INSTITUTO SUPERIOR TÉCNICO
Universidade Técnica de Lisboa

Automatic Speech Translation

Nuno Miguel Machado Grazina

Dissertação para obtenção do Grau de Mestre em
Engenharia Informática e de Computadores

Presidente:	Professora Doutora Maria dos Remédios Vaz Pereira Lopes Cravo
Orientador:	Professora Doutora Maria Luísa Torres Ribeiro Marques da Silva Coheur
Co-orientador:	Professora Doutora Isabel Maria Martins Trancoso
Vogais:	Professora Doutora Amália Mendes

Novembro 2010

Agradecimentos

O meu muito obrigado à Professora Luísa Coheur por ter tido confiança em mim desde o ínicio, pelos conselhos, pela orientação, pelas correcções, pelos incentivos e pela disponibilidade invulgar de me ajudar a dar o meu melhor. O meu muito obrigado à Professora Isabel Trancoso pelas suas intervenções sempre construtivas e pelas ideias de que eu nunca me teria lembrado.

Gostaria ainda de agradecer a todos no L2F cuja ajuda e trabalho foram muito preciosos, Hugo, Fernando, Wang e Ana, e em especial ao Tiago por toda a paciência, boa disposição, pelas respostas a inúmeras perguntas e pelas soluções de inúmeros problemas.

Obrigado a todos os meus amigos pelos ”já falta pouco” e pelos ”vai correr tudo bem” mesmo quando não faziam a mínima ideia do que eu estava a falar.

Finalmente não podia deixar de agradecer aos meus pais pela ajuda e apoio constantes, por sempre me terem dado tudo e por terem acreditado sempre que eu seria capaz. E à Carolina por ter estado sempre do meu lado mesmo quando não tinha razão, por ter abdicado de algumas sestas para ler e perceber documentos infindáveis e por ser capaz de me fazer sempre acreditar que sou melhor do que aquilo que penso.

Muito obrigado a todos os co-autores deste trabalho!

Lisboa, November 12, 2010
Nuno Miguel Machado Grazina

Abstract

Global communication and understanding play an increasingly vital role in shaping our economic and social environment. However, language differences are a great barrier in achieving true global understanding and knowledge sharing. Human translators are often unavailable or are prohibitively expensive, and cannot deliver much needed information in a timely and usable manner. Thus, if a system capable of automatically performing translations with the same accuracy levels as a human being was to be created, it would be greatly beneficial in allowing true cross-lingual communication. The field of Spoken Language Translation aims at developing such systems, but despite seeing large improvements in the last few years, it still fails to achieve its goals for the majority of languages and real scenarios.

This work's main objective is to build and perform experiments on automatic translation systems for unlimited domains such as Broadcast News and presentations and for the specific case of the Portuguese-English language pair.

This work comprises a historic overview of the Spoken Language Translation field, focused primarily on Machine Translation, which presents the evolution of the main technologies and resources involved in the process of translating spoken language, followed by a review of several state-of-the art techniques aimed at improving translation accuracy. And it also includes the description of the work carried out with the objective of facing some of the current challenges in this field, such as dialect adaptation of Brazilian resources to European Portuguese, in order to improve the training process of a European Portuguese-English translation system, and various translation experiments under different conditions followed by an error analysis with the goal of understanding how these errors relate to the overall translation quality itself.

Resumo

A comunicação e compreensão globais, desempenham cada vez mais um papel vital naquilo que é o nosso contexto económico e social. No entanto, as diferenças entre línguas são uma grande barreira para alcançar a verdadeira compreensão e partilha de conhecimento globais. Tradutores humanos estão, frequentemente, indisponíveis ou são excessivamente dispendiosos, e não são capazes de gerar informação necessária de forma atempada e usável. Desta forma, se um sistema capaz de efectuar, automaticamente, traduções com os mesmos níveis de precisão de um humano fosse desenvolvido, traria enormes benefícios ao permitir, efectivamente, uma verdadeira comunicação entre diferentes línguas. A área da Tradução de Língua Falada visa desenvolver sistemas deste tipo, mas apesar de ter sofrido grandes evoluções nos últimos anos, ainda não é capaz de atingir os seus objectivos para a maioria das línguas e de cenários reais.

O foco deste trabalho é construir e efectuar experiências em sistemas de tradução automática para domínios ilimitados como Notícias Televisivas e apresentações orais, e para o caso específico do par de línguas Portugês-Inglês.

Este trabalho é constituido por uma revisão histórica da área da Tradução de Língua Falada, focada principalmente na Tradução Automática, que descreve a evolução das tecnologias e dos recursos envolvidos no processo de tradução de língua falada, seguida de uma revisão de diversas técnicas usadas, actualmente, para melhorar os resultados de tradução. Inclui ainda a descrição do trabalho levado a cabo com o objectivo de enfrentar alguns dos desafios neste campo, tais como adaptação de dialecto de recursos Brasileiros para Português Europeu de modo a melhorar o processo de treino de um sistema de tradução Português Europeu-Inglês, e uma série de experiências de tradução sob diferentes condições seguida de uma análise dos erros presentes com o intuito de compreender de que forma estes erros se relacionam com a tradução final em si.

