolver neste trabalho, será explicado em mais detalhe nos capítulos seguintes.
Até muito recentemente esta era a abordagem que melhores resultados apresentava. No entanto, recentemente foram apresentadas melhorias com a utilização de informação sintáctica (Marcu 2006).

Foram feitas muitas tentativas para a obtenção de melhorias significativas nos resultados da tradução automática baseada em métodos estatísticos através da introdução de mais informação linguística, mas infelizmente a maioria não revelou resultados positivos. No entanto, foram obtidos alguns melhoramentos através da introdução de informação morfológica (Gispert 2006).

Actualmente a maioria das tarefas realizadas no âmbito da tradução automática baseada em métodos estatísticos não contempla um vocabulário muito extenso (até 10000 palavras) (Ney 2005). No entanto, a utilização de uma maior quantidade de vocabulário é um desafio que está já a ser enfrentado por algumas instituições, especialmente com o desenvolvimento do projecto TC-STAR (http://www.tc-star.org/).

Podem ser encontradas mais informações acerca da história da tradução automática em (Hutchins 1986) e sobre o seu futuro em (Hutchins 2001) e (Knight 2005).

2.2 Tipos de Sistema de tradução

2.2.1 Sistemas de tradução baseados em regras

Geralmente, os sistemas de tradução baseados em regras criam uma representação intermédia simbólica, com base do texto entrada, através da qual o texto na língua destino é gerado.
No entanto, este tipo de sistemas apresentam algumas desvantagens relevantes, nomeadamente:

- É necessária muita quantidade de informação linguística;
- É muito difícil escrever regras que cubram toda a língua.

Os sistemas de tradução automática podem ser classificados de acordo com a sua estratégia de tradução em sistemas directos, sistemas de transferência e sistemas interlíngua.

2.2.1.1 Sistemas directos

Os sistemas directos envolvem emparelhamento de palavras, seguido de reordenações na frase gerada, de acordo com a ordem das palavras na língua de destino. Esta foi uma estratégia utilizada inicialmente por muitos sistemas de tradução. Grande parte dos sistemas estatísticos actuais são directos.

2.2.1.2 Sistemas de transferência

Este tipo de sistemas envolve uma primeira fase de análise morfológica e sintáctica da frase de entrada (e por vezes semântica), de modo a obter uma representaçao intermédia (entre a língua de orígem e destino). Após esta fase de análise, o processo de tradução é realizado através da conversão da representação final obtida, numa representação pertencente à língua destino, com a utilização de dicionários bilingue e regras gramaticais..

2.2.1.3 Sistemas interlíngua

Nos sistemas interlíngua, as frases na língua fonte são analisadas e transformadas para uma língua ou representação neutra, através da qual é gerada a frase alvo, por vezes após alguma

manipulação da representação ou língua neutra. A grande promessa deste tipo de sistemas é a independência entre as línguas fonte e destino, ou seja, para cada língua, são apenas necessárias uma fase de análise para a interlíngua e uma fase de geração para a língua pretendida.

Estas noções são ilustradas utilizando o triângulo de *Vauquois* representado na figura 2.1. O triângulo ilustra na direcção vertical a quantidade de esforço necessária para a análise e geração e, na horizontal a quantidade de esforço necessária para a transferência (transferência de uma língua para a outra). No topo do triângulo o esforço para a transferência é mínimo enquanto que para a análise e geração é máximo (Trujillo 1999).

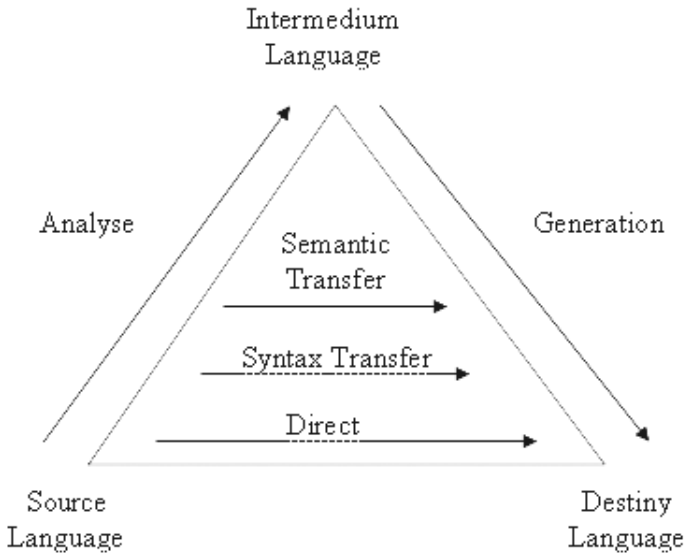

Figura 2.1: Triângulo de *Vauquois*

Uma forma de melhorar a qualidade destes sistemas de tradução é modificar ou adicionar regras de geração ou de análise. No entanto, este método requer muita informação linguística e não é garantido que os resultados sejam melhores. Ao serem alteradas certas regras, podem não só ser modificadas frases incorrectas para frases correctas, como também podem ocorrer efeitos negativos em frases correctas (Charoenpornsawat 2002).

Dado que o foco desta dissertação está em sistemas de tradução automática baseados em métodos estatísticos, o tema da tradução baseada em regras não será abordado de um

modo aprofundado. Para mais informações acerca deste tema, pode consultar-se, por exemplo, o livro (D Jurafsky 2000).

2.2.2 Sistema de tradução que utilizam métodos estatísticos

Nos sistemas estatísticos, ao invés de serem utilizadas regras escritas por linguistas, são utilizadas distribuições de probabilidades. Quando se pretende construir um sistema estatístico, quanto mais conhecimento à priori for utilizado, melhores resultados, em geral, se podem obter (Tetko 1995). Este conhecimento permite definir modelos de probabilidade mais apropriados para o objectivo pretendido. Por exemplo, a utilização de textos de treino constituídos por frases e respectivas traduções podem ser utilizados como conhecimento à priori. Um dos grandes objectivos da tradução automática baseada em métodos estatísticos é o cálculo de probabilidades dos vários intervenientes no processo de tradução, de modo a obter um sistema que devolva traduções aceitáveis (Ney 2005).

2.2.2.1 Vantagens e Desvantagens

Os sistemas baseados em métodos estatísticos apresentam como vantagens os seguintes aspectos (Ney 2005) (Koehn 2005b):

1. Utilização de funções de distribuição de probabilidade, que permitem ao sistema estimar qual a melhor tradução para uma determinada frase de entrada, tendo em conta o contexto em que esta se encontra inserida. O valor da probabilidade é obtido através da combinação de múltiplos valores individuais, obtidos por exemplo, através da intervenção do modelo de linguagem, modelo de tradução ou modelo de distorção. Com a utilização de probabilidades deixa de ser necessário o desenvolvimento manual de regras linguísticas (sistemas baseados em regras);

2. Existem algoritmos muito eficientes que aprendem automaticamente estas probabilidades, sem a intervenção humana;

3. Ao contrário da aproximação baseada em regras, a tradução automática baseada em métodos estatísticos não requer o desenvolvimento manual das regras linguísticas, o que para além de ser muito caro, tem a desvantagem adicional de normalmente estas não poderem ser generalizadas para outras linguagens.

No entanto, a aproximação estatística não apresenta apenas vantagens. Os seguintes aspectos podem ser apontados como desvantagens da utilização de métodos estatísticos:

1. Uma vez que os sistemas são treinados com base em exemplos de determinado domínio temático e vocabulário, não apresentam uma boa capacidade de adaptação para outros domínios temáticos, sendo necessário retreiná-lo para cada um deles;
2. A maior parte dos sistemas actuais ainda não lidam explicitamente com a sintaxe das frases;
3. A linguagem matemática utilizada (nos algoritmos referidos nas vantagens destes sistemas) é muito complexa, o que dificulta o melhoramento dos resultados da tradução automática baseada em métodos estatísticos (Knight 2006).

2.2.2.2 Modelação

A aproximação estatística pode ser explicada da seguinte forma:
Dada uma frase F numa linguagem fonte, pretende-se traduzi-la para uma frase E, numa determinada linguagem destino.
É considerada a probabilidade de distribuição $P(F \mid E)$ de todos os pares possíveis (F, E) e é seleccionada a frase traduzida Ê com a maior probabilidade, ou seja,

$$E = argmax_E P(E \mid F) \tag{2.1}$$

Através do teorema de Bayes tem-se:

$$P(E \mid F) = \frac{P(E \wedge F)}{P(F)} = \frac{P(F \mid E).P(E)}{P(F)} \tag{2.2}$$

$$E = argmax_E \frac{P(F \mid E).P(E)}{P(F)} \tag{2.3}$$

Como $P(F)$ é constante obtém-se o seguinte:

$$E = argmax_E P(F \mid E).P(E) \tag{2.4}$$

Em que:

- $P(F \mid E)$ representa o *modelo de tradução* - probabilidade da frase F ser a tradução da frase E.

- $P(E)$ representa o *modelo de fluência* - probabilidade da frase traduzida pertencer à língua E.

A utilização do teorema de Bayes, permite assim aplicar o princípio de dividir para conquistar. Deste modo, a utilização independente destes dois modelos pode não produzir resultados satisfatórios, o mesmo não acontecendo quando são utilizados em conjunto, uma vez que as frases mais correctas serão favorecidas ao ser feito o seu produto.

A utilização de modelos de probabilidade tem como objectivo promover uma ligação entre os dados de entrada e os dados de saída que serão produzidos pelo sistema de tradução. Uma vez fixado o modelo, os seus parâmetros (distribuição de probabilidades) serão estimados com base em exemplos. É neste ponto que o conhecimento linguístico poderá ajudar a melhorar os modelos existentes (Ney 2005).
Os modelos de probabilidades modelam também a forma como se calcula a probabilidade da tradução de cada frase através das probabilidades de cada palavra.

2.3 Modelos de Tradução

Nos modelos de tradução apresentados a seguir, uma frase de uma língua fonte pode produzir a mesma frase na língua alvo por diferentes formas, devido aos diferentes tipos de alinhamento que podem ser utilizados. Um exemplo de um alinhamento não é mais do que um conjunto de ligações entre palavras fonte e palavras alvo, em que cada palavra alvo se encontra ligada a uma só palavra fonte. Pode observar-se na figura 2.2 um alinhamento entre '*an old house*' e 'uma casa velha' (Knight 2006).
A palavra 'NULL' é sempre introduzida no início de cada frase fonte para o caso de existir uma ou mais palavras em relação à frase gerada.
Supondo então que a frase fonte contém l palavras e que a frase gerada contém m palavras, existem $(l+1)^m$ alinhamentos possíveis.
Deste modo, $P(F \mid E) = \sum_a P(F, a \mid E)$, em que a representa um alinhamento. Cada modelo

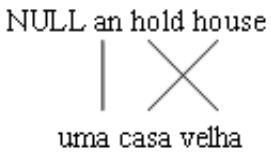

Figura 2.2: Exemplo de um alinhamento possível

define a probabilidade de um alinhamento, todos de modo diferente.

Cada modelo de tradução sugere uma possível computação para a probabilidade condicional $P(F \mid E)$, a qual é designada por verosimilhança (em inglês *likelihood*) da tradução (F, E). Esta verosimilhança não é mais do que uma função com uma grande quantidade de parâmetros livres, que devem ser estimados num processo ao qual chamamos de *treino*.

2.3.1 Modelos de tradução IBM

Em seguida serão descritos os modelos de tradução IBM com algum detalhe, mas para uma informação mais precisa, pode consultar-se (Brown 2003), um artigo fundamental que descreve com bastante detalhe os métodos de tradução baseados em estatística e onde foram introduzidos os chamados modelos IBM. Este artigo provocou o renascimento do interesse sobre os métodos estatísticos utilizados na tradução.