Keywords
Palavras Chave

Keywords

Spoken Language Translation

Automatic Speech Recognition

Machine Translation

Linguistic Resource Improvement

Dialect Adaptation

Error Analysis

Palavras Chave

Tradução de Língua Falada

Reconhecimento Automático de Fala

Tradução Automática

Melhoria de Recursos Linguísticos

Adaptação de Dialectos

Análise de Erros

Table of contents

1 Introduction **1**

1.1 Work objectives 2

1.2 Document structure 2

2 Related Work **5**

2.1 Historic overview 5

2.1.1 Early work 5

2.1.2 Interlingua approach 6

2.1.3 Statistical approach 7

2.1.4 Recent work 8

2.1.5 Evaluation campaigns 9

2.1.6 Avaliable corpora 10

2.1.7 Available tools 10

2.1.8 Evaluation metrics 11

2.2 Improvements in machine translation 12

2.2.1 Machine translation evaluation using automatic sentence segmentation 12

2.2.2 Integration between ASR and MT 13

2.2.3 Domain adaptation 16

2.2.4 Other improvements 20

3 Adaptation of Brazilian Portuguese texts to European Portuguese **25**

3.1 Background in exploiting similar languages to improve translation quality 26

3.2 Main differences between EP and BP 27

3.2.1 Third person clitics 27

3.2.2 Pronominal clitics 28

3.2.3 Gerunds 28

3.2.4 Time distances 28

3.2.5 Using the verbs *ter* and *haver* to express existence 28

3.2.6 Article before a possessive 29

3.2.7 Way of addressing someone 29

3.2.8 Lexical differences 29

3.3 Adaptation from BP to EP (BP2EP) 29

3.3.1 Linguistic resources and tools 30

3.3.1.1 MARv 30

3.3.1.2 The CETENFolha corpus 31

3.3.1.3 The CETEMPublico corpus 31

3.3.1.4 A list of contrastive EP-BP pairs 31

3.3.1.5 An extensive list of EP verbs and all respective forms 31

3.3.1.6 TED talks and respective BP and EP translations 31

3.3.2 BP2EP 32

3.3.2.1 Handling of clitics (third person and pronominal) 32

3.3.2.2 Handling of gerunds 34

3.3.2.3 Handling of time distances and expressions denoting existence 35

3.3.2.4 Handling of the lack of article before the possessive 36

3.3.2.5 Handling of way of addressing 36

3.3.2.6 Handling of lexical differences 38

3.4 Evaluation 39

3.4.1 Objective evaluation 40

3.4.2 Subjective Evaluation 42

3.4.3 Final adaptation experiment 45

4 Error analysis for speech-to-speech machine translation 47

4.1 Background in translation error analysis 48

4.2 Early experiments 49

4.2.1 Corpora and resources 49

4.2.2 Experimental settings 49

4.2.2.1 Broadcast News setup 50

4.2.2.2 Talk setup 53

4.3 Types of errors 54

4.3.1 Missing words 55

4.3.1.1 Missing filler words 55

4.3.1.2 Missing content words 55

4.3.2 Word order 56

4.3.3 Incorrect words 57

4.3.3.1 Lexical choice 57

4.3.3.2 Disambiguation 58

4.3.3.3 Incorrect forms 58

4.3.3.4 Extra words 59

4.3.3.5 Idiomatic expressions 60

4.3.4 Unknown words 62

4.3.5 ASR errors 62

4.3.6 Punctuation and capitalization 63

4.4 Influence of errors on system performance 64

5 Conclusions and future work 69

5.1 Future work in BP2EP 70

5.2 Future work in error analysis 71

I Appendices 81

A Examples of speech translation systems and projects 83

A.1 JANUS 83

A.2 NESPOLE 85

A.3 Verbmobil 86

A.4 Babylon/TRANSTAC 88

A.4.1 Mastor 88

A.4.2 CMU TRANSTAC and IraqComm 89

A.5 GALE 90

A.6 Lecture translator 92

B Systems results in recent editions of IWSLT and WMT 95

C Running BP2EP 97

D Subjective evaluation of BP2EP 99

List of Figures

1.1 Traditional architecture of a SLT system 2

3.1 System architecture of BP2EP 32

4.1 Baseline system architecture 50

4.2 System architecture with punctuation and capitalization module 51

4.3 Simplified error taxonomy for error classification used in this work 54

B.1 Results of several systems in IWSLT editions of 2007 and 2008 95

B.2 Results of several systems in WMT editions of 2007 and 2008 95

B.3 Results of several systems in WMT edition of 2009 95

List of Tables

3.1 Common spelling differences between EP and BP 39

3.2 Results of the first evaluation using BLEU 40

3.3 Results of the final objective evaluation using BLEU 41

3.4 Composition of the set of evaluation sentences 42

3.5 Results of the evaluation using the subjective method 43

3.6 Results of the final adaptation experiment 45

4.1 BLEU Results of BN experiments, with and without ASR, for the whole test text 50

4.2 BLEU Results of BN experiments, with and without ASR, only for the spontaneous part of the test text 50

4.3 BLEU Results of BN experiments, with and without ASR, only for the planned part of the test text 51

4.4 BLEU Results of BN experiments using capitalized models and the punctuation and capitalization module 52

4.5 BLEU Results of BN experiments using lowercased models, offline segmentation and the punctuation and capitalization module 52

4.6 BLEU Results of TED talks experiments, with and without ASR, for the whole test text . . 53

4.7 BLEU Results of TED talks experiments using lowercased models, offline segmentation and the punctuation and capitalization module 54

4.8 Statistics of the test TED talk translation 64

4.9 Frequency for each type of errors present in the TED talk test text 64

4.10 Partial and total BLEU results for the test TED talk after correcting each type of errors . . 65

4.11 Partial and total BLEU results for the test TED talk after using the corrected text as an extra reference 67

1 Introduction

In an increasingly globalized world, mutual understanding and communication play a vital role. In order to achieve global competitiveness and effectiveness, governments, companies, research organizations continuously face the challenge of depending on high-quality information delivered in a natural, effective and timely manner. The main issue is that, given this global setting, most of the needed information will not be available in one's own native language, thus not having any real application if it cannot be interpreted (Waibel & Fügen, 2008).

Now more than ever, the ability of building cross-lingual communication channels and global interaction is an important driving force in shaping our economic environment and, ultimately, our lives. However, language is still a great barrier that needs to be overcome. Although there are some common languages to certain groups, such as English or Spanish, people's skills in these languages may vary greatly, which prevents any true communication. Human translators are often unavailable or highly expensive, with costs rising to hundreds of millions of Euros per year in organizations such as the European Commission whose members speak twenty-three different languages. But if effective solutions were to be developed which would enable real language integration and equality, economic and social benefits would be very high.

These are the challenges that the field of Spoken Language Translation (SLT) has been working to solve in the past two decades. SLT encompasses three component technologies: automatic speech recognition (ASR), machine translation (MT) and speech synthesis (TTS). But despite the increasing access to faster and larger computational resources, advanced machine learning techniques and virtually unlimited web resources, the current state of these core technologies still fail to produce sufficiently accurate and completely usable results, which stresses the need of further research in order to enable true human-human cross-lingual communication. Figure 1.1 shows the traditional cascaded architecture of a SLT system.

Several SLT systems and projects have been developed throughout the years with distinct objectives, approaches and scopes. PT-STAR (Speech Translation Advanced Research to and from Portuguese) is one of these SLT projects. It is part of the CMU-Portugal cooperation agreement, jointly conducted by the Language Technologies Institute (LTI) from Carnegie Mellon University (CMU) and a Portuguese consortium formed by the Spoken Language Systems Lab (L2F) from INESC-ID Lisboa, the

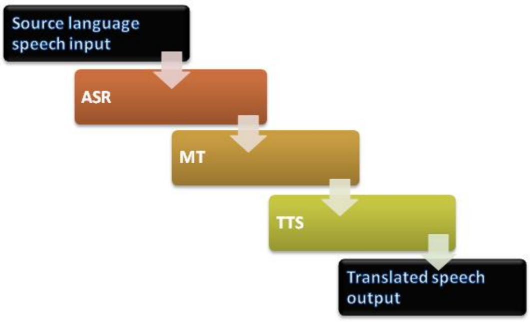


Figure 1.1: Traditional architecture of a SLT system

Center of Linguistics of the University of Lisbon (CLUL) and the University of Beira Interior (UBI).

Its main objective is to improve SLT's component technologies and their interfaces in order to develop Speech-to-Speech (S2S) systems for the English- European Portuguese language pair. It encompasses three major tasks: interface between ASR and MT components, interface between MT and TTS components, and the MT engine itself.

The work described in this dissertation is developed within the PT-STAR framework and focuses heavily on MT. Moreover, since one of the most used forms of disseminating knowledge and fostering the discussion of new ideas is through lectures and presentations, the translation of such speeches, along with Broadcast News, are the main targets of this research.

1.1 Work objectives

The objectives of this work are the following:

- To review various state-of-the art approaches aimed at improving the quality of speech translation;
- To apply the learned concepts and available resources, such as Europarl (see Section 2.1.6) and Moses (see Section 2.1.7), by performing baseline experiments on two systems: a) a Broadcast News (BN) translation system b) an oral presentation (talk) translation system;
- To adapt BP texts to EP in order to improve the training process;
- To perform an error analysis on the developed systems with the goal of understanding how errors influence automatic metric scores.

1.2 Document structure

This document is structured in the following chapters:

- Chapter 2 presents a review of related work. It has two main sections: an historic overview of the SLT field and a review of local improvements within the speech translation process.
- Chapter 3 presents an overview of related work in exploiting similar languages to improve translation quality and focuses on the development of the approach used to adapt BP resources to EP.
- Chapter 4 presents an overview of related work on translation error analysis, describes early baseline experiments under different conditions and performs an analysis of the errors produced by these experiments and discusses their influence on evaluation scores.
- Chapter 5 presents the conclusions drawn with this work and how they can be taken into account to develop further improvements.

2 Related Work

This chapter presents an overview of the field of SLT and describes various generalized types of speech translation technologies, approaches and systems that relate to the objectives of this work. It comprises two subsections: an historic overview of technologies and approaches used in SLT and a description of several improvements in specific areas within the SLT field.

2.1 Historic overview

This section presents an historic overview on the field of SLT. It describes the main challenges encountered until today as well as the most used MT paradigms used in this period and some of the resources and techniques currently available to researchers.

2.1.1 Early work

Interest in speech translation began in the 1980s and early 1990s, and although these early systems were very limited, they proved that speech translation was indeed possible. In order to advance the state-of-the-art of component technologies, several initiatives were launched. Among these early initiatives, the one that stands out the most is perhaps the Consortium for Speech Translation Advanced Research (C-STAR) which provided a major thrust in international academic and industrial cooperation and whose members are responsible for developing many of the approaches and technologies that shape today's speech translation landscape.

While proving that SLT was indeed possible, early systems were still very limited since the user had to follow a very strict speaking style and had to know exactly which sentences were allowed by the system. These shortcomings rendered the systems unusable as humans do not speak in a well behaved manner, and cannot constrain themselves to a limited set of sentences and syntactic patterns.

In the early 1990s, a way of turning SLT into an usable technology began to be explored. Although humans speak in a spontaneous way and, generally, cannot deal effectively with imposed language constraints, there are many regular and daily tasks where the domain discourse is inherently limited. By exploiting these limitations, investigators hoped to create a practical way of using SLT. Such domains can be, for instance, appointment negotiation, travel planning and its sub-domains (hotel booking, car rental), medical assistance, force protection and military missions, and many others.

Investigators also focused their attention in other important aspects of these systems: accuracy, speed and human-factors. They had to provide acceptable translations in a reasonable amount of time (close to real-time), while addressing usability issues to allow the system to be effectively used in a real situation.

Although the domains of application were limited, the user usually speaks in an unrestricted, uncontrolled and spontaneous way, and in order to guarantee the quality of the translation, technology had to be developed to deal with the spontaneity of the speech input, both in the recognition and in the translation itself. With these systems, two MT paradigms emerged: at first the Interlingua approach, which is knowledge-driven and rule-based, and later the direct statistical approach (SMT).

2.1.2 Interlingua approach

The Interlingua approach relies on the process of mapping a source language fragment into an abstract and language-independent representation, and generating the target language translation from this intermediate representation. This knowledge-based abstraction consists on a set of syntactic and semantic rules that are the interface between analysis of the source language and the generation of the target language.

By creating an intermediate representation, this approach eliminates the need of direct translations and stimulates the parallel and independent development of translation interfaces between different source languages and the Interlingua, and language-specific analysis systems, allowing the simultaneous translation from many source languages to many target languages, without the need of repeating all of the work for each language pair.

Several Interlinguas were developed (Levin et al., 2002), (Lonsdale et al., 1994), some outside the speech-to-speech translation context, but nevertheless, most follow similar aspects: they try to convey meaning of sentences by representing knowledge (meaning and concepts) using symbols and establishing relations between these symbols and their roles in a given sentence.

For limited-domain systems such as JANUS and NESPOLE (see Appendices A.1 and A.2), the Interlingua approach provided a seemingly suitable way of achieving good translations as the concepts and their relations are restricted to the domain and can be easily enumerated and represented.

2.1.3 Statistical approach

Statistical machine translation (Brown et al., 1993) emerged in the early 1990s as the result of investigation led by IBM and refers to a MT paradigm that treats automatic translation as a machine learning problem. This means that the basic idea behind it is that by applying a learning algorithm to a large collection of translated text (parallel corpus), a machine can learn how to translate previously unseen text.

Several tutorials on this subject are available such as (Knight, 1999) and (Knight & Koehn, 2003) and a thorough survey has been proposed in (Lopez, 2007).

There are three major components to a SMT system: the translation model (TM), the language model (LM) and the decoder. The translation model has the goal of matching a source language string (*f*) to its target language counterpart (*e*) by estimating the conditional probability $p\left(f|e\right)$; the language model has the purpose of evaluating how well a sentence is written in the source or target language by estimating *p(e)*; the decoder applies an algorithm that given a source language string *f*, estimates the most likely target language string *ê*, using the Bayes's rule as the product of the translation model $p\left(f|e\right)$ and the language model *p(e)*:

$$\hat{e} = argmax_e[p(e|f)] = argmax[p(f|e)p(e)] \tag{2.1}$$

A major issue in this technique is how to obtain accurate translation models based on limited information present in parallel texts. $p\left(f|e\right)$ cannot be automatically estimated and, instead, it has to be learned from the limited parallel corpus. Thus, the maximization expectation (EM) algorithm is employed to perform the parameter estimation from information which is unobserved in the training data. Still, a simple estimation alone is not enough to obtain high quality translation models. One must also account for structural differences between source and target languages, namely word alignment and reordering as the following and very popular example clearly illustrates:

English sentence: *Mary did not slap the green witch.*

Spanish sentence: *Maria no daba una bofetada a la bruja verde.*

In the example there are words that are aligned with more than one word (*slap*, *daba una bofetada*) and others that appear in different orderings (*green witch*, *bruja verde*). The IBM and HMM alignment models are widely employed to automatically extract these alignments and widely used as basic setups, but several heuristics described in (Och & Ney, 2000) and (Koehn et al., 2003) have been developed to extend these models.

However, performing translation at just word level still does not produce accurate translations as phrasal context and syntactic information are likely to be lost. Thus, more robust techniques of extracting bilingual alignments are needed.

Phrase-based SMT is one such technique and aims at complementing word alignments by segmenting the input into smaller sentence-like units, translating them individually and finally reordering them. Most importantly, this allows for the system to perform many-to-many translations and to capture local contexts in translation. Popular phrase-based models are the Word alignment induced phrase model (Koehn et al., 2003), the Alignment template (Och & Ney, 2004) and the Joint phrase model (Marcu & Wong, 2002).

Syntax-based SMT tries to solve some of the phrase-based approach shortcomings, mainly the inability of performing word reordering for syntactic reasons and the use of syntactic models to disambiguate syntax-dependent translations. Still, approaches based on inverse transduction grammars and syntax trees (D. Wu, 1997), hierarchical finite state transducers (Alshawi et al., 1998) and syntax-based translation and language models (Yamada & Knight, 2001), (Charniak et al., 2003) face even greater obstacles than phrase-based systems since it is hard to obtain parallel data on syntactic transfer, there are few foreign language-specific syntactic parsers and often there are great differences in syntactic structure between languages.

As for the language model, the standard approach is based on n-grams, where one only needs to perform n-gram counts to compute their probabilities. Other approaches have been used such as syntactic LMs and the use of non-parallel corpora, mainly by counting n-gram frequencies on the web (Soricut et al., 2002) or using suffix-trees (Munteanu & Marcu, 2002).

There are several techniques for decoding such as a greedy one (Germann, 2003), building the translation by expanding hypotheses using a beam-search algorithm with pruning strategies to reduce the search space, using the same beam-search approach in building word graphs which allows hypotheses recombination, finite state transducers (Alshawi et al., 1997), (Knight & Al-Onaizan, 1998), and tree parsing (Yamada & Knight, 2002).

2.1.4 Recent work

From 2003, much effort has been put into overcoming the barrier of unrestricted and domain unlimited speech translation, by launching projects, such as DARPA[1]'s Transtac and GALE (see Appendices A.4 and A.5), aimed at developing techniques and algorithms capable of producing high quality translations of completely unrestricted speech where the possible vocabulary and the domain variety are virtually unlimited, for instance, broadcast news and political speeches. This lack of domain constraints has

[1] Defense Advanced Research Projects Agency (http://www.darpa.mil/)

greatly limited the application of knowledge-driven translation approaches such as Interlinguas, and instead, the statistical approach at machine translation gained popularity, despite its drawbacks, due to its ability to naturally cover the unlimited semantic variations present in unrestricted speech, and has become today's most popular MT paradigm.

These DARPA initiatives were pioneer in stimulating academic and industrial research and competition since, unlike early speech translation projects, they were open-call programs to which any research institute or company could apply with their own systems, with well-defined objectives and requisites, tight schedules and extensive evaluations. This constant competition greatly contributed to rapid tecnhological and scientific evolution in speech translation systems.

Effort has also been put into lowering the entry barriers to new investigators by providing publicly available large multilingual corpora and open-source tools to quickly deploy full SMT prototypes, in addition to the highly competitive open-call strategy already used in DARPA programs.

A specific case of domain-unlimited translation is perhaps one of the most challenging fields in today's SLT. The simultaneous translation of lectures, seminars and presentations further broadens the scope of spontaneous speech translation as it encompasses both the needs of early systems and the need of developing new algorithms and technologies to solve new problems (Wölfel et al., 2008).

As any spontaneous speech translation system, simultaneous translation of lectures has to deal with the speakers specific voice and speaking style and with issues inherent to the spontaneity of speech, mostly disfluencies, stuttering, accents, self corrections and any kind of ill-formed phrasings, but it also has to deal with the fact that lectures often have very specialized and technical vocabularies, which stresses the need of creating ways of addressing domain adaptation not only at translation time but at recognition time as well.

Another vital aspect of simultaneous translation of lectures is the fact that it cannot be performed by an offline system. While in other open-domain systems the whole process is done offline, simultaneous lecture translation must be done at real-time with little to none latency times, as a foreign listener will not be able to follow the speaker and effectively understand the lecture if he has to wait for the translation even if it is for a short time. Since the system cannot wait for the end of sentences or topics to stop the recognition step and start the translation component, it brings forward the challenge of segmenting the speech in a manner that on one hand allows for translation of meaningful speech chunks, and on the other hand does not increase latency times.

An example of such a system is the Lecture Translator developed by the University of Karlsruhe (see Appendix A.6).

2.1.5 Evaluation campaigns

IWSLT[2] and WMT[3] have been, in the last few years and until today, a major driving force for the advancement of state-of-the-art in MT systems by holding shared evaluation tasks where many research and industrial entities submit and evaluate their own systems. Participants can freely build their systems, based on whichever paradigms they prefer, and using their own state-of-the-art techniques to improve overall system quality; only the basic training corpora and evaluation metrics are common to every participant. This kind of competition stresses the need of readily available training, development and test corpora and tools.

2.1.6 Avaliable corpora

Many public corpora have been made available, but perhaps the most used and most popular is Europarl (Koehn, 2005). Europarl was specifically created to provide SMT systems developers a large amount of aligned bilingual data. It compiles the European Parliament sessions from 1996 to 2006 and includes translations in eleven different European languages, with sizes ranging from about 26M words to 44M. It also includes preprocessing and alignment tools to help developers in transforming the corpus into usable data by standard SMT systems.

Other parallel corpora are also widely used, although not being as general-purpose and as large as Europarl. This is the case of the BTEC[4] and CLDC[5] corpora. The former provides multilingual aligned corpus with basic spoken expressions in the travel domain usually found in phrasebooks for tourists, while the latter is a balanced Chinese-English corpus in several domains with about 156k sentence pairs.

Apart from parallel corpora, there are also monolingual corpora such as Hansard[6] and Cortes which gather the British and Spanish Parliament sessions and contain 57.7M and 50.4M words, respectively.

Still, the available bilingual texts are not enough to keep up with the advancement of systems and technologies and do not even exist for the majority of language pairs. Since building large enough corpora is a very expensive and time-consuming task, there has been some investigation on how to build comparable corpora (Munteanu & Marcu, 2005). Comparable corpora are texts that are not necessarily parallel, but similar in content and addressing overlapping topics, thus turning the Internet into an easily exploitable resource, especially by crawling through news agencies' web pages such as BBC, France Press or CNN where multilingual news feeds are available.