De um modo resumido, nos modelos IBM 1 e 2, em primeiro lugar é escolhido o tamanho da frase alvo. Posteriormente, para cada posição da frase alvo, é tomada a decisão de como a ligar à frase fonte e qual a palavra da língua alvo que será colocada nessa posição. No modelo 1 são assumidas todas as ligações possíveis entre a frase fonte e frase alvo, implicando deste modo que a ordem das palavras nas duas frases não afecte o valor da probabilidade $P(F \mid E)$. Já no modelo 2, a probabilidade $P(F \mid E)$ depende da ordem das palavras nas duas frases. Estes modelos permitem obter alinhamentos entre pares de palavras de duas línguas, mas muitas das vezes são insatisfatórios.

Nos modelos 3, 4 e 5, a frase alvo é gerada, escolhendo para cada palavra na frase fonte, em primeiro lugar o número de palavras na frase alvo que irão estar ligadas a ela (fertilidade

da palavra), posteriormente a identidade (tradução) dessas palavras na língua alvo e, finalmente, a posição na frase alvo que essas palavras irão ocupar (distorção da palavra). É esta última etapa que determina as ligações entre as frases fonte e alvo e é também neste ponto que estes três modelos diferem entre si. No modelo 3, tal como no modelo 2, a probabilidade de uma ligação depende das posições a que estão ligadas e no tamanho das frases alvo e fonte. No modelo 4, a probabilidade de uma ligação depende adicionalmente dos identificadores das palavras ligadas nas duas línguas e também das posições de quaisquer outras palavras na língua alvo que estejam ligadas à mesma palavra na língua fonte.

Devido à dificuldade em estimar directamente os parâmetros dos modelos mais complexos, o modelo 1 é utilizado para gerar estimativas iniciais para os parâmetros do modelo 2. O modelo 2 e restantes modelos são sucessivamente iniciados pelos parâmetros do modelo anterior e retreinados utilizando um algoritmo do tipo EM (Expectation-Maximization). Para mais informações sobre algoritmos do tipo EM, consultar (Dempster 1977).

Relativamente ao treino, os modelos IBM 1 e 2 são treinados de forma eficiente através da utilização de algoritmos do tipo EM. Os modelos IBM 3, 4 e 5 são de implementação extremamente complexa. Actualmente a ferramenta de treino mais utilizada é o Giza++ (Josef & Ney 2003).

Quanto à procura, é frequente a utilização do decoder *ReWrite* (Germann 2001) (Germann 2003), útil até ao modelo IBM 4.

2.3.2 Sistemas baseados em Segmentos

Segmentos são pares de segmentos de frases escritos em línguas diferentes mas que são equivalências de tradução. A tradução é efectuada através da selecção de um conjunto de segmentos compatíveis com a frase de entrada, reordenando posteriormente as palavras na língua alvo.

A qualidade dos sistemas de tradução baseados em métodos estatísticos melhorou consideravelmente com a introdução da tradução por segmentos por parte de vários investigadores. Por

exemplo, (Yamada 2001), onde a tradução de segmentos foi utilizada em sistemas de tradução baseados em sintaxe, (Daniel Marcu 2002) introduziu um modelo de probabilidades conjuntas para a tradução de segmentos e, nos sistemas de tradução estatísticos baseados em palavras da CMU e IBM foi também introduzida a capacidade de tradução de segmentos (Philip Koehn 2003).

Quanto ao treino, os sistemas baseados em segmentos são treinados com base em textos paralelos alinhados, sendo utilizados modelos de tradução baseados em palavras para alinhar cada par de frases no corpus, a nível das palavras. Este alinhamento é melhorado através de heurísticas específicas ou através da combinação de múltiplos alinhamentos. Os segmentos do par de línguas em causa são extraídos destes últimos alinhamentos. É também frequente a combinação de vários modelos (Franz Och 2004) (Och 2003) utilizando o princípio da máxima entropia (Berger 1996). Neste caso, os parâmetros da combinação são estimados, minimizando directamente o erro de tradução quando medido usando medidas automáticas, como o BLEU ou o NIST.

De um modo sucinto, as frases (qualquer sequência de palavras) dadas como entrada são segmentadas (separadas) em segmentos. Cada segmento é traduzido na língua alvo, e finalmente são reordenados de modo a gerar frases dessa língua (Koehn 2005b).

Devido à elevada complexidade que a tarefa de procura (escolha da melhor frase traduzida) apresenta, torna-se crucial a implementação de algoritmos eficientes. Os sistemas de tradução baseados em segmentos utilizam essencialmente dois algoritmos de procura, A* e *Dynamic Programing Beam*. Estes algoritmos seleccionam a melhor frase gerada de um conjunto enorme de frases geradas (Ney 2005).

Dynamic Programing Beam

Com esta estratégia, as palavras da frase fonte são processadas da esquerda para a direita, permitindo ainda a troca de posições de palavras até um certo grau. O espaço de procura é explorado em largura-primeiro. Este algoritmo de procura é baseado em programação dinâmica e deste modo aplica várias técnicas de poda, de modo a restringir o número de

hipóteses a considerar (Nicola Ueffing 2001).

A*

Em A*, todas as hipóteses de procura (frases hipóteses de tradução) são guardadas numa fila de prioridades (Nicola Ueffing 2001). A procura A* pode ser descrita da seguinte forma:

1. Iniciar a fila de prioridades com a hipótese vazia;
2. Remover a hipótese com maior pontuação da fila de prioridades;
3. Se esta hipótese for uma hipótese objectivo esta hipótese é devolvida e é terminado o processo de procura;
4. Caso contrário, produzem-se todas as extensões a esta hipótese e colocam-se na fila de prioridades;
5. Voltar ao segundo passo.

Relativamente aos modelos de tradução palavra a palavra (modelos IBM), os baseados em segmentos apresentam as seguintes vantagens (Quirk 2006):

- Não composição - Os segmentos permitem a tradução, de segmentos fixos ou não composicionais, em uma unidade. Deste modo é simplificada a implementação da tradução e melhorada a sua qualidade;
- Reordenação local - Visto que os pares de segmentos consistem em subfrases memorizadas do conjunto de treino, é muito provável que sejam produzidos reordenamentos locais (ao nível do segmento) correctos;
- Informação de contexto - Mesmo quando uma palavra gerada na língua alvo pareça estar correcta, a introdução de informação de contexto introduzida pelo segmento normalmente melhora a qualidade de tradução.

2.3.3 Sistemas baseados em sintaxe

Apesar dos sistemas baseados em segmentos terem apresentado vários melhoramentos relativamente ao modelos IBM, existem ainda alguns problemas por resolver, como por exemplo:

- Não descontinuidade nos segmentos - Nenhum segmento em qualquer língua pode conter descontinuidades, o que acontece em diversas línguas, como por exemplo no francês, em que a negação (*ne pas*) é feita de modo descontínuo, uma vez que o verbo aparece entre estas duas palavras;
- Reordenação global - Uma vez que não existe uma estratégia eficiente de reordenação global das frases geradas, é necessário considerar todas as permutações possíveis. Este é um processo que não utiliza informação sintáctica da língua de destino e é também computacionalmente pesado.

Os sistemas baseados em sintaxe procuram resolver estes problemas, que são ainda apresentados pelos modelos baseados em segmentos.

É sabido que pares de linguagens com ordens de palavras muito diferentes, como por exemplo o Inglês e o Japonês, não são muito bem modeladas através dos modelos de tradução existentes. De modo a incorporar aspectos estruturais da linguagem, pode por exemplo, em primeiro lugar, efectuar-se o pré-processamento da frase de entrada através de um analisador sintáctico. São realizadas então diversas operações em cada nó da árvore de *parse*, sendo estas, operações de reordenamento de nós filho, introdução de palavras extra em cada nó, e finalmente a tradução das palavras que se encontram nas folhas. A figura 2.3 ilustra um exemplo da aplicação destas operações (Yamada 2001).
Relativamente à operação de reordenamento, esta é utilizada para modelar a tradução entre linguagens que apresentam palavras muito diferentes, como por exemplo o caso do Inglês e do Japonês. A operação de introdução de palavras extra é utilizada para capturar diferenças linguísticas em casos sintácticos específicos, por exemplo, em Inglês e Francês são utilizadas posições estruturais para cada caso específico.

Vantagens dos sistemas baseados em sintaxe:

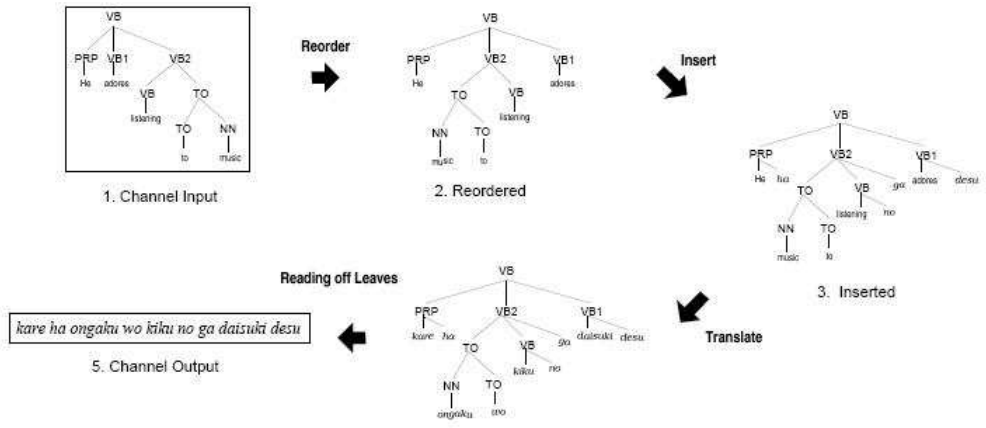

Figura 2.3: Operações de Reordenamento, Introdução e Tradução

- Algumas transformações durante a tradução (reordenação, introdução e remoção de palavras) podem ser explicadas de uma forma mais simples com a utilização de conceitos sintácticos;
- A análise sintáctica da frase de entrada fornece conhecimento adicional que pode ser explorado;
- A utilização de sintaxe na frase traduzida permite a utilização de modelos de linguagem sintácticos que ajudam a garantir que a saída do sistema seja gramatical (Koehn 2003).

Até muito recentemente, os sistemas que apresentavam melhores resultados eram os baseados em segmentos. Só muito recentemente foram apresentadas melhorias com a utilização de informação sintáctica (Marcu 2006).

Relativamente ao treino e procura, existem ainda diversos problemas por resolver para a criação de *decoders* eficientes, sendo esta uma área que se encontra em grande actividade.

2.4 Avaliação

A avaliação da tradução automática por parte de humanos, para além de ser muito cara, leva muito tempo a ser efectuada, tornando-se num grande problema para quem desenvolve sistemas de tradução automática. No entanto, é necessário fazer uma monitorização contínua aos

resultados apresentados pelos sistemas, de modo a detectar problemas/traduções incorrectas, para que de algum modo os sistemas possam estar constantemente a ser actualizados e melhorados, produzindo resultados cada vez mais fiáveis.
Assim, na última década foram desenvolvidas medidas de erro independentes em relação à língua e que tornam o processo de avaliação mais rápido e mais barato.

2.4.1 WER (*Word Error Rate*)

A medida WER (Nicola Ueffing 2004) é muito utilizada em sistemas de reconhecimento de fala e baseada na distância de Levenshtein. A distância de Levenshtein é uma medida de similaridade entre duas *strings* e é computada através do número de substituições, inserções e remoções necessárias para converter a *string* gerada na *string* referência. Para mais informações sobre a distância de Levenshtein consultar (http://www.merriampark.com/ld.htm) (http://www.iar.unicamp.br/suporte/php/function.levenshtein.html).

2.4.2 PER (*Position Independent Word Error Rate*)

Um aspecto em falta na medida de erro WER é sem dúvida o facto de não permitir o movimento de palavras ou blocos (Nicola Ueffing 2004). A ordem das palavras de duas frases traduzidas pode ser diferente mesmo que sejam ambas traduções correctas. De modo a superar este problema, a medida de erro PER compara as palavras nas duas frases sem ter em conta a ordem das palavras. As palavras sem ligação representam erros de substituição, palavras que faltam representam erros de remoção e palavras adicionais representam erros de introdução.

2.4.3 BLEU (*Bilingual Evaluation Understudy*)

A medida BLEU (Papinemi 2001) é uma métrica de avaliação automática que permite comparar o número de palavras comuns entre várias frases candidatas e frases de referência. Deste modo, a primeira tarefa do BLEU é simplesmente comparar o número de palavras partilhadas entre os candidatos e as referências que existem. Quanto mais palavras forem partilhadas por um candidato, melhor é a sua tradução. O próximo passo é determinar a precisão e, para isso, são contadas as palavras dos candidatos que aparecem nas referências e divide-se pelo número de palavras da frase candidata.

Esta é uma medida baseada no emparelhamento de *n-grams*[1], em que, ao contrário da medida PER, onde apenas é verificado se cada palavra da frase traduzida se encontra presente nas frases correctas, é verificado se sequências de palavras (sequências até 4 palavras) se encontram nessas mesmas frases. É feita uma contagem do número de *n-grams* comuns entre a frase traduzida e a frase que supostamente é a tradução correcta (frase referência) e, é atribuído um valor a essa frase através da média geométrica entre as várias contágens.

2.4.4 NIST (*National Institute of Standards and Technology*)

A medida NIST (Doddington 2002) é uma variante do BLEU, onde é também utilizado o emparelhamento de *n-grams*. Ao contrário do BLEU, é utilizada a média aritmética de *n-grams* em vez da média geométrica para a contagem do número de *n-grams* comuns entre a frase traduzida e as frases que supostamente são traduções correctas.

2.5 Reconhecimento e Tradução da fala

Um sistema automático de tradução de fala é composto essencialmente por dois sistemas sequenciais: um de reconhecimento de fala para texto e outro de tradução do texto para o idioma desejado. Uma vez que estas áreas apresentam ainda problemas que não estão resolvidos, a tradução automática da fala está ainda longe de ser concretizada com sucesso. No entanto, mesmo que existissem sistemas de reconhecimento de fala e tradução de texto perfeitos, uma vez que a fala é linguisticamente diferente do texto escrito (devido à espontaneidade que o orador utiliza nas suas palavras, à falta de tempo para a preparação da frase, às hesitações e paragens sucessivas e também devido à própria linguagem utilizada na fala ser totalmente diferente da utilizada na escrita), a sua tradução continuaria a ser um problema de difícil resolução.

No Laboratório de Língua Falada (L^2F) existem alguns sistemas de reconhecimento de fala, nomeadamente o AUDIMUS (Meinedo 2003), um sistema de reconhecimento de fala Portuguesa no domínio das notícias televisivas. Existe também um sistema de tradução de

[1]Sub-Sequência de N items de uma frase. Os items podem ser letras ou palavras. N-gram de dimensão 1 é designado por *unigram*, dimensão 2 é designado por bigram, dimensão 3 é designado por trigrama e dimensão 4 ou superior é simplesmente designado por N-Grama.

fala para texto (Diamantino Caseiro 2006) mas para um domínio mais restrito, nomeadamente para marcação de quartos de um hotel. No entanto estes são sistemas que apresentam ainda alguns problemas por resolver.

Entre vários sistemas de tradução de fala existentes, o sistema desenvolvido pelo projecto Europeu TC-STAR[2] (http://www.tc-star.org/) é aquele que talvez se assemelhe mais com o desenvolvido durante esta tese, devido a contemplar um vocabulário extenso (mais de 60000 palavras) e também por ser um projecto cujo domínio (tradução de sessões do Parlamento Europeu) se aproxima mais das notícias televisivas. O projecto TC-STAR contempla essencialmente três línguas, Inglês, Espanhol e Chinês. O principal objectivo deste projecto é efectuar uma pesquisa profunda acerca da tradução automática da fala, de modo a reduzir a distância de performance entre a tradução automática e humana.
A maior parte dos outros sistemas de tradução de fala existentes são de domínio limitado, por exemplo, domínio turístico ou marcação de reuniões (Wahlster 2000).

Relativamente ao estado da arte em sistemas de reconhecimento de fala no domínio das notícias televisivas, os sistemas de reconhecimento de língua inglesa apresentam taxas de erro menores do que 13% (Nguyen 2005) em 10 vezes tempo real, utilizando a medida automática WER, e menores do que 16% (Matsoukas 2005) em tempo real. Em relação ao sistema AUDIMUS, este apresenta uma taxa de erro de 11.3% em condições de estúdio (pouco ruído de fundo, boa qualidade de som) e, uma taxa de WER de 23.5% em todas as condições (Rui Amaral 2006). Relativamente ao projecto TC-STAR, os melhores sistemas apresentam uma taxa de erro no reconhecimento da fala (Inglês e Espanhol) na ordem dos 10% (Bisani 2006) (Mostefa 2006).
Ainda neste projecto, em relação à qualidade da tradução da fala, os valores da última avaliação estão apresentados na tabela 2.1 (Mostefa 2006).
Um exemplo destes dois tipos de entrada para o sistema de tradução é ilustrado na tabela 2.2 em que **ASR** corresponde a texto directamente da saída do sistema de reconhecimento de fala e **Texto** corresponde a texto final editado, sem disfluências próprias da fala espontânea, incluindo pontuações e distinção entre maiúsculas e minúsculas.
Devido ao facto do sistema de tradução de fala ser composto por dois sistemas (Reconhe-

[2] *Technology and Corpora for Speech to Speech Translation*

Direcção	Entrada	BLEU	NIST	PER	WER
Inglês → Espanhol	ASR	38.7	8.73	35.8	49.8
Inglês → Espanhol	Texto	46.3	9.66	31.1	41.2
Espanhol → Inglês	ASR	41.5	9.12	32.3	46.6
Espanhol → Inglês	Texto	53.3	10.5	25.6	35.1
Chinês → Espanhol	ASR	15.0	5.61	59.7	80.0
Chinês → Espanhol	Texto	12.5	5.40	62.7	83.6

Tabela 2.1: Taxas de erro na tradução da fala do projecto TC-STAR

ASR: and i'm times and starting to know what frank sinatra must have felt like **Texto:** I am starting to know what Frank Sinatra must have felt like,

Tabela 2.2: Tipos de entrada para o sistema de tradução

cimento e Tradução), resultam alguns problemas que revelam alguma urgência em serem resolvidos, como por exemplo, a propagação de erros ocorridos na fase de reconhecimento da fala. Se a frase reconhecida não for a mais correcta, é muito provável que a frase traduzida também não seja. Abordagens a este problema vão desde englobar estes dois passos num só, tornando estas duas operações numa única operação atómica, até à utilização de múltiplas hipóteses do reconhecedor sob a forma de listas de frases ou grafos de palavras.

2.6 Conclusão

A tradução automática da fala é uma área que tem vindo a ser estudada e melhorada progressivamente através da utilização de diferentes abordagens nos últimos 50 anos. Na última década, grande parte destes avanços resultam da utilização de métodos estatísticos, sendo um dos grandes desafios actuais a introdução de informação linguística. Só muito recentemente foram obtidas melhorias através da introdução de informação sintáctica. Um outro desafio muito importante é a tradução de novas modalidades como a fala.

3 O sistema de tradução

Nesta secção apresenta-se uma descrição geral do sistema de tradução, bem como das ferramentas e corpora utilizadas.

3.1 Descrição geral do Sistema

De seguinda descrevem-se as etapas de criação do sistema de tradução usado nesta tese, nomeadamente as etapas de:

- Criação de modelo de linguagem;
- Normalização do corpus de treino;
- Treino do sistema;
- Filtragem da tabela de segmentos;
- Afinação de parâmetros;
- Avaliação.

3.1.1 Modelo de linguagem

O modelo de linguagem estatístico atribui uma probabilidade a uma sequência de palavras, através da distribuição de probabilidades. A utilização deste modelo é muito frequente em tarefas de reconhecimento e tradução de fala, onde este modelo tenta prever a próxima palavra numa sequência de palavras. Estimar a probabilidade de sequências de frases num corpora pode tornar-se uma tarefa bastante complicada. Por esta razão, estes são modelos de aproximação, utilizando o modelo *N-grams*. Quando uma frase é dada como entrada no tradutor são geradas várias traduções possíveis dessa mesma frase. A função posterior do modelo

de linguagem é atribuir um determinado peso a cada uma dessas traduções, escolhendo a que obteve maior valor. Supostamente, essa é uma frase que apresenta a maior probabilidade de se encontrar correcta, de acordo com a língua de destino.

3.1.2 Normalização do corpus de treino

O corpus de treino é constituído por um conjunto de frases escritas em duas línguas distintas, em que cada par de frases corresponde à tradução de uma na outra (equivalências de tradução). Este corpus, após normalizado, serve de treino ao sistema de tradução. Tarefas típicas desta etapa são transformação de abreviaturas para os seus nomes completos, bem como numeração romana, números decimais, datas, símbolos monetários, etc. Deste modo, é garantido que o vocabulário a utilizar no treino do sistema, é compatível com o vocabulário proveniente do reconhecedor de voz. A figura 3.1 demonstra o processo de normalização do corpus.

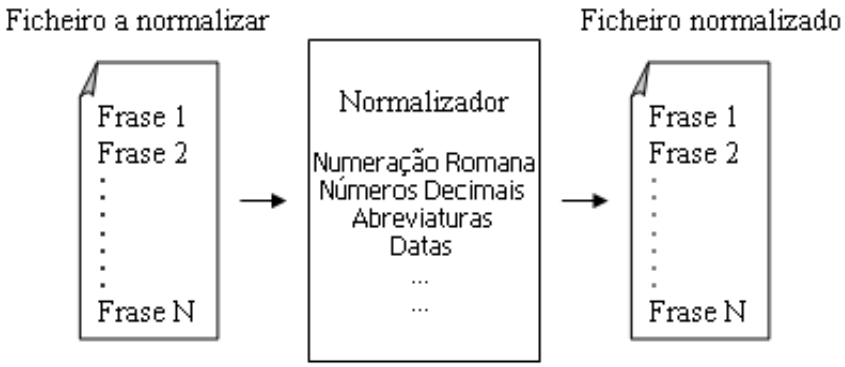

Figura 3.1: Normalização do corpus de teino

Exemplo:

- Frase não Normalizada - Exmo. dr. Rui Alves, venho por este meio informá-lo que dia 7 de Janeiro será promovido a eng.
- Frase Normalizada - excelentíssimo doutor rui alves venho por este meio informá-lo que dia sete de Janeiro será promovido a engenheiro

3.1.3 Treino do sistema

Após a construção de um modelo de linguagem e do corpus de treino a utilizar, o treino do sistema é efectuado, por exemplo, através da ferramenta Moses (ver Secção 3.2). Inicialmente são obtidos vários alinhamentos entre as palavras das duas línguas, sendo atribuída uma probabilidade a cada um desses alinhamentos, através da combinação das diferentes probabilidades obtidas pelas *features standard* que o sistema Moses contém. É gerada uma tabela de segmentos que contém todos esses alinhamentos e probabilidades entre palavras das duas línguas. A figura 3.2 ilustra o processo de treino de um sistema de tradução.