[2] International Workshop on Spoken Language Translation (2009 edition: http://mastarpj.nict.go.jp/IWSLT2009/)

[3] Workshop on Machine Translation (2009 edition: http://www.statmt.org/wmt09/)

[4] Basic Travel Expression corpus

[5] Chinese Linguistic Data Consortium

[6] Aligned Hansards of the 36th Parliament of Canada (http://www.isi.edu/natural-language/download/hansard/)

2.1.7 Available tools

There are, currently, several MT open-source tools from which, clearly the most popular are SRILM (Stolcke, 2002), GIZA++ (Och & Ney, 2003) and Moses (Koehn et al., 2007).

GIZA++ is an extension of the Egypt toolkit and is a training tool that creates words alignments between source and target languages in the training data. Moses is a beam-search decoder for phrase-based SMT which is part of a comprehensive toolkit that also includes a set of software and scripts aimed at building a complete state-of-the-art SMT system. These include phrase-table extraction, parameter tuning, translation and automatic evaluation. While other tools such as Berkeley aligner, IRSTLM (Federico et al., 2008) and the Joshua decoder (Li et al., 2009) are available, GIZA++, SRILM and Moses are still a widely used combination for building and training a full SMT system.

A decoder such as Moses is able to score the candidate translations by using a log-linear interpolation model that combines the weights of the features used to characterize each candidate. These features can either be extracted from TMs, LMs and reordering models (RM). Usually, Minimum Error Rate Training (MERT) (Och, 2003) is used to optimize these weights on a held-out development dataset, which is intended to represent the whole training set, with the objective of maximizing the scores of evaluation metrics.

2.1.8 Evaluation metrics

Most of early SLT efforts were the subject of end-to-end evaluations, placing a great deal of attention on the usability issue. On the speech recognition and translation components, early systems were evaluated by human experts, professional translators, linguists or simply bilingual speakers, as to whether the translations were adequate and fluent, which led to very subjective results, because many times the same sentence or fragment can be translated in many different ways and still be correct. While different people may have diverging opinions on the quality of a translation, combining their evaluations provided a way of unbiasing the results, but ultimately, the evaluations were slow and costly, and reflected the evaluators own views instead of being objective and repeatable experiments.

With the increasing need to evaluate systems in a more objective, quicker and inexpensive way, in the early 2000s some automatic evaluation metrics were developed.

The Word Error Rate (WER) is a very common metric to evaluate speech recognition and MT performance. It derives from the Levenshtein distance string alignment algorithm and uses dynamic programming to align the recognized word sequence with a reference word sequence, where the WER result is the percentage of words present in the reference that had to be inserted, substituted or deleted in the recognized sequence.

The Bilingual Evaluation Understudy (BLEU) (Papineni et al., 2002) is one of the most widely used metrics in MT evaluation. The main idea behind it is that humans can translate the same sentence in various ways and still be correct and can immediately identify when a translation is wrong. By comparing a candidate MT result, typically at a sentence-level, to high quality human references, BLEU assesses how close the candidate is to them. This proximity is calculated by a modified n-gram precision measure, where precisions from unigrams to 4-grams are averaged over the whole translated corpus.

It has been shown that BLEU has a high level of correlation with human judgement, and therefore, has been widely accepted as the standard metric in MT evaluation and as a benchmark for other automatic metrics.

Some other metrics have been developed based on BLEU and trying to overcome some of its shortcomings. For instance, NIST[7] is similar to BLEU in almost every aspect, but with a significant difference: while BLEU scores each matched n-gram equally, NIST assigns higher weights to more rare n-grams, for example, BLEU would score the bigrams *to the* and *suggests improvements* equally, but NIST would assign a higher score to *suggests improvements* as it is less likely to occur.

The Metric for evaluation of translation with explicit ordering (METEOR) (Banerjee & Lavie, 2005), was also designed to fix some of BLEU's flaws. It uses a unigram recall measure along with precision, and is designed to produce good scores at sentence level, where BLEU uses sentence-level scores to obtain good corpus-level results. METEOR includes other features as a stemmer module to allow not only exact word matching but also stem matching, and a synonymy module that uses WordNet[8] to solve BLEU's lack of semantic support.

2.2 Improvements in machine translation

This section presents a review of several techniques aimed at improving specific steps within the S2S translation process, focusing mainly on the MT component and its integration with the ASR one. Among these techniques are improvements at domain adaptation, ASR-MT interface, automatic evaluation and several recent approaches used by top-ranked systems in IWSLT and WMT evaluations.

2.2.1 Machine translation evaluation using automatic sentence segmentation

Evaluation metrics such as BLEU expect that for each source segment the MT systems produces exactly one target segment corresponding to the translation. Thus, it is crucial that manually translated references have that exact same number of segments. Usually, a segment corresponds to a sentence but a

[7] National Institute of Standards and Technology (http://www.nist.gov/index.html)

[8] WordNet (http://wordnet.princeton.edu/)

segment may only be a sequence of words without proper punctuation.

When dealing with speech translation, typically, the concept of sentence is not very clear. The speaker may leave the sentence incomplete, hesitate, correct himself, make long pauses or speak for a long time without any interruptions. Still, human translators are able to, subjectively, understand where a sequence of utterances may be segmented in order to produce sentence-like meaningful segments, but on the other hand, ASR systems cannot perform this. It is possible to force ASR modules to produce an output segmentation based on timestamps defined by a human transcriber, but this approach assumes that this transcription has to be performed earlier, which is, arguably, an unrealistic scenario. In an online system, the correct transcription and its timestamp are unknown to the ASR and segmentation is performed based on prosodic analysis and language model training. In these conditions, ASR and the corresponding MT output will likely be segmented in a different way than the human references, rendering automatic evaluation metrics unusable.

Because performing a manual segmentation is a very lengthy and expensive method, an algorithm that, based on a set of human-made references, automatically re-segments the ASR output in order to match the number of segments in the references while still maintaining coherence between the source-language segment and its corresponding translation, is proposed in (Matusov, Leusch, et al., 2005).

The algorithm is based on a modified Levenshtein distance to calculate the alignment between a sequence of candidate words for the whole document and a sequence of several translation references. The algorithm performs an alignment between the words in each candidate segment and the corresponding segments from each reference, and stores a pointer to which of the reference segments the alignment had the smallest cost. The reference segmentation is transferred to the document of candidate translation segments, simply by backtracking the decisions made by the Levenshtein algorithm for each segment.

Experiments made on the 2004 IWSLT BTEC Chinese-to-English evaluation, by running the algorithm on twenty different systems, showed that WER, BLEU and NIST metrics correlate, at least, as well as before to human judgement, and although scoring slightly lower, the score is not seriously affected by the automatic segmentation, with decreases varying between 0 and 0.3 points. This method greatly improves efficiency in automatic evaluation without seriously compromising the systems' translation quality.

2.2.2 Integration between ASR and MT

The standard approach at SLT consists in a simple cascaded architecture in which the ASR text output is simply passed as the input to the MT component. This structure assumes that the ASR engine produces adequate and fluent transcriptions and that the translations are produced accordingly. However, this is an incorrect assumption as ASR systems do not perform flawlessly and, instead, the MT component usually receives the single best hypothesis generated by the ASR and propagates its errors to the translated text.

It is desirable that later processing steps, as the MT, are able to retain as much information about the input as possible in order to generate more accurate translations, which is hardly accomplished with the traditional decoupled architecture where ASR and MT work independently. To tackle this problem, pioneer attempts at modelling a tight integration between ASR and MT have been devised (Ney, 1999), (Casacuberta et al., 2002) with the objective of making better use of the information present in the original speech input, passing it to subsequent steps and delaying the decision about the final translation as much as possible. Current state-of-the-art approaches borrow heavily from these early efforts.

(Quan et al., 2005) described a way of applying N-best hypothesis lists instead of the standard single-best hypothesis. Instead of simple text outputs and inputs, ASR produces a word graph from which an N-best list of hypothesis is extracted and passed to the MT engine. Similarly, the MT engine produces a new word graph from which another M-best list is extracted. Thus, for each spoken utterance, the system has N*M different translation hypothesis, which are a much more extensive knowledge source than a single hypothesis.

Each entry in the hypothesis list is characterized by eight different scores, two from the ASR and six from the IBM-4 translation model present in the MT component, which are combined in a log-linear interpolated model that effectively integrates both components' features. The weights in the model are tuned in order to minimize WER and maximize BLEU and used to rescore the final hypothesis list and extract the best translation.

Experiments show an improvement from 39.66 to 41.22 BLEU when rescoring a 100*100best list from the Italian-English pair in the BTEC corpus.

(Zhang et al., 2004) followed a similar approach by using an N-best hypothesis strategy and performing an integration of features from both the ASR and MT components in a log-linear model. Two scores from the ASR and five from the MT component are combined along with five other features such as Part-of-Speech (POS) language models and example matching scores.