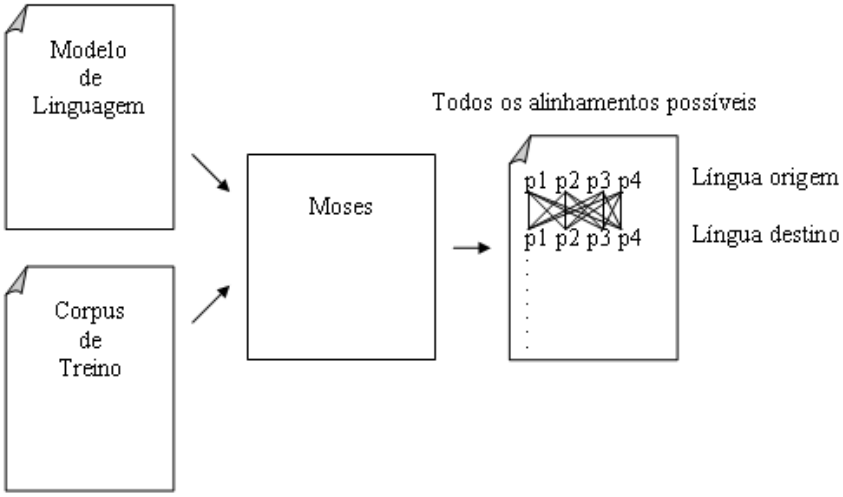

Figura 3.2: Treino do sistema

3.1.4 Filtragem da tabela de segmentos

O corpus de treino a utilizar necessita de conter uma grande quantidade de pares de frases (cerca de 1.000.000) e dado que tal não existe para o âmbito das notícias televisivas, foi tomada a decisão de continuar a utilizar-se o corpus baseado nas sessões do Parlamento Europeu. Uma vez que tal acontece e, de modo a que a tabela de segmentos faça parte integrante do domínio das notícias televisivas, é feita uma listagem das palavras utilizadas no vocabulário do reconhecedor. Essa mesma listagem é utilizada como filtro da tabela de segmentos, para que esta, no final, não contenha palavras desnecessárias, diminuindo substancialmente a carga computacional. A figura 3.3 ilustra o processo de filtragem da tabela de segmentos, gerada durante o

treino do sistema de tradução.

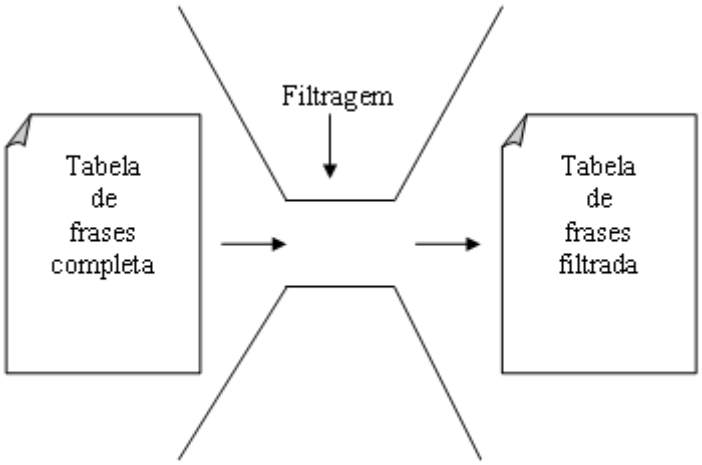

Figura 3.3: Filtragem da tabela de segmentos

3.1.5 A afinação de parâmetros (Tuning)

Neste passo é desempenhada (pelo sistema Moses) a tarefa de combinar diferentes valores para as *features* utilizadas pela ferramenta Moses, durante sucessivas iterações de tradução, até que o valor de BLEU, obtido até ao momento, estabilize e comece a convergir para um valor final. O vector final de pesos é então utilizado pela ferramenta Moses, altura em que passará a poder ser efectuada a tradução das frases pertencentes ao conjunto de teste. Esta é uma das etapas que mais tempo consome (para além da construção dos alinhamentos através da ferramenta Giza++ (ver secção 3.2)) e também uma das mais importantes de todo o processo de construção de um sistema de tradução eficaz. Neste passo, é indispensável utilizar um conjunto de desenvolvimento baseado no domínio em que estamos a trabalhar, nomeadamente as notícias televisivas. O treino do sistema com este conjunto permite que alguns parâmetros sejam afinados, melhorando os resultados de tradução, iteração após iteração, de modo a que, no final, quando o sistema for submetido a um corpus de teste, este apresente bons resultados de tradução e também um valor de BLEU (métrica descrita no capítulo 2) elevado.

3.1.6 Avaliação

Após o sistema ter sido construído, é necessário proceder à avaliação do seu desempenho. Deste modo, são utilizadas algumas métricas, por exemplo, a métrica BLEU. Quanto maior for o seu valor, melhor qualidade apresentam as frases obtidas pelo tradutor construído.

3.2 Ferramentas utilizadas

Nesta secção serão descritas as ferramentas utilizadas nas diferentes etapas da tradução, nomeadamente no modelo de linguagem e no treino do sistema.

3.2.1 Modelos de linguagem

O SRILM (Stolcke 2002) é uma ferramenta utilizada para criar e aplicar modelos de linguagem baseados em métodos estatísticos, principalmente para utilizar em reconhecimento de fala. O SRILM tem sido desenvolvido no laboratório *SRI Speech Technology and Research* desde 1995. A ferramenta alcançou destaque e maior notoriedade através do *WorkShop* de Verão em 1995, 1996 e 1997 realizadas na Universidade de *Johns Hopkins*.

SRILM é uma ferramenta que consiste nos seguintes componentes:

- Um conjunto de bibliotecas C++ que implementam modelos de linguagem;
- Um conjunto de programas executáveis de modo a possibilitar determinadas tarefas, tais como o treino de modelos de linguagem baseados em métodos estatisticos, bem como o seu teste e segmentação, etc;
- Um conjunto de vários scripts que facilitam as tarefas desta ferramenta.

O MKCLS (Och 1995) (Och 1999) é utilizada para realizar o treino de classes de palavras através da utilização do critério de máxima verosimilhança[1]. As classes de palavras resultantes são especialmente adaptadas para modelos de linguagem ou modelos de tradução baseados em métodos estatísticos.

[1]Método estatístico utilizado para calcular a melhor forma de adaptação de um modelo matemático a dados. A modelação de dados do mundo real por estimativa da máxima verossimilhança oferece um modo de estimar os parâmetros do modelo para que seja proporcionado um óptimo ajuste.

3.2.2 Treino do sistema

O Moses (Koehn 2007) é um sistema de tradução automático baseado em métodos estatísticos que permite de uma forma automática treinar modelos de tradução para qualquer par de línguas, sendo apenas necessário um conjunto de textos paralelos. A ferramenta Moses consiste num conjunto de componentes necessários para realizar o pré-processamento de informação, treino do modelo de linguagem e treino do modelo de tradução. Contém também ferramentas para a realização do *tuning* desses modelos (através da utilização de *minimum error rate training*) e para a avaliação das frases traduzidas (através da métrica BLEU). O GIZA++ (Och & Ney 2003) é um dos principais componentes do Moses. Esta é uma ferramenta utilizada para a realização do treino de modelos de tradução baseados em métodos estatisticos. GIZA++ é uma extensão do programa GIZA (parte da ferramenta *EGYPT*) que foi desenvolvido pela equipa de sistemas de tradução estatísticos durante o *WorkShop* de Verão em 1999 no Centro de Linguagem e Processamento de Fala na Universidade de *Johns-Hopkins*. GIZA++ inclui uma grande variedade de características adicionais. As extensões do GIZA++ foram desenvolvidas e escritas por *Franz Josef Och*. Mais informações acerca da ferramenta Moses podem ser consultadas em (Koehn 2007).

Características do GIZA++ que não estão presentes no GIZA:

- Implementa na totalidade o modelo de alinhamento IBM 4 descrito em (Brown 2003);
- Implementa o modelo de alinhamento IBM 5;
- Implementa o modelo de alinhamento *hidden Markov model* (HMM): Treino de *Baum-Welch*[2], algoritmo de *Forward-Backward*[3], palavra nula[4], dependência em classes de palavras;
- Implementação de uma variante do modelo de alinhamento IBM 3 e IBM 4 que permite o treino de -p0[5];

[2]Descobrir parâmetros desconhecidos do modelo HMM.

[3]Algoritmo de programação dinâmica para a computação da probabilidade de uma frase na língua destino. A sua complexidade temporal é de $O(2.T.N^T)$, onde T é o tamanho da frase e N o número de símbolos do alfabeto da língua em questão.

[4]Quando uma palavra não emparelha com nenhuma outra da língua destino, é emparelhada com a palavra *NULL*.

[5]Probabilidde de inserção da palavra NULL após uma palavra.

- Treino bastante mais eficiente para os modelos de fertilidade;
- Permite, de um modo mais fácil, adicionar novos parâmetros;
- O cálculo da perplexidade (ver Secção 4.4) para os modelos IBM 1, 2 e HMM foi melhorado.

3.3 Os Corpora utilizados

Uma das tarefas importantes a realizar para o sucesso deste trabalho foi sem dúvida o estudo do corpus a utilizar para o treino dos sistemas de tradução no âmbito das notícias televisivas. Relativamente à tradução automática no ambiente das sessões do Parlamento Europeu, não existem grandes problemas, uma vez que estas sessões são traduzidas para diversas línguas, existindo deste modo uma quantidade mais que suficiente de texto paralelo para treinar um sistema de tradução neste contexto. Como nota de observação, pode adiantar-se que existem cerca de 135.670 palavras em Português no corpus do Parlamento Europeu e cerca de 74.475 palavras em Inglês neste mesmo corpus. Estes corpus foi obtido no site (www.statmt.org/) e corresponde à versão 3, desenvolvida por Cameron Shaw Fordyce, Josh Schroeder e Philipp Koehn.

No entanto, este é um trabalho cujo objectivo passa por realizar um sistema de tradução baseado nas noticias televisivas. Como tal, foi necessário fazer um estudo acerca de corpora existente neste domínio. Após ter efectuado o reconhecimento de fala para texto de diversos telejornais (Meinedo 2008), chegou-se à conclusão de que o número de palavras distintas existentes nos noticiários portugueses é cerca de 57.565 e, no caso dos noticiários em Inglês, o número total de palavras é cerca de 34.828. Caso todas estas palavras estivessem contidas no corpus do Parlamento Europeu não haveria problema relativamente à sua tradução, uma vez que todo o corpus do Parlamente Europeu se encontra traduzido para diversas línguas. Mas tal não acontece. Após ter feito um estudo acerca do número de palavras portuguesas dos noticiários que não estão presentes no corpus do Parlamento Europeu, chegou-se à conclusão de que são cerca de 14.521 palavras. Analogamente, mas agora para o caso do Inglês, chegou-se à conclusão que cerca de 11.036 palavras dos noticiários em Inglês não estão presentes no corpus do Parlamento Europeu.

Este foi um dos principais problemas a serem resolvidos. A falta de corpora baseado no âmbito das notícias televisivas para treinar o sistema influenciará bastante a sua avaliação. Quanto mais corpora nesse domínio for utilizado no treino do sistema, mais o âmbito das notícias televisivas será abrangido, e consequentemente melhores resultados no final serão obtidos.

Para a resolução deste problema, inicialmente pensou-se em filtrar frases, tanto em Português como em Inglês, que contivessem as palavras em falta para que depois fossem traduzidas manualmente. O problema surgiu precisamente na tradução manual. Como não houve a possibilidade de se contratar um tradutor profissional, a solução passou por uma tentativa de tradução própria, de cerca de 500 frases. No final, o resultado não atingiu as espectativas esperadas, pois realizar esta tarefa por alguém sem qualificações técnicas tornou-se muito difícil. Dado que as notícias televisivas são emitidas numa linguagem específica, a tradução obtida não auferia as qualidades necessárias para resolver o problema da falta de corpus de treino no contexto das notícias. Chegou-se então à conclusão que este não seria o melhor método para tentar contornar o problema das palavras em falta no corpus do Parlamento Europeu.