Experiments performed on the Japanese-English pair in the BTEC corpus have shown that best results are achieved when the number of recognition hypotheses is lowered as much as 5, thus relieving the system of the expected computational burden of processing large hypothesis lists, and when the number of translation hypothesis is 1000. Results show an improvement of 5 in BLEU scores from a simple and unoptimized IBM-4 model to an improved log-linear model with ten combined features and optimized weights. By using the same models and an enhanced one with the total twelve features along with an N-best strategy, scores rise by 3 to a BLEU score of 62.1.

Another approach at coupling ASR and MT, although less straightforward than N-best lists, is the explicit use of word and phrase lattices. Its main purpose is to directly translate spoken input by passing a graph structure that completely represents its hypothesized words by the ASR component to the MT one, thus not losing any information in the process and delaying decisions on the final translation. Among its advantages are larger search spaces and detailed phrase-level information such as acoustic and language model scores. If the search space becomes too large, the use of confusion networks is generally preferred as it is a more compact representation of a lattice. While the approach described in (Quan et al., 2005) uses word lattices, as the output of each component, it does not use them as input of the MT engine. Thus, by making decisions on extracting N-best hypothesis from the lattices on early steps, potentially important information on the spoken utterance is lost.

(Matusov, Kanthak, & Ney, 2005) have proposed a method of directly translating spoken utterances by using word lattices and weighed finite state transducers (WFST) to integrate ASR and MT. Both structures are graph-based which suggests a straightforward integration between acoustic scores and translation modelling. They attempt to estimate a WFST model for translation by creating a representation of the spoken utterance that corresponds to a path from the initial state of the transducer to any of its final states that is labelled with a sequence of source words and that translates it to a sequence of target words, possibly with different lengths. The final goal is to extract the best path, from the alignment that represents all possible translations of the word lattice. It is assumed that probabilities only depend on the immediate predecessor m words, thus the translation model is an m-gram bilingual language model, from which the source side can also be used as the source language model for acoustic recognition and scoring. Each arc in the lattice is scored as the conditional probability of the acoustic signal given a source word hypothesis. Word reordering is also performed as this approach produces target words in the same order as the source ones, which often leads to incorrect translations. Therefore, the use of a target language model and a sliding fixed-sized window for IBM-based reordering ensures translation fluency and adequacy.

Experiments were performed on the Italian-English BTEC corpus and comparisons were made between a standard 1-best decoupled structure and an integrated approach using word lattices and acoustic scores. Results demonstrate BLEU improvements of 2.3.

(Mathias & Byrne, 2006) further explore the use of lattices to develop a framework aimed at extending word lattices into phrase lattices which can be easily used as an improvement of existing text-based translation systems.

Speech translation is carried out by a generative model called Translation Template Model which consists of five distinct steps, each of them producing a probability distribution. Together they form a joint distribution over the source and target training data and over intermediate source and target phrases. These component distributions are formulated so that they can be implemented by WFSTs and a target language acceptor is created by a lattice of translations resulting from simple FST compositions of all distributions. The translation is then extracted from the composed WFST as the path from the initial state to a final state with the least cost. In practice, the composition roughly corresponds to three separate translation steps: the acceptor of the translation (which is the word lattice itself), source word to target phrase translation and source language model.

The main issue is that the ASR component does not produce phrase lattices, and therefore, the composition must be extended in order to include source phrase to target phrase translations instead of source words to source phrases. A Target Phrase Segmentator is formulated as yet another WFST and used as a preprocessing step to transform the word lattice into a phrase lattice. The resulting phrase lattice is used in the FST composition instead of the previously described word lattice. However, extracting phrase lattices from word lattices is still an unsolved modelling problem and the used approach includes less than perfect phrase boundary detection in what is hypothesized by the ASR and the use of external software to count sub-sequences in a WFST.

Experiments carried out with the translation task of Broadcast News from Mandarin to English show improvements in the BLEU score of only 0.2 from the 1-best setup to the use of ASR phrase lattices.

2.2.3 Domain adaptation

One of the most important aspects of translating speech in unlimited domains is the system's ability of maintaining translation quality when the speech domain is different from the one that the system has been trained with. Since the most used approach is the statistical one, systems' performance is likely to suffer severe decreases when tested on data from domains outside the training corpus. On the 2006 WMT translation task, BLEU scores suffered a 10 point decrease when systems were tested on the News Commentary corpora instead of the Europarl corpus which had been used for training. On the 2007 and 2008 editions, the organization provided a smaller News Commentary bilingual training corpus in addition to the standard Europarl and ran separate evaluations, which increased the interest of investigators in developing domain adaptation techniques. The 2009 edition was completely dedicated the evaluation of performance on the News Commentary domain on systems trained with out-of-domain data.

(Civera & Juan, 2007) performed experiments on how to apply the technique of mixture modelling (McLachlan & Peel, 2000) to translation models in SMT, taking advantage of its ability to model heterogeneous datasets by defining partitions and learning specific probability distributions that better fit each of them. By using this technique, an extension of the well-known HMM alignment model is proposed, capable of dealing with context sensitive training data, effectively creating distinct translation models that capture domain-specific contexts and translation processes and producing specialized alignments. The goal of this method is to restrict the search for translation candidates, by using only the translation model which is specific to the domain of each input source language string. Previous work had already been carried out on extending IBM alignment models with mixture modelling, but the HMM had also been identified as suitable for extensions and integration in a phrase-based system.

By using Moses as a phrase-based decoder and extracting the HMM mixture alignments from a concatenation of the Europarl and News Commentary corpora, the use of this technique shows little advantage over the standard HMM model when tested on an Europarl text, but when tested on a more domain sensitive set, such as News Commentary, BLEU increases by almost 1 point.

Other experiments on domain adaptation were performed but agreed on key aspects:

- One could concatenate both training corpora, but it is expected that just because of their relative sizes, the large out-of-domain corpus would simply overwhelm the smaller one, rendering it useless in the training process. Thus, a more robust way of combining the training data is needed, one that is able to influence the decoder into assigning better scores for domain-related translation hypotheses. This is performed by training separate in and out-of-domain language and translation models and giving the in-domain models' features greater weights, or interpolating them using either a linear or a log-linear approach.
- Obtaining bilingual corpora is much more difficult than obtaining an in-domain monolingual corpus either in source or target languages. Also in some domains it is easier to obtain monolingual data in the form of dictionaries containing domain-specific expressions and their translations. Even if such dictionaries are not available, it is fairly easier to compile them manually than creating a bilingual corpus.

For the 2007 WMT shared task, the work carried out by the University of Edinburgh (Koehn & Schroeder, 2007) tested several baseline systems: for instance, only out-of-domain training data, only in-domain training data, simple concatenation of both, two translation models or two language models. The weights were optimized by a try and error approach, in which several weight ranges were tested, and the better ones were chosen simply based on those observations. On the development set and using Moses as the decoder, results showed an increase of 2.5 BLEU from standard out-of-domain training (25.11) to the use of two separate translation models (27.64). The submitted system used an interpolated language model approach and obtained drops of 2-4 BLEU points from an out-of-domain to News Commentary test set in language pairs such as French-English or English-German, but scored higher on the Spanish-English pair (33.26 to 34.17).

Other investigators developed a unifying framework of the two approaches (H. Wu et al., 2008). Their objective was to combine an in-domain translation dictionary with an out-of-domain translation model, to combine an in-domain language model with an out-of-domain one, and to create an in-domain bilingual corpus by simply translating a source-language in-domain corpus and improve translation quality by adding this corpus to the training data and rebuilding the translation model. Following transductive learning proposed in (Ueffing, Haffari, & Sarkar, 2007), the translation quality is incrementally improved by repeatedly translating the in-domain source-language corpus with the improved models until there is no improvement to be made.

An algorithm was developed to integrate the several possible approaches and consisted in creating a phrase table and a LM from the out-of-domain corpus, and another phrase-table from an in-domain dictionary. Both phrase-tables are combined. If a target-language in-domain corpus is available, an in-domain LM is constructed and combined with the first LM and integrated in the standard log-liner model used in the baseline setup. If an in-domain source-language corpus is available, the already built log-linear model and the available SMT system are used to artificially translate the corpus and create a bilingual in-domain corpus. The new corpus is added to the training data and used to successively improve the quality of the log-linear model until there is no more improvement possible.

Since there is no parallel corpus to estimate dictionary probabilities, they can be assigned by three approaches: uniform probability where different translations for the same word have equal probability, for instance, if a word has n possible translations, each of the translations have a 1/n probability of occurring; constant probability where all translations have a fixed probability of occurring; corpus probability where if an in-domain source-language corpus exists, a bilingual corpus is artificially created as described before, and translation probabilities are estimated in the process. Phrase-tables can be combined during translation by Moses, using a discriminative model, where all possible translations are searched in both phrase-tables and used for translation expansion. Phrase-tables can also be combined by a simple mixture model based on linear interpolation. Language models can also be combined by linear or log-linear interpolations.

One experiment was run on the Chinese-English pair on the 2006 IWSLT evaluation and another was run on the English-French domain adaptation shared task of the 2007 WMT.