A solução adoptada foi a de treinar o sistema com o corpus do Parlamento Europeu existente, sendo que quando fosse altura de afinar os parâmetros do sistema com um conjunto de desenvolvimento, seria utilizado um corpus exclusivamente baseado no contexto das notícias televisivas, corpus este que foi construído e traduzido na íntegra. Assim, uma vez que os recursos relativamente ao âmbito das notícias televisivas são poucos, foi obtido, através do *website* da *euronews* (http://www.euronews.net/), um conjunto de textos paralelos em Português e Inglês, de modo a criar dois documentos que contivessem uma quantidade suficiente de frases para serem utilizados como corpus de desenvolvimento e avaliação do sistema de tradução. Foi construído um conjunto de 914 frases, em que 457 frases estão escritas em Português e, a outra metade, corresponde à tradução dessas mesmas frases. Na tabela 3.1 podem ver-se exemplos das frases retiradas do *website* da *euronews* e a respectiva tradução.

Este é um conjunto de frases comparável mas que não corresponde a traduções exactas.

É um aspecto que será analisado na Seccção 4.3, onde serão apresentados os respectivos problemas e soluções adoptadas para os mesmos.

Frase em Português	Frase em Inglês
Logo depois do seu governo ter recebido um voto de confiança no Parlamento	Just after your government received a vote of confidence in parliament
Como é que responde a isto?	How do you respond to that?
O que quis dizer com isso?	Do you think that's fair?

Tabela 3.1: Exemplo de pares de frases retiradas do *website* da *Euronews*.

Relativamente ao corpus utilizado para a construção do modelo de linguagem, tal como acontece com o corpora de desenvolvimento e avaliação, este foi construído com base em textos pertecentes ao âmbito das notícias televisivas. Contudo, na Secção 4.4 é explicado qual o modelo de linguagem final utilizado no sistema de tradução construído, nomeadamente um modelo de linguagem interpolado entre dois modelos distintos, um relativo a notícias televisivas (MacIntyre 1998) e um outro relativo a notícias de jornais (Graff 1995).

4 Tradução de Texto e Avaliação

4.1 Introdução

Como já havia sido referido, um sistema de tradução automático de fala é composto por dois sistemas, um de reconhecimento de fala para texto e um outro de tradução de texto para texto. Neste capítulo, serão descritas as diversas experiências efectuadas para a criação de um tradutor de texto para texto para o par de línguas Português - Inglês e para o contexto das notícias televisivas, bem como os resultados obtidos, problemas encontrados e formas de os resolver.

4.1.1 Parlamento Europeu

Uma vez que não existia material suficiente para se obter um tradutor baseado única e exclusivamente nas notícias televisivas, começou-se inicialmente por desenvolver um tradutor com âmbito no Parlamento Europeu. Esta experiência serviu para obter referências relativamente aos resultados obtidos em outros sistemas no âmbito do Parlamento Europeu e, em paralelo, permitiu um primeiro contacto com as ferramentas a utilizar para a criação de um sistema de tradução automático baseado em métodos estatísticos.

O primeiro tradutor desenvolvido para o par de línguas Português - Inglês foi então baseado exclusivamente nas sessões do Parlamento Europeu. O modelo de linguagem foi criado através de textos exclusivos do Parlamento Europeu, assim como o corpus de treino utilizado para o treino do sistema de tradução. Para a realização do *Tuning* do sistema, foram também utilizados textos paralelos baseados no Parlamento Europeu. Relativamente ao

conjunto de teste utilizado, este foi igualmente baseado nas sessões do Parlamento Europeu. Quanto à avaliação do sistema de tradução, foi utilizada a métrica BLEU, tendo o sistema obtido um valor de BLEU de 0,3531 com a saída gerada, entrada e texto de referência, ambos *tokenizados* (cada elemento da frase constitui um *token*) e com todas as palavras escritas em letras minúsculas. A este conjunto de condições será atribuído a designação de condição A.

Para obter um valor de BLEU dentro das mesmas condições que o sistema *Europarl*, foi necessário efectuar o *recase*[1] (sistema que automaticamente passa de minúsculas para maiúsculas a primeira palavra da frase, nomes próprios, etc) da entrada, saída e referência, bem como a *detokenização* (inverso do processo de *tokenização*). Nestas condições, obteve-se um valor de *Bleu* de 0,3445. A este conjunto de condições será atribuído a designação de condição B.
Na tabela 4.1 é ilustrado o sistema acima descrito, bem como os seus valores de BLEU.

	BLEU
Condição A	0,3531
Condição B	0,3445

Tabela 4.1: Sistema de tradução, baseado exclusivamente nas sessões do Parlamento Europeu e respectivos valores de BLEU.

4.1.2 Parlamento Europeu Filtrado

Numa segunda experência no quadro da tradução do Parlamento Europeu, a *phrase-table* gerada a partir do treino do sistema foi filtrada, de modo a conter apenas as palavras que se encontrem presentes na lista de palavras das notícias televisivas (esta é uma lista que contém um conjunto de palavras que abrange quase todo o vocabulário utilizado nas notícias televisivas, com cerca de 57.565 palavras). Devido à filtragem da tabela de segmentos, muitas das palavras utilizadas no parlamento foram suprimidas e, uma vez que o corpus de teste utilizado para avaliar o sistema segundo a métrica de BLEU foi baseado no parlamento, então os resultados obtidos, como seriam de esperar, baixaram um pouco, relativamente ao sistema

[1]Ferramenta treinada geralmente com o corpus utilizado no treino do modelo de linguagem. O sistema converte todas as palavras do corpus para minúsculas e efectua alinhamentos, atribuindo as respectivas probabilidades. Desta forma, este sistema aprende quais as palavras cuja primeira letra deve ser convertida em maiúscula e converte também, por omissão, a primeira letra de cada frase em maiúscula.

baseado por completo nas sessões do parlamento. O sistema obteve assim um valor de Bleu de 0,2699 com a saída gerada, entrada e referência, ambos *tokenizados* e com todas as palavras escritas em letras minúsculas. Após efectuar o *recase* e a *detokenização*, o valor obtido foi de 0,2643. Na tabela 4.2 é ilustrado o sistema criado, a sua descrição e valores de BLEU obtidos.

	BLEU
Condição A	0,2699
Condição B	0,2643

Tabela 4.2: Sistema de tradução, baseado nas sessões do Parlamento Europeu mas já com a tabela de segmentos filtrada apenas com as palavras pertencentes aos corpus de desenvolvimento.

4.2 Baseline

Apesar de ter sido desenvolvido um sistema de tradução de Português para Inglês que apresenta valores de BLEU relativamente semelhantes aos obtidos por outros sistemas, este sistema continua ainda a pertencer ao domínio das sessões do Parlamento Europeu. É então necessário fazer a sua adaptação para o domínio das notícias televisivas, abrangendo várias tarefas a realizar, que serão mencionadas em seguida.

Em primeiro lugar é necessário adaptar o vocabulário, de modo a que este seja compatível com o reconhecedor, ou seja, foi necessário realizar a normalização dos textos paralelos utilizados no treino do tradutor, bem como a sua *tokenização* e elaboração da passagem de todas as palavras de maiúsculas para minúsculas.

É igualmente necessário filtrar a tabela de segmentos gerada no treino do modelo, para que apenas sejam contemplados os segmentos cujas palavras se encontrem na lista de palavras utilizadas no corpus de desenvolvimento a utilizar. A tabela de segmentos é uma tabela que contém sequências de palavras escritas numa dada língua e diversas traduções possíveis para cada sequência dessas palavras, sendo atribuída a cada tradução uma certa probabilidade, de acordo com os valores das diferentes *features standard* que o sistema Moses contém. Como exemplo de algumas *features* podem ser destacadas o modelo de linguagem (atribui um deter-

minado valor à frase de acordo com a sua fluência na língua destino) e modelo de distorção (atribui um determinado valor à frase relativamente à quantidade de reordenamentos feitos). Muitas das vezes é frequente existirem várias traduções iguais para o mesmo segmento com *scores* diferentes, uma vez que essas traduções podem ter sido obtidas de formas diferentes, como por exemplo, podem ter sido efectuados reordenamentos diferentes ou podem ter sido efectuadas traduções de diferentes segmentos e, no final, o segmento gerado ter sido igual a um outro.

Foi realizada uma nova experiência, mas desta vez com mais alguma adaptação ao domínio pretendido. Neste caso, o corpus de treino do modelo continuou a ser o do Parlamento Europeu (devido à falta de recursos), mas relativamente ao modelo de linguagem, corpus de desenvolvimento e teste, foram já utilizados corpus baseados nas sessões das notícias televisivas. Tanto o corpus de desenvolvimento como o de teste foram extraídos, na íntegra, do *website* da *euronews*. Relativamente à tabela de segmentos, esta foi também adaptada ao âmbito das notícias televisivas, tendo sido filtrada, de modo a conter apenas segmentos construídos com palavras pertencentes ao corpus de desenvolvimento. No final, foi obtido um valor de BLEU de 0,1705 com a saída gerada, entrada e referência, ambos tokenizados e com todas as palavras escritas em letas minúsculas. Após efectuar o *recase* e a *detokenização*, o valor obtido foi de 0,1650. A tabela 4.3 descreve o sistema criado, a sua descrição e valores de BLEU obtidos.

	BLEU
Condição A	0,1705
Condição B	0,1650

Tabela 4.3: Sistema de tradução, com a tabela de segmentos filtrada, corpus de desenvolvimento, teste e modelo de linguagem baseados nas notícias televisivas e corpus de treino baseado nas sessões do Parlamento Europeu.

4.3 Experiência 1

Após realizar a experiência descrita anteriormente e, uma vez que os resultados de BLEU não foram os desejados, chegou-se à conclusão de que os corpora que estavam a ser utilizados para *Tuning* e teste poderiam não estar consistentes, uma vez que as frases existentes são comparáveis, mas não necessariamente traduções directas. Na tabela 4.4 ilustra um exemplo

desta natureza.

Frase em PT: Ennio Morricone vai ser galardoado com um Óscar pelo conjunto da sua carreira. **Frase em EN:** Ennio Morricone is on his way to LA to receive an honorary Oscar.

Tabela 4.4: Frases comparáveis, mas não necessariamente traduções directas

Foi então tomada a decisão de se proceder à alteração deste corpus mas apenas nas frases escritas em Português. Deste modo, as frases de referência em inglês mantêm-se a salvo de alterações com base no conhecimento já existente sobre o sistema de tradução. Em todas as alterações efectuadas foi sempre mantida a ordem pelas quais as palavras se apresentavam na frase, bem como todas as abreviaturas, procurando ao máximo efectuar o menor número de alterações necessárias. Deste modo, foram apenas realizados ajustes das frases Portuguesas relativamente às Inglesas, como por exemplo a introdução de palavras que se encontravam nas frases em Inglês, mas que não estavam presentes nas frases em Português, ou a eliminação de palavras que se encontravam nas frases em Português mas que não estavam presentes nas frases da língua Inglesa. O corpus final passou então a ser mais coerente, tendo cada frase a respectiva tradução correcta.

A tabela 4.5 ilustra alguns exemplos das alterações efectuadas ao corpus de desenvolvimento e teste.

O primeiro exemplo demonstra um caso em que existem conceitos que não estão presentes na frase em Português ('la' e 'is on his way') e também outros que estão a mais ('conjunto da sua carreira'). Foi então decidido remover estas palavras a mais e introduzir a palavra 'la' bem como a tradução de 'is on his way'.
No segundo exemplo é demonstrado um caso em que a palavra em Português não tem o mesmo significado que a palavra em Inglês. Neste caso a palavra 'isso' foi substituída pela palavra 'desqualificar'.
No terceiro exemplo é ilustrado um caso em que as siglas que se encontravam no corpus, tal como foi retirado do *website* da *euronews*, foram mantidas na íntegra. Adicionalmente, em semelhança com o exemplo anterior, foi efectuada uma substituição da palavra 'coisas' pela

Frase antiga em PT: Ennio Morricone vai ser galardoado com um Óscar pelo conjunto da sua carreira. **Frase nova em PT:** Ennio Morricone **está a ir para la** para receber um Óscar honorário. **Frase em EN:** Ennio Morricone is on his way to LA to receive an honorary Oscar. **Frase antiga em PT:** O que quis dizer com isso? **Frase nova em PT:** O que quer dizer com **desqualificar**? **Frase em EN:** What do you mean by disqualify? **Frase antiga em PT:** Para a Europa, o acordo entre a Euronext e a NYSE traz um certo número de coisas. **Frase nova em PT:** Para a Europa, o acordo entre a Euronext e a NYSE traz um certo número de **benefícios**. **Frase em EN:** The agreement between the New York Stock Exchange and Euronext has a number of benefits for Europe.

Tabela 4.5: Alteração do corpus de desenvolvimento e teste

palavra 'benefícios'.

Após terem sido realizadas as alterações acima descritas, foi construído um sistema de tradução adaptado o mais possível ao domínio das notícias televisivas, sendo que, inevitavelmente, o corpus de treino foi baseado na íntegra nas sessões do Parlamento Europeu, mas tanto o corpus utilizado para a construção do modelo de linguagem, como o corpus utilizado para *Tuning* (desenvolvimento) e corpus utilizado para teste, foram exclusivamente baseados no âmbito das notícias. Após o treino do sistema e a criação da respectiva tabela de frases, esta foi filtrada apenas com o conjunto de palavras que se encontram no corpus de desenvolvimento, de modo a ser optimizada e adaptada ao vocabulário das notícias. No final, foi obtido um valor de BLEU de 0,4776, com a saída gerada, entrada e referência, ambos tokenizados e com todas as palavras escritas em letras minúsculas. Após efectuar o *recase* e a *detokenização*, o valor obtido foi de 0,4722.

A tabela 4.6 descreve o sistema criado, a sua descrição e valores de BLEU obtidos.

	BLEU
Condição A	0,4776
Condição B	0,4722

Tabela 4.6: Sistema de tradução, com tabela de segmentos filtrada, corpus do modelo de linguagem baseado nas notícias televisivas, corpus de *Tuning* (desenvolvimento) e teste alterados e, corpus de treino baseado nas sessões do Parlamento Europeu.

4.4 Experiência 2

Como já havia sido referido nas secções anteriores, o modelo de linguagem é indispensável nos sistemas de tradução baseados em métodos estatísticos. Para que sejam correctamente construídos, é necessário uma grande quantidade de textos, de preferência pertencentes ao domínio da tarefa. Neste caso, não existe material suficiente para a construção de um modelo de linguagem baseado exclusivamente nas notícias televisivas. Uma vez que também se encontra fora de hipótese a utilização de corpus baseado exclusivamente em jornais (por não pertencer ao domínio da tarefa), foi tomada a decisão de se construirem dois modelos de linguagem, um baseado nas notícias televisivas e outro baseado em textos de jornais, para que depois se proceda à sua interpolação, resultando um modelo mais rico baseado nas notícias e, ao mesmo tempo, que contém expressões utilizadas frequentemente nos jornais que possam também ser utilizadas por vezes nas notícias televisivas.

Relativamente à criação do modelo de linguagem das transcrições televisivas, como se pode observar na tabela 4.7, foram construídos 8 modelos de linguagens com diferentes *cut-off* (a contagem do *n-gram* apenas é contabilizada caso este ocorra pelo menos um determinado número de vezes). Os números 1, 2, 3 e 4 representam a dimensão do *n-gram* e N representa a dimensão total do modelo. Para cada modelo de linguagem gerado, foi determinada a sua perplexidade (PPL) para o conjunto de desenvolvimento a utilizar. Quando se cria um modelo de linguagem é importante determinar a sua qualidade. A melhor forma para o fazer é utilizar a métrica WER. No entanto, este método não é muito eficiente, uma vez que necessita de uma grande quantidade de processamento de informação, tornado isto num processo muito demorado. É então necessário utilizar uma outra alternativa, como por exemplo a utilização do valor de perplexidade. A perplexidade é uma medida utilizada muitas vezes para medir

a qualidade de um determinado modelo de linguagem, uma vez que testa a capacidade de um modelo para prever um texto desconhecido (texto não utilizado no corpora de treino). A perplexidade de um modelo de linguagem relativamente a um texto de n palavras é definida pela fórmula 4.1.

$$PPL = 2^{LP}, ondeLP = (\frac{1}{n})logP(W_1...W_n) \tag{4.1}$$

P é a probabilidade de uma sequência de n palavras dada pelo modelo de linguagem. Quanto maior for o tamanho do modelo, menor será o valor de perplexidade no conjunto de treino. Esta relação no entanto sofre uma alteração brusca, na medida em que não vale a pena utilizar um valor de perplexidade mais baixo caso o tamanho do modelo sofra um grande aumento, pois isto terá impacto na capacidade de processamento. Relativamente à criação do modelo de

Cut-Off	1	2	3	4	N	PPL
2-4-4	51.023	687.913	471.656	471.656	1.590.597	159,621
2-3-4	51.023	687.913	700.717	380.005	1.819.658	156,672
2-3-3	51.023	692.135	732.371	617.183	2.092.712	155,928
2-2-2	51.023	723.343	1.493.769	1.425.671	3.693.806	154,64
1-2-2	51.023	1.855.266	1.493.769	1.425.671	4.825.719	**151,453**
1-1-3	51.023	1.855.266	6.936.411	617.183	9.459.883	151,29
1-1-2	51.023	1.855.266	6.936.411	1.425.671	10.268.371	151,814
1-1-1	51.023	1.855.266	6.936.411	11.428.928	20.271.628	153,182

Tabela 4.7: Perplexidade dos modelos de linguagem das notícias televisivas, para diferentes *cut-off*.

linguagem para os jornais, como é visível através da tabela 4.8, foram construídos 5 modelos de linguagem com diferentes *cut-off*. Tal como no caso anterior, foi também determinada a perplexidade de cada um dos modelos de linguagem.

Cut-Off	1	2	3	4	N	PPL
50-50-50	51.023	558.970	695.679	339.642	1.645.314	157,23
10-10-10	51.023	1.846.059	3.806.538	2.819.012	8.522.632	137,15
5-10-10	51.023	3.148.956	3.806.538	2.819.012	9.825.529	134,15
5-5-10	51.023	3.148.956	8.443.910	2.819.012	14.462.901	**132,607**
5-5-5	51.023	3.148.956	8.443.910	8.221.924	19.865.813	132,948

Tabela 4.8: Perplexidade dos modelos de linguagem dos jornais, para diferentes *cut-off*.

Tanto na tabela 4.7 como na 4.8 se verifica que à medida que os *cut-off* vão sendo mais pe-

quenos, o tamanho total do modelo de linguagem e o valor de perplexidade aumenta, até um certo ponto que começa a convergir e, até mesmo, a regredir. Esta é a altura em que os *cut-off* são demasiado rigorosos e o modelo de linguagem começa a ficar demasiadamente adaptado ao conjunto de treino, o qual está a ser utilizado para se obterem as perplexidades. Desta forma, relativamente ao modelo de linguagem das notícias televisivas, foi escolhido o modelo com *cut-off* de 1-2-2, uma vez que os modelos com *cut-off* de 1-1-3, 1-1-2 e 1-1-1 não apresentam uma melhoria no valor de perplexidade que justifique o grande aumento da dimensão final do modelo de linguagem. Relativamente ao modelo de linguagem dos jornais, foi escolhido o que apresenta *cut-off* de 5-5-10, pois é aquele que apresenta melhor valor de perplexidade tendo em conta a dimensão final do modelo de linguagem. Após terem sido escolhidos os dois modelos a interpolar, é calculado o valor de *lambda* (através de um script pertencente à ferramenta SRILM) a ser utilizado no processo de interpolação. O valor de *lambda* obtido foi de 0,497919 e o modelo de linguagem interpolado e respectiva perplexidade encontra-se descrito na tabela 4.9. Este valor de *lambda* minimiza a perplexidade do conjunto de desenvolvimento. Foi uma tarefa realizada com sucesso, pois além do valor de perplexidade do modelo de linguagem interpolado ter sido mais baixo relativamente aos dois modelos utilizados, o novo modelo passou a conter não só expressões utilizadas nas notícias televisivas, como também de notícias de jornal. A figura 4.1 ilustra o processo de interpolação descrito e utilizado para a criação do modelo de linguagem final. A interpolação é dada pela fórmula 4.2, em que P corresponde à interpolação dos modelos P_1 e P_2.

$$P(s) = \lambda P_1(s) + (1 - \lambda)P_2(s) \tag{4.2}$$

Cut-Off	1	2	3	4	N	PPL
1-2-2 — 5-5-10	51.023	3.863.377	8.822.249	3.671.319	16.407.968	**112,894**

Tabela 4.9: Perplexidade do Modelo de Linguagem Interpolado.

4.5 Experiência 3

Após a construção do modelo de linguagem interpolado, este foi introduzido no sistema de tradução anterior, substituindo o existente. Foi novamente realizado o *Tuning* do sistema com o mesmo corpus de desenvolvimento obtendo uma nova gama de pesos a utilizar pela fer-

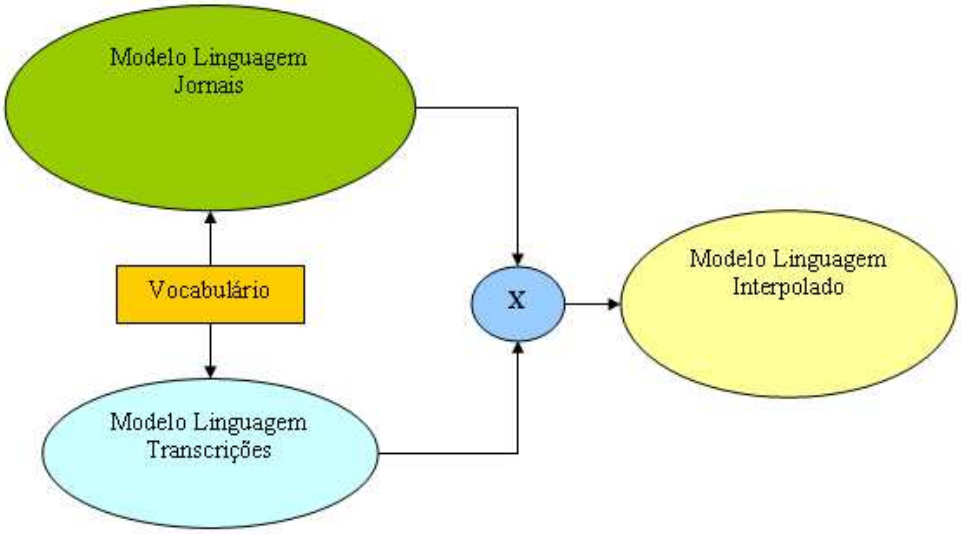

Figura 4.1: Interpolação dos Modelos de Linguagem

ramenta Moses. Ao ser utilizado o mesmo corpus de teste para obter uma nova avaliação do sistema com este modelo de linguagem, foi obtido um valor de BLEU de 0,4861 com a saída gerada, entrada e referência, ambos tokenizados e com todas as palavras escritas em letas minúsculas. Após efectuar o *recase* e a *detokenização*, o valor obtido foi de 0,4799. A tabela 4.10 descreve o sistema criado, a sua descrição e valores de BLEU obtidos e, na tabela 4.11 ilustram-se alguns exemplos de traduções obtidas por este sistema de tradução, a entrada utilizada e a respectiva frase de referência.

	BLEU
Condição A	0,4861
Condição B	0,4799

Tabela 4.10: Sistema de tradução, com tabela de segmentos filtrada, modelo de linguagem interpolado, corpus de *Tuning* (desenvolvimento) e teste alterados e, corpus de treino baseado nas sessões do Parlamento Europeu.