For the Chinese to English translation, the out-of-domain training was performed with the CLDC data and the in-domain training was performed with the BTEC corpus. Two dictionaries were used: one based on the LDC Chinese-English Translation Lexicon Version 3.0 and the other manually constructed. For the English to French task, the out-domain corpus was Europarl, the in-domain corpus was the News Commentary corpus distributed in WMT 07 and an in-domain dictionary was manually created.

Results showed that by using the all of the described in-domain resources, corpora, dictionaries and transductive learning, on the Chinese to English task, the BLEU score was improved by 8.16 points from 13.59 to 21.75. Results also show that using in-domain translation dictionaries and in-domain monolingual corpora, the score on the English to French domain adaptation task was improved by 3.36

points from 25.44 to 28.80.

(Bertoldi & Federico, 2009) also worked on the idea that monolingual in-domain resources can be automatically translated by a standard SMT configuration (for instance, Moses) and combined with the rest of the training data. Just as had been argued in (Koehn & Schroeder, 2007) earlier, a simplistic approach would be to concatenate the artificially generated in-domain corpus to the out-of-domain one, but this way, the smaller corpus would not improve the training process.

Instead, the combination is performed by extending the decoder to handle multiple translation, reordering and language models. Moses can search for translation hypothesis in both the intersection and the union of all of the available phrase-tables, and regarding the LMs it simply searches, separately, each LM for the most likely translation hypothesis.

Experiments were conducted on the English-Spanish portion of the UN corpus, as the out-of-domain data, Europarl as the in-domain corpus and two synthetically produced parallel versions of the Europarl data, under the assumption that there were only monolingual versions available, one for English and the other for Spanish, by using a system trained only on the UN corpus. Results show that, when training all of the models on the artificially generated parallel corpora and assigning optimal uniform weights when interpolating the translation and reordering models, the system obtains a BLEU score of 23.68. Results also showed an improvement of 5 points when using two language models (in-domain and out-of-domain) over a single out-of-domain language model, from 22.60 to 27.83, and by combining translation models obtained from bilingual synthetic data and out-of-domain data, scores further rise to 28.10. The conclusion drawn is that although combining translation models produces slightly better scores, using two or more language models provide greater gains in translation quality in different domains. Work performed by (Nakov & Hearst, 2007) had also reached the same conclusion.

On the 2009 edition, the National University of Singapore (NUS) (Nakov & Ng, 2009b) entered with an English-Spanish MT system.

In addition to a standard phrase-based SMT setup, non-standard experiments were performed:

- Two different 5-gram LMs were used, one trained in an in-domain News Commentary corpus and the other trained in the out-of-domain Europarl corpus.
- Two phrase-tables and two lexicalized reordering tables were created and merged. A small phrase-table was trained on the in-domain data and a larger phrase-table was trained on the out-of-domain Europarl corpus and they were merged by simply adding to the in-domain phrase-table all of the pairs not present in it but present in the larger Europarl phrase-table. For merging the lexicalized reordering tables, the procedure was the same.

- Cognates were used to improve the quality of phrase pairs. From a linguistic point of view, cognates are words derived from the same root, but following computational linguistics research, these investigators defined cognates as words that are mutual translations in different languages and have similar orthography. The LCSR similarity measure was combined with competitive linking to extract cognates from the training data. The extracted cognates were finally added, artificially, as pairs to the bilingual training corpus in order to influence the IBM models in favor of the domain-specific alignments.

- Finally, improved de-tokenization, recasing and post-editing were used. While the decoder may be unable to translate compound words (which are linguistically correct) such as *self-assured* or *well-rehearsed*, it will probably translate the words separately. Therefore, the tokenizer was changed to split compound words on the character - and de-tokenizer was also changed accordingly. The standard recaser was also improved as it left the unknown words lowercased. Most of the unknown words were named entities (people, organizations, places), which in Spanish are capitalized, thus the improved version of the recaser simply runs over the standard version's output and sets the casing of the unknown words to their original English casing, as well as forcing each sentence to start with a capital letter. Post-editing consisted of rule-based heuristics designed to solve some of the most common errors such as duplicate punctuation or numbering style differences between English and Spanish.

These experiments were conducted as an extension, and in a similar fashion, to the work already carried out by the same author for the 2008 WMT edition (Nakov, 2008), as a member of the Berkeley University. When combining both LMs and translation models, using improved recasing and tokenization and after tuning the weights, the system scored 37.13, which is an improvement of around 3.6 over the baseline configuration.

For the 09 NUS system, results showed a 6 point improvement from the baseline settings to the complete non-standard setup when tested on lowercased data, achieving a BLEU score of 24.4, while scoring 22.37 when performing improved post-casing and post-editing and ranking as the second best system.

2.2.4 Other improvements

With the goal of improving quality in MT output, several techniques, other than domain adaptation and tight coupling of components, have been submitted to recent editions of IWSLT and WMT evaluations, generally included in SMT systems, and have shown promising results.

In the 2007 editions of IWSLT and WMT some of the best scoring single systems came from FBK[9], RWTH[10], CMU and NRC[11] .

The FBK system (Bertoldi et al., 2007) makes use of confusion networks to handle punctuation input before translation rather than following the common approach of recasing and punctuating the output translations. It also exploits the effects of adding language resources to the baseline setup, namely lowercased and truecased LMs, an external 5-gram LM provided by Google and the use of multiple phrase-tables in scoring partial translation candidates.

The CMU system (Lane et al., 2007) includes source language punctuation recovery after recognition, and the incorporation of domain knowledge by rescoring N-best translation hypothesis with additional domain-specific feature-scores and by running clustering techniques to learn each topic distribution over the entire training set. The application of hierarchically structured models to SMT such as the use of probabilistic synchronous context-free grammars and grammar induction is used to extract syntax rules from the training data and augment the standard SMT decoding with syntactic information.

The RWTH submission (Mauser et al., 2007) differs from the former since it combines several SMT systems into a unified one. It combines five systems: three phrase-based decoders, an n-gram component and a syntax-driven hierarchical module. A confusion network is generated for each system where the best hypothesis it produces is aligned with the n-best hypothesis from all other systems. Finally, the five resulting confusion networks are merged into a single lattice which is rescored with system-specific weights and from which the consensus translation is extracted.

In the same year, only in WMT, the PORTAGE system (Ueffing, Simard, et al., 2007) from NRC also achieved high results by improving the training and decoding processes of a standard phrase-based SMT system. These improvements include phrase-tables enhanced with additional feature scores, higher order n-gram LMs and enhanced N-best list rescoring with extra features during decoding.

The 2008 IWSLT edition showed a clear tendency towards system combination as the best way of achieving high scores, which had already been present in earlier editions of WMT and have increasingly been given more importance in subsequent years. Whether to simply try to experiment with the strengths and weaknesses of different translation paradigms and software or to achieve a more principled approach at combining translation hypotheses, two main techniques have been proposed: the use of confusion networks to combine systems' outputs and extract the best path as the consensus translation, and the rescoring of joint n-best lists from all systems with global system-specific feature scores.

The TCH[12] submission (Wang et al., 2008) combines a rule-based MT system and a SMT system

[9] Fondazione Bruno Kessler

[10] Rheinisch-Westfälische Technische Hochschule Aachen

[11] National Research Council Canada

[12] Toshiba (China) Research and Development Center

where the former is used as a first translation step to generate an artificial corpus to be used in the training of the latter. For better MT performance, other techniques are applied to generate improvements at training level by creating better and cleaner training resources such as domain-specific LMs, bilingual dictionaries to improve word alignments, handling of named entities and improved punctuation recovery.

The ICT[13] system (Liu et al., 2008) is also a combination of several systems: a linguistically syntax-based system (syntax augmented) that learns rules in order to extract strings from compact parsing trees, a formally syntax-based system that uses a maximum entropy reordering model, a phrase-based system and another formally syntax-based system based on hierarchical phrases. The N-best translations from each system are merged and reranked based on a sentence level linear combination of global features' weights.

The CASIA system (He et al., 2008) from NLPR[14] uses a very similar approach only with three phrase-based systems instead of just one. The main difference occurs at combination time where N-best lists of translation hypothesis from each system are aligned with each other in order to be merged and transformed into a confusion network. Finally, the decoding of the confusion network produces a new N-best list which is rescored by a log-linear combination of global features' weights.

In the 2007 WMT edition, the ISL system (Paulik et al., 2007) was a combination of a phrase-based system and syntax-augmented one. Joint N-best hypotheses from each system are again merged and rescored with the weights of global features. A distinctive characteristic from other systems is the use of the same rescoring method but combining translations from different source languages into the same target language.

In the 2008 edition, the University of Washington focused on improving reranking of N-best translation lists by developing an enhanced version of MERT in order to better optimize feature weights (Axelrod et al., 2008). The LIMSI[15] system (Déchelotte et al., 2008) further explores N-best lists reranking with the use of in and out-of-domain LMs interpolated with a Neuronal Network LM based on the CSLM method. The TALP-UPC[16] system used an n-gram-based SMT system instead of the more popular phrase-based one. The translation model is a 5-gram bilingual LM which, unlike phrase-based systems, accounts for translation context and dependencies, and linguistically motivated reordering is applied on the decoding step via a LM of reordered source-language POS tags.