A utilização de um modelo de linguagem interpolado a partir de dois modelos distintos (Jornais e Notícias) mas que se complementam foi benéfica, uma vez que permitiu uma melhoria do valor de BLEU de 0,4776 para 0,4861 (no caso da entrada, saída gerada e referência, ambos tokenizados e com todas as palavras escritas em letras minúsculas) e de 0,4722 para 0,4799

Frase de entrada: não acho que devemos estar preocupados com o que o ivanov lavrov ou outros líderes militares russos dizem **Frase obtida:** i do not think that we should be concerned with what ivanov lavrov or other russian military leaders say . **Frase de referência:** i do not think we should be concerned with what ivanov lavrov or other russian military leaders say .
Frase de entrada: o massacre de oito mil muçulmanos em julho de mil novecentos e noventa e cinco depois da enclave sob protecção do governo foi um genocídio **Frase obtida:** the massacre of eight thousand muslims in july 1995 after the enclave under protection of the government was genocide . **Frase de referência:** the massacre of 8000 muslims in july 1995 , after the enclave under the protection of the govern , was genocide .
Frase de entrada: por vezes o tribunal tem sido criticado pelo tempo que demora a resolver um caso **Frase obtida:** sometimes the court has been criticized by the time it takes to solve a case . **Frase de referência:** sometimes the court has been criticised by the time it takes to solve a case .

Tabela 4.11: Traduções obtidas e respectivas frases de entrada e referência

(após efectuar o *recase* e a *detokenização*). Com este trabalho efectuado e juntando as alterações efectuadas ao corpus de desenvolvimento e teste, comprova-se a necessidade de se utilizarem modelos de linguagem baseados no domínio da tarefa em questão (preferêncialmente interpolados, caso não exista material de texto suficiente, como é o caso das notícias televisivas) e também conjuntos de texto paralelo que estejam coerentes, uma vez que um grande número de pequenas incoerências, levam a resultados muito maus, tanto na qualidade das frases obtidas, como no valor de BLEU.

4.6 Experiência 4

Todas as experiências realizadas demostraram uma perca no valor de BLEU (ainda que relativamente pequena) ao efectuar o *recase* das frases obtidas pelo *decoder*. Desta forma, decidiu-se refinar o modo como esta ferramenta está a ser treinada, procurando minimizar ao máximo a perca do valor de BLEU. Juntaram-se dois corpora de diferentes domínios, um relativo a jornais e, um outro relativo a transcrições de notícias televisivas. Deste modo, a ferramenta passou a considerar mais informação, no momento da decisão de efectuar a passagem de letras

minúsculas para minúsculas.

O texto adicionado (texto de jornais) é semelhante às transcrições das notícias televisivas mas, uma vez que continua a pertencer a um domínio diferente, os resultados obtidos pela ferramenta não melhoraram, antes pelo contrário, verificou-se um descida do valor de BLEU de 0.4799 para 0.4790. Com esta experiência pode então concluir-se que este tipo de tarefa apenas pode ser melhorada caso o texto a adicionar faça parte do domínio da tarefa, caso contrário, o resultado pode não melhorar (e até piorar como foi o caso), passando a utilizar-se mais informação linguística, o que torna a tarefa computacionalmente mais pesada e sem resultados positivos.

Uma vez que se verifica sempre uma perca no valor de BLEU ao realizar o *recase* das frases obtidas pelo *decoder*, mesmo quando este é baseado única e exclusivamente em corpus do domínio das notícias televisivas, realizou-se uma outra experiência afim de tentar encontrar alguma solução que apresentasse resultados positivos. O sistema de tradução foi então treinado com um conjunto de treino, em que todas as palavras das frases em Inglês não se encontram escritas em letras minúsculas. Desta forma, as frases traduzidas pelo *decoder* são já constituídas por palavras escritas em minúsculas e também em maiúsculas. Foi realizada esta experiência com os mesmos corpora utilizados no sistema anterior. No final, o resultado de BLEU não foi satisfatório. Verificou-se que algumas palavras eram bem traduzidas para maiúsculas, mas outras não, mantendo-se em letras minúsculas. Este resultado justifica-se pelo facto do corpus paralelo de treino não se encontrar totalmente coerente relativamente à capitalização das palavras. O valor de BLEU obtido pelo sistema de tradução que aprende como capitalizar as palavras está descrito na tabela 4.12.

BLEU
0,4071

Tabela 4.12: Sistema de tradução com recaser automático.

4.7 Experiência 5

Com o objectivo de melhorar ainda mais o valor de BLEU obtido, nesta secção é apresentado um último passo, em que um conjunto de 1.000 hipóteses geradas pelo melhor sistema de tradução descrito na secção anterior é reavaliado com algumas *features* novas.

As novas *features* utilizadas foram as seguintes:

- Diferença do número de palavras entre a frase em Português e a frase em Inglês
- POS (*Part Of Speech*) - Utilização de regras de correspondência entre o par de línguas Português e Inglês e de padrões de penalização na língua Inglesa.

Estas *features* foram combinadas com os *scores* obtidos numa primeira passagem de acordo com um modelo logarítmico. Foram feitas diversas combinações de pesos de modo a maximar o valor de BLEU, através do algoritmo POWELL (Powell 1998).

4.7.1 *Feature* de contagem do número de palavras

Relativamente à *feature* da diferença do número de palavras, foram feitas várias experiências, utilizando a combinação entre número de palavras da frase em Português, número de palavras da frase em inglês e diferença do número de palavras. Na tabela 4.13 estão descritas as várias experiências e o respectivo valor de BLEU obido.

Sistema	BLEU
Sistema Final	0,4861
Contagem de palavras em Inglês	0,5053
Contagem de palavras em Português	0,4927
Diferença do número de palavras	**0,5055**
Contagem de palavras em Português e Inglês	0,5053
Contagem de palavras em Inglês e diferença de palavras	0,5019
Contagem de palavras em Português e diferença de palavras	0,5054
Contagem de palavras em Português e Inglês e diferença do número de palavras	0,5019

Tabela 4.13: Contribuição da *feature* de contagem do número de palavras no processo de *rescoring*.

Através da análise da tabela 4.13 pode concluir-se que todas as combinações desta *feature* contribuem para um aumento do valor de BLEU. Contudo, é a diferença do número de palavras aquela que mais contribui.

4.7.2 *Feature Part-Of-Speech* (POS)

Relativamente a esta *feature* foram introduzidos dois conceitos:

f1 - Computação de semelhanças entre as etiquetas de POS, assumindo que o número de entidades (tais como nomes próprios), deve ser estável tanto na frase em Português como na sua tradução. Para cada frase em Português e Inglês, é feita uma contagem de determinadas etiquetas. Desta forma, a feature *f1* é calculada através da fórmula 4.3, onde pt_i representa as contagens da etiqueta número i em Português e en_i as contagens das etiquetas correspondentes em Inglês.

$$f1 = \sqrt[2]{\sum_{n=1}^{\#pos} (pt_n - en_n)^2} \tag{4.3}$$

A tabela 4.14 demonstra as equivalências entre as etiquetas em Português e Inglês.

Português	Inglês
Ncms Ncfs	NN
Ncmp Ncfp	NNS
V=ip3s V=ip3p V=f=1s	VBZ VVP VB
A=pfs	JJ
A=pms	JJ
Pd=ms	DT
S=s	DT

Tabela 4.14: Equivalência de etiquetas de POS entre Português e Inglês

Nos dois conceitos inerentes a esta *feature*, as frases geradas (*n-best*) e as frases de entrada são analisadas pelo *treetagger* [2] que as transforma numa sequência de etiquetas.

f2 - Computação de padrões de penalização, assumindo que determinadas sequências de etiquetas (padrões) são muito invulgares (tais como DT DT, onde DT é um determinante). De modo a calcular *f2*, os padrões de penalização para Inglês considerados estão na tabela 4.15.

[2] http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger

Padrão Inglês
DT DT
VV VV
IN IN
DT NN JJ

Tabela 4.15: Padrões de penalização para Inglês

Depois dos padrões de penalização estarem definidos, estes são procurados em cada frase gerada. Sempre que for encontrado um padrão numa determinada frase, é atribuída uma penalização. *f2* é a soma das penalizações.
Esta é uma *feature* que merece um estudo mais aprofundado, uma vez que permite a introdução de determinadas regras linguísticas nas frases geradas, podendo levar a que se obtenham frases mais bem estruturadas, de acordo com a língua de destino.

A tabela 4.16 contém o resultado da utilização desta feature e o respectivo BLEU obtido.

Sistema	BLEU
Sistema Final	0,4861
POS	**0,4967**

Tabela 4.16: Contribuição da *feature* POS.

A tabela 4.16 demonstra que a utilização destas *features* (conceitos *f1 e f2*) contribuiu para a maximização do valor de BLEU.

No final feito realizado um último rescoring com a utilização das duas *features* que mais contribuíram para o aumento do valor do BLEU, mais propriamente a *feature* da diferença do número de palavras e a de POS.

O valor de BLEU final está documentado na tabela 4.17. Este valor sofreu um aumento após o rescoring ter contabilizado estas duas *features* em simultâneo.

Sistema	BLEU
Sistema Final	0,4861
Diferença do número de palavras e POS	**0,5088**

Tabela 4.17: Contribuição das *features* de diferença do número de palavras e POS.

4.8 Ponto de Comparação

De modo a obter uma comparação com outros sistemas de tradução, foi tomada a decisão de traduzir todas as frases do corpus de teste utilizado para avaliar o sistema descrito, através do sistema de tradução fornecido pelo motor de busca *google*. Desta forma, foram então introduzidas todas as frases em Português neste sistema e foram obtidas as respectivas traduções. No final, foi registado um valor de BLEU de 0,4102. No caso da tradução obtida pelo *google*, o próprio sistema efectua o *recase* e a *detokenização*, pelo que apenas é apresentado um valor de BLEU. A tabela 4.18 descreve o sistema utilizado para efectuar a tradução e os respectivos valores de BLEU obtidos. A tabela 4.19 demonstra algumas traduções obtidas pelo *google* e outras pelo sistema de tradução desenvolvido.

Sistema	BLEU
Google	0,4102

Tabela 4.18: Sistema de tradução do google e respectivos valores de BLEU obtidos com o mesmo corpus de desenvolvimento.

Com todo este trabalho e esforço, foi possível obter um sistema de tradução no contexto das notícias televisivas que fornece traduções com uma melhor qualidade relativamente ao sistema de tradução fornecido pelo motor de busca do *google*, tal como indicado pela métrica de BLEU.

4.9 Resumo

De um modo resumido, inicialmente foi construído um sistema de tradução baseado exclusivamente nas sessões do Parlamento Europeu. Foi aplicado o princípio de dividir para conquistar, tendo sido introduzidas, sequencialmente, várias adaptações para que o sistema fosse capaz

Entrada	*Google*	Sistema Desenvolvido
em janeiro a alemanha assume a presidência da união	germany in january to assume the presidency of the union	**in january** , germany assumes the presidency of the union
por outro lado é preciso tomar em conta a ausência de liderança europeia nesta área	secondly we must take into account the lack of european leadership in this area	**on the other hand** , we must take into account the absence of european leadership in this area
mas esta é a história e esta é a razão pela qual pretendo modificar	but this is history and this is why i want change	but this is the history and **this is the reason** why i want to change

Tabela 4.19: Traduções obtidas pelo *Google* Vs Sistema de tradução.

de modelar o domínio das notícias televisivas. Começou-se por filtrar a tabela de sequências de modo a esta conter apenas sequências de palavras existentes no domínio das notícias televisivas. Foi contruído um modelo de linguagem através da interpolação de dois modelos, nomeadamente um de jornais e um outro de notícias televisivas. O conjunto de *Tuning* (desenvolvimento) foi refinado, de modo a conter traduções coerentes e correctas e, por fim, foi utilizada uma lista das *n-best* traduções de cada frase, tendo sido introduzidas novas *features*, de modo a que a escolha da melhor frase pudesse ser efectuada ainda com mais informação.