In the 2009 WMT edition, the LIMSI submission (Allauzen et al., 2009) built separate systems, one based on a standard Moses setup and a contrasting one based on a similar n-gram setup as described

[13] Institute of Computing Technology Chinese Academy of Sciences

[14] National Laboratory of Pattern Recognition, China

[15] Computer Sciences Laboratory for Mechanics and Engineering Sciences

[16] Center for Language and Speech Technologies and Applications, Technical University of Catalonia

earlier. A similar POS-based context aware and reordering system is also included in both setups. Experiments showed that both setups produced very similar scores, although the n-gram approach proved to scale up into very large corpora in a more natural way.

The 2009 WMT edition also had a shared task dedicated exclusively to the increasingly emerging topic of system combination for MT. Therefore, many participants focus only on how they perform the combination rather than describing each of the component systems.

RWTH also participated in this specific task with a combination of nine German-English MT systems and ten other cross-lingual MT systems with English as the target language (Leusch et al., 2009). The combination algorithm performs a statistical word alignment between translation candidates produced by each system, then a primary system is selected and all systems' hypotheses are reordered regarding the single-best hypothesis from the primary one, a confusion network is generated from this alignment, the process is repeated for each system as the primary one, and finally all resulting confusion networks are merged in a single lattice where the weight of each arc results from a simple weighed voting process. These weights can also be rescored with a LM trained on the outputs of the component MT systems, which results in a bonus towards n-grams present in original hypothesis and in most cases in the original phrases. To generate the best consensus translation, one only needs to extract the path with the least cost in the rescored confusion network.

The CMU system combination (Hildebrand & Vogel, 2009), on the other hand, uses the already mentioned approach of joining N-best lists produced by the component systems in a single list and then rescores it with global system-specific weights. The main issue addressed in this system is the fact that hypotheses in the same sub-list tend to be more similar to each other than to hypotheses present in other sub-lists, which leads to hypotheses from larger lists to score higher in n-gram-based features, even if there is a higher quality candidate in a shorter sub-list. To solve this problem, a sub-list size normalization technique, capable of adapting to various n-best list sizes, was developed and employed. This method consists in normalizing the n-gram agreement feature score. This feature represents the percentage of translations hypotheses in which a given n-gram is present and the technique modifies its score by normalizing the count of hypotheses containing each n-gram with the size of the sub-list it came from.

The system combination from BBN Technologies (Rosti et al., 2009) further tries to improve the use of confusion networks to extract consensus translations. It builds upon the use of Translation Error Rate (TER) software to perform preliminary alignments between pairs of translation candidates, and its limitations such as the fact that when choosing a hypothesis as the primary one, and aligning all others with it may result in the same word in different secondary hypotheses to be inserted in different positions, which leads to the need of having specific handling of each primary hypothesis and each insertion, the fact that it only matches words with identical surface forms and the fact that uses often non-optimal heuristics in determining block shifts. The combination algorithm makes use of sentence specific alignment orders where the order is computed by the edit distance of the hypothesis to an incrementally built confusion network (starting on the primary hypothesis). This means that the hypothesis with the lowest edit cost is aligned. It also uses flexible word matching based on WordNet to find all possible synonyms and words with identical stems in a set of hypotheses. Finally, it uses heuristics to find optimal word blocks that can be shifted without increasing the edit distance by resorting to a paraphrase list. This system combination was ranked first in every task in entered with results similar to those achieved by Google, only surpassed by RWTH's cross-language combination.

3 Adaptation of Brazilian Portuguese texts to European Portuguese

SMT training takes place over large parallel corpora and produce better translations as the training data becomes simultaneously larger and more specialized in a given translation domain. Obtaining parallel corpora with European Portuguese on one of the sides is, currently, a very hard task since parallel European Portuguese-English resources are very scarce, especially when considering such specific and varied domains as the ones addressed in oral presentations, and very expensive to build.

Despite the fact that there are few European Portuguese (EP) usable resources, Brazil, which is another Portuguese-speaking country, has a much larger translation community than Portugal and so, there are much more parallel resources, which are too valuable not to be exploited. Examples of such resources are the ones available at the MIT OpenCourseWare (OCW)[1] website and at the TED talks website[2].

This seemingly perfect solution has a problem: although Portuguese is spoken both in Portugal and Brazil, there are important differences between both EP and Brazilian Portuguese (BP) varieties, and in order to use these available resources in improving the training of translation and language models, language normalization towards the European variety is needed before training.

So, if such texts could be automatically adapted to the European variety, one could expect to quickly increase the volume of parallel training corpora and, therefore, easily improve translation quality.

This chapter describes the approach which was applied in the development of a tool capable of, automatically, transforming BP texts into EP, and how it was evaluated in order to assess how close to the European variety these texts become after being transformed by this tool.

Section 3.1 presents a quick overview of related work about exploiting language similarities, Section 3.2 summarizes the main differences between BP and EP, Section 3.3 describes the proposed approach for BP-EP adaptation and Section 3.4 presents the evaluation methodology and shows the results.

[1] MIT OpenCourseWare (http://ocw.mit.edu/OcwWeb/web/home/home/index.htm)
[2] TED - Ideas worth spreading (http://www.ted.com/talks)

3.1 Background in exploiting similar languages to improve translation quality

Few works are available on the topic of improving MT quality by exploring similarities in dialects, varieties and closely related languages, but the ones available stress the importance of MT among closely related dialects or varieties, since in the majority of the cases, MT is performed among very different languages.

(Altintas, 2002) states that developing an MT setup between similar languages is much easier than the traditional approaches, and that by putting aside issues like word reordering and most of the semantics, which are probably very similar, it is possible to focus on more important features like grammar and the translation itself. This also allows the creation of domains of closely related languages which may be interchangeable and that, in this particular case, would allow, for instance, the development of MT systems between English and a set of Turkic languages instead of only Turkish.

This system uses a set of rules, written in the XEROX Finite State Tools (XFST) syntax, which is based on FSTs, to apply several morphological and grammatical adaptations from Turkish to Crimean Tatar. Results showed that these rules and FST based approach does not cover all variations possible in these languages, and that in some cases, there is no way of adapting the text without an additional parser capable of determining if an adaptation not covered by the rules, should be performed.

Other authors have developed very similar systems with identical approaches to translate from Czech to Slovak (Hajič et al., 2000) and from Spanish to Catalan (Forcada et al., 2001).

(Scannell, 2006) also developed a system to translate Irish to Gaelic Scottish, built with a series of small piped components, such as a POS-tagger, a Naïve Bayes word sense disambiguator and a set of lexical and grammatical transfer rules, based on bilingual contrastive lexica and grammatical differences.

On a different level, (Nakov & Ng, 2009a) describes a way of building MT systems for low resource languages by exploring similarities with closely related and high resource languages. More than allowing translation for low resource languages, this work also aims at allowing translation from groups of similar languages to other groups of similar languages just like stated earlier. This method proposes the merging of bilingual texts and phrase-table combination in the training phase of the MT system.

Merging bilingual texts from similar languages (on the source side), one with the low resource language and the other (much larger) with the high resource language, provides new contexts and new alignments for words existing in the smaller text, increases lexical coverage on the source side and reduces the number of unknown words at translation time. Also, words can be discarded from the larger corpus present in the phrase-table simply because the input will never match them (the input will be in the low resource language). This approach is heavily based on the existence of a high number of cognates between the related languages.

Phrase-table combination can also be achieved after building two separate phrase-tables, each one extracted from each of the bilingual texts, either by using them as alternate decoding paths, by using extra features to indicate from which phrase-table the translation should be fetched, or by interpolating both tables, giving higher weights to the options present in the table extracted from the low resource text. This approach gives preference to translations originated by the higher weighed table and like the previous approach offers a greater number of translation hypothesis.

Experiments performed when both approaches are combined and extended by handling of transliterations between the similar languages, show that when improving Indonesian-English translation with larger Malaysian texts, a gain of up to 1.35 BLEU points is achieved. Similar experiments when improving Spanish-English translation with larger Portuguese texts also show a raise of up to 2.86.

3.2 Main differences between EP and BP

Although both these Portuguese varieties are very similar and understood by both Portuguese and Brazilian natives, in order to adapt Brazilian Portuguese texts to European Portuguese, one must handle several important morphologic and syntactic differences.

These differences comprise the use of pronominal and third person clitics, constructions using the gerund verbal form, the way of addressing someone, the use of an article before a possessive, the use of the verbs *ter* and *haver* to express time distance or when used with the same meaning as *to exist*, and finally, lexical differences in expressing the same meaning, whether being small spelling changes or completely distinct expressions (Mateus et al., 2003).