5 Conclusões e Trabalho Futuro

Relativamente a trabalhos futuros que possam melhorar a qualidade de sistemas de tradução deste género, algumas abordagens sobre as OOVWs (out of vocabulary words - palavras fora do vocabulário) podem ser desenvolvidas, de modo a contornar e ao mesmo tempo resolver este problema. Pode utilizar-se um dicionário onde são inseridas todas as palavras e respectivas traduções. Este método pode ser eficaz e rápido de desenvolver, mas apenas no caso em que este número de palavras não se revele muito elevado. Neste caso, cerca de 14521 palavras em Português não se encontram presentes no corpus do Parlamento Europeu e o mesmo se passa para cerca de 11036 palavras em Inglês, sendo por isso impensável transcrever todas estas palavras e respectivas traduções. Uma solução possível pode passar pela escrita dos verbos contidos no corpus de treino do sistema em todos os tempos verbais, bem como a sua tradução. Para isso, pode ser utilizado como auxílio o *website* (http://www.verbix.com), onde é possível obter todas as formas verbais de qualquer verbo. Uma outra opção para resolver o problema das palavras fora do vocabulário passa apenas por algumas palavras serem copiadas para a saída, no caso das palavras que se encontrem simultaneamente no corpus de treino Português e Inglês, ou seja, palavras que não tenham tradução, como por exemplo alguns nomes próprios.

Relativamente a conclusões que podem ser retiradas deste trabalho, destacam-se as seguintes:

- A utilização de um corpus de treino com um âmbito equivalente ao objectivo pretendido (neste caso não foi possível a utilização de um corpus da área de notícias televisivas);
- A utilização de um modelo de linguagem do âmbito do trabalho, interpolado com um outro modelo de um contexto semelhante (neste caso foi utilizado um modelo de linguagem de textos de jornal);

- A criação e utilização de um corpus de desenvolvimento limpo, ou seja, com frases bem estruturadas e traduzidas correctamente;
- A utilização de algum processamento pós-tradução, como é o caso da utilização de *features* descritas no capítulo anterior, que possam maximizar o valor de BLEU do sistema, ao ser escolhida a melhor frase de entre as n possíveis.

A conjugação de todos estes pressupostos resultaram na construção de um sistema de tradução no contexto das notícias televisivas com um valor de BLEU de 0.5088.

Relativamente ao corpora utilizado, também este sofreu algumas alterações ao longo das várias experiências realizadas. O corpus utilizado para o treino do sistema de tradução foi sempre baseado do âmbito das sessões do Parlamento Europeu uma vez que não existem recursos suficientes para o contexto das notícias televisivas. Quanto ao corpus utilizado para a construção do modelo de linguagem, tal como aconteceu no caso anterior, começou por ser utilizado um corpus baseado nas sessões do Parlamento Europeu (Koehn 2005a), sendo que mais tarde, passou a ser utilizado um corpus baseado exclusivamente no domínio das notícias televisivas (MacIntyre 1998) e acabando por no final, ser construído um modelo de linguagem através da interpolação de dois modelos de domínios diferentes, nomeadamente das notícias televisivas e dos jornais (Graff 1995). Quanto ao corpora utilizado para o *tuning* e teste do sistema foi sempre baseado no âmbito das notícias televisivas. No entanto, este corpora sofreu algumas correcções, de forma a que o sistema produzisse uma tradução com melhor qualidade e consequentemente apresentasse um melhor valor de BLEU. A tabela 5.1 ilustra o tipo de corpora a utilizar e as respectivas descrições, para o domínio das notícias televisivas.

Para uma melhor compreensão das experiências realizadas neste trabalho, a tabela 5.2 descreve todas essas experências, apresentando uma breve descrição das mesmas e os respectivos valores de BLEU obtidos.

Tipo de corpus	Descrição
Modelo de Linguagem	Corpus constituído por um conjunto de frases, baseadas no domínio das notícias televisivas e jornais, somente escritas na língua destino.
Treino do sistema	Corpus paralelo, baseado no âmbito das sessões do Parlamento Europeu.
Desenvolvimento	Corpus paralelo, baseado no contexto das notícias televisivas.
Teste	Corpus paralelo, baseado no âmbito das notícias televisivas.

Tabela 5.1: Corpora utilizado e respectivas descrições.

		BLEU	
	Descição da Experiência	Condição A	Condição B
Experiência 1	Sistema base, com modelo de linguagem baseado nas notícias televisivas, corpus de treino baseado nas sessões do Parlamento Europeu e corpora de *tuning* e teste refinados	0.4776	0.4722
Experiência 2	Interpolação do Modelo de Linguagem	-	-
Experiência 3	Sistema com Modelo de Linguagem Interpolado	0.4861	0.4799
Experiência 4	Aumento do corpus de treino do sistema de *Recaser* com textos de jornais	0.4799	0.4790
	Sisema de capitalização automática	-	0.4071
Experiência 5	Processamento das traduções obtidas (novas *features*)	-	0.5088

Tabela 5.2: Experiências realizadas.

Bibliography

Berger, A. (1996, March). A maximum entropy approach to natural language processing. In *Computational Linguistics*.

Bisani, M. (2006). The 2006 rwth parliamentary speeches transcription system. lehrstuhl fur informatik 6 - computer science dept. Aachen University, Germany.

Brown, P. (2003). The mathematics of machine translation: Parameter estimation. In *Computational Linguistics*, Volume 19, pp. 263–310.

Charoenpornsawat, P. (2002). Improving translation quality of rule-based machine translation. Thailand. Information Research and Development Division. Nation Electronics and Computer Technology Center.

D Jurafsky, J. M. (2000). Speech and language processing. Publisher: Prentice Hall.

Daniel Marcu, W. W. (2002). A phrase-based joint-probability model for statistical machine translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing, EMNLP*.

Dempster, A. P. (1977). Maximum likelihood from incomplete data via the em algorithm. In *Journal of the Royal Statistical Society, B*, Volume 39, pp. 1–38.

Diamantino Caseiro, I. T. (2006, July). A specialized on-the-fly algorithm for lexicon and language model composition. transactions on audio, speech, and language processing. Volume 14, pp. 1281–1291.

Doddington, G. (2002). Automatic evaluation of machine translation quality using n-gram co-occurrence statistics.

Franz Och, D. G. (2004, May). A smorgasbord of features for statistical machine translation. In *HLT / NAACL*, Boston.

Germann (2001). Fast decoding and optimal decoding for machine translation. In *Proceedings of ACL-01*, Toulouse, France.

Germann (2003). Greedy decoding for statistical machine translation in almost linear time. In *Proceedings of HLT-NAACL*, Edmonton, Canada.

Gispert (2006, August). Improving statistical word alignment with morpho-syntactic transformations. In *Fintal 2006 5th International Conference on Natural Language Processing*, Turku, Finland.

Graff, D. (1995). Ldc catalog number ldc95t21, isbn 1-58563-053-5.

Hutchins, J. (1986). Machine translation: past, present, future. In *Computers and their Applications*, Ellis Horwood, Chichester, UK.

Hutchins, J. (2001, September). Towards a new vision of mt. In *MT Summit VIII*, Santiago de Compostela, Galicia, Spain.

Josef, F. & O. H. Ney (2003). A systematic comparison of various statistical alignment models. *Computational Linguistics 29*, 2003.

Knight, K. (2005). Machine translation in the year 2004. In *Proceedings of ICASSP*.

Knight, K. (2006). Translation with finite-state devices. CA 90292.

Koehn, P. (2003, December). Noun phrase translation. University of Southern California.

Koehn, P. (2005a, September). Europarl: A parallel corpus for statistical machine translation, mt summit 5, the 10th machine translation summit. Phuket, Thailand. Association for Computational Linguistics.

Koehn, P. (2005b). Introduction to statistical machine translation.

Koehn, P. (2007, June). Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, Prague, Czech Republic, pp. 177–180. Association for Computational Linguistics.

MacIntyre, R. (1998). Ldc catalog number ldc98t31.

Marcu, D. (2006). Spmt: Statistical machine translation with syntactified target language phrases. 4640 Admiralty Way, Suite 1210, Marina del Rey. Language Weaver Inc. CA 90292.

Matsoukas, S. (2005). The english broadcast news transcription system. In *Proceedings interspeech2005*.

Meinedo, H. (2003). Audimus.media: A broadcast news speech recognition system for the european portuguese language. Spoken Language Systems Lab, INESC-ID. Lisboa, Portugal.

Meinedo, H. (2008). *Audio Pré-Processing and Speech Recognition for broadcast news*. Ph. D. thesis, IST/UTL, Lisbon, Portugal.

Mostefa, D. (2006). Evaluation of automatic speech recognition and speech language translation within tc-star: Results from the first evaluation campaign. Paris, France. Evaluations and Language resources Distribution Agency (ELDA).

Ney, H. (2005). One decade of statistical machine translation: 1996:2005. In *Human Language Technology and Pattern Recognition*, Germany. Lehrstuhl informatik VI-Computer Science Department: RWTH Aachen.

Nguyen, L. (2005). The recognition systems for english broadcast news and conversational telephone speech. In *Proceedings interspeech2005*.

Nicola Ueffing, H. N. (2001, March). An eficient a* search algorithm for statistical machine translation. In *Proceedings of the ACL-2001, Workshop on Data-Driven Machine Translation*, Toulouse, France, pp. 55–62.

Nicola Ueffing, H. N. (2004). Lehrstuhl f¨ur informatik vi. bayes decision rules and confidence measures for statistical machine translation. In *Computer Science Department RWTH*, University Ahornstrasse 55, 52056, Aachen, Germany.

Och, F. (2003, July). Minimum error rate training in statistical machine translation. In *Procedings of the 41st Annual Meeting of the Association for Computational Linguistics*, Information Sciences Institute. University of Southern California, 4676 Admiralty Way, Suite 1001, Marina del Rey, CA 90292.

Och, F. J. (1995). Maximum-likelihood-schätzung von wortkategorien mit verfahren der kombinatorischen optimierung. Germany. Universität Erlangen-Nürnberg.

Och, F. J. (1999, June). An efficient method for determining bilingual word classes. In *Ninth Conf. of the Europ. Chapter of the Association for Computational Linguistics*, Bergen, Norway, pp. 71–76.

Och, F. J. & H. Ney (2003). A systematic comparison of various statistical alignment models. Volume 29, pp. 19–51.

Papinemi, K. (2001). Bleu: a method for automatic evaluation of machine translation. IBM Research Report.

Philip Koehn, D. M. (2003). Information sciences institute, department of computer science. University of Southern California.

Powell, M. J. D. (1998). Direct search algorithms for optimization calculations. *Acta Numerica 7*, 287–336.

Quirk, C. (2006). Do we need phrases? challenging the conventional wisdom in statistical machine translation. Redmond, Washington, USA. Microsof Research, One Microsoft Way.

Rui Amaral, H. M. (2006, November). Automatic vs. manual topic segmentation and indexation in broadcast news. In *IV Jornadas en Tecnologia del Habla*, pp. 123–128.

Stolcke, A. (2002, September). Srilm - an extensible language modeling toolkit. In *Proc. International Conference on Spoken Language Processing*, Volume 2, pp. 901–904. Denver, CO.

Tetko (1995). Neural network studies and comparison of overfitting and overtraining. In *Information Computer Science*, pp. 826–833.

Trujillo, A. (1999). Translation engines: Techniques for machine translation. New York, pp. 5–6, Cap1. Springer.

Wahlster, W. (2000). Verbmobil: Foundations of speech-to-speech translation. New York. Springer-Verlag.

Yamada (2001). A syntax-based statistical translation-model. In *Proceedings of ACL 39*.