3.2.1 Third person clitics

BP has dropped the third person clitics and has, instead, replaced them with constructions using the pronouns *ele* (*he*) or *ela* (*she*), or the respective plurals, after a verb.

- Example: I saw **him** on the street.
- BP: Eu vi **ele** na rua.
- EP: Eu vi**-o** na rua.

3.2.2 Pronominal clitics

Usually, in PB, they are placed before a verb, while in EP they are placed in a post verbal syntactic position, linked together with the verb with the use of a hyphen.

- Example: Tell **me** something.
- BP: **Me** diga uma coisa.
- EP: Diga**-me** uma coisa.

3.2.3 Gerunds

Gerunds may have the syntactic function of progressive or secondary predicate. While BP uses, in fact, gerunds in both situations, EP usually uses constructions with the infinitive form, although it may, in some cases, also use the gerund form.

- Example: He was **running**.
- BP: Ele estava **correndo**.
- EP: Ele estava **a correr**.

3.2.4 Time distances

To express time distances, BP uses the verbs *fazer* (literally *to do*) and *ter* (literally *to have*), while EP uses the verb *haver* (roughly *to be* or *to have*).

- Example: He has been in Paris for two years.
- BP: Ele está em Paris **faz** dois anos.
- BP: Ele está em Paris **tem** dois anos.
- EP: Ele está em Paris **há** dois anos.

3.2.5 Using the verbs *ter* and *haver* to express existence

In BP, the verb *ter* is used to express existence, while in EP the verb *haver* is used instead.

- Example: **There is** fire in that house.
- BP: **Tem fogo** naquela casa.
- EP: **Há** fogo naquela casa.

3.2.6 Article before a possessive

In BP, usually there is no article before a pre noun possessive, while in EP, the article is always present.

- Example: I do not know **your** wife.
- BP: Eu não conheço **tua** mulher.
- EP: Eu não conheço **a tua** mulher.

3.2.7 Way of addressing someone

In BP, *você* (*you*) is used as a personal pronoun when addressing someone, in the majority of situations, instead of *tu* (*you*) in EP.

- Example: **You wish** to receive useful packages.
- BP: **Você deseja** receber pacotes úteis.
- EP: **Tu desejas** receber pacotes úteis.

3.2.8 Lexical differences

Just like *color* and *colour*, or *gas* and *petrol* in American and British English varieties, between EP and BP there are also lexical differences whether they are small spelling changes (the majority tends to fade with the recent introduction of a new orthographic agreement between Brazil and Portugal) or completely different expressions.

- Examples: project, bus
- BP: projeto, ônibus.
- EP: projecto, autocarro

3.3 Adaptation from BP to EP (BP2EP)

To automatically handle the mentioned differences and try to adapt BP texts to EP, BP2EP has been developed. Since the main differences are few and easily enumerable, and since there were only a few parallel EP-BP texts available, a rule-based approach was preferred instead of a more flexible statistical approach.

3.3.1 Linguistic resources and tools

In order to build and extract rules for BP2EP, and to evaluate it, the following resources and tools were used.

3.3.1.1 MARv

MARv (Ribeiro et al., 2003) is a POS tagger that runs through a text and identifies to which morphological class a word belongs to. It also provides additional information such as number, genre, verbal forms, moods and tenses and in some cases, the corresponding lemma.

Considering the sentence

E coloco novas imagens porque eu aprendo mais sobre isso.

(And I put new pictures because I learn more about it.)

MARv produces the following output:

```
E|e|4240|4240|Cc********|hmm|

coloco|colocar|4242|4247|V*ip1s_***|hmm|

novas|novo|4249|4253|A****pfp**|hmm|

imagens|imagem|4255|4261|Nc***pf***|hmm|

porque|porque|4263|4268|Cs********|hmm|

eu|eu|4270|4271|Pp**1s_*ns|hmm|

aprendo|aprender|4273|4279|V*ip1s_***|hmm|

mais|mais|4281|4284|R******p**|hmm|

sobre|sobre|4286|4290|S****__**s|hmm|

isso|isso|4292|4295|Pi**_sm*__|hmm|

.|.|4296|4296|O*********|hmm|+SENT|
```

For example, for the second word, the result means that the word *coloco* (*(I) put*) is originated by the word *colocar* (*to put*), it is a verb (which means that the root word is the infinitive form of the verb) in its present tense and indicative mood and for the first singular person.

3.3.1.2 The CETENFolha corpus

The CETENFolha corpus[3] is a collection of texts extracted from the Brazilian newspaper *Folha de São Paulo* and contains about 24 million words.

3.3.1.3 The CETEMPublico corpus

The CETEMPublico corpus[4] is a collection of texts extracted from the Portuguese newspaper *O Público* and contains about 180 million words.

3.3.1.4 A list of contrastive EP-BP pairs

This list, built at L^{2}F, contains about 4000 entries of EP expressions and their respective BP counterparts.

3.3.1.5 An extensive list of EP verbs and all respective forms

This list, developed at L^{2}F, contains over 150000 entries of EP verbal forms and additional information such as the infinitive of the verb, mood, tense, number and person as the following example shows:

expuseste / expor / cat:VERB / subcat:MAIN / tense:PAST / number:SINGULAR / person:2 / mood:INDICATIVE

The previous example shows that the verbal form *expuseste* (*(you) exposed*), is originated by the verb *expor* (*to expose*), it is in the past tense of the indicative mood, and corresponds to the second singular person.

3.3.1.6 TED talks and respective BP and EP translations

The Brazilian translations of three TED talks were used to test and evaluate the system. The system would run through them and perform the adaptation to EP.

The corresponding Portuguese translations were used as references for evaluation, as well as human adaptations of the Brazilian translations.

The TED talks which were used, were the following:

- TED Talk number 1 (development): ”Al Gore on averting climate crisis”, which consists in 140 sentences;

[3] CETENFolha corpus (http://www.linguateca.pt/cetenfolha)

[4] CETEMPublico corpus (http://www.linguateca.pt/cetempublico)

- TED Talk number 2 (test): "Amy Smith shares simple, lifesaving design", which consists in 146 sentences;
- TED Talk number 7 (test): "David Pogue says "Simplicity sells", which consists in 303 sentences;

Furthermore, for a final experiment, the remainder of the BP translations and respective adaptations to EP were used to train two separate translation systems. These texts consist in 312 translations, with the average length of about 100 sentences.

3.3.2 BP2EP

The starting point for this system is MARv. It runs through the Brazilian text and produces a morphological classification for each word. This is very useful information for the process of building transformation rules from BP to EP, since, as described earlier, most of the main differences depend on whether specific classes of words are present in the sentence.

For example, by using MARv to tag each word in the BP text with its morphologic category and important additional information, handling of clitics, lack of articles before the possessive and gerunds is immediately possible, without the need of using any additional resources. Figure 3.1 shows the arquitecture of the BP2EP system.

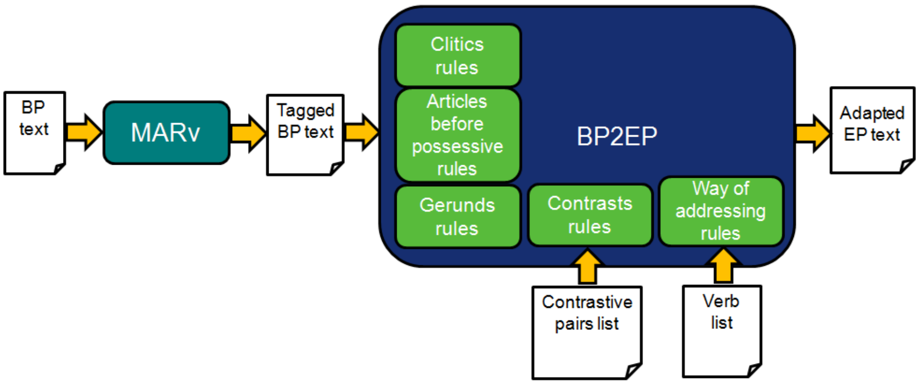


Figure 3.1: System architecture of BP2EP

3.3.2.1 Handling of clitics (third person and pronominal)

In the case of third person clitics, whenever a verb is followed by a third person personal pronoun, depending on the pronoun, a similar rule is applied as shown in the following example.

if pronoun is "ele"

verb + "ele" → verb + "-o"

Since there are only six different variations of third person clitics, there are six different rules to handle these cases, where instead of *ele* (*he*), other pronouns are present and the transformation using the corresponding third person clitic is applied.

As an example, the following BP sentence fragment

dobro **ele** *no meio*

(I fold **it** *in half)*

would become

dobro-**o** *no meio*

If the sentence was instead

dobro **elas** *no meio*

which has the same meaning, only with a plural and feminine pronoun instead of a singular and masculine, the result would be

dobro-**as** *no meio*

With pronominal clitics, if a personal pronominal form appears immediately before a verb, a rule similar to the following is used.

Pronominal form + verb → verb + "-" + pronominal form

As an example, the following BP sentence

E o homem **o** *disse.*

(And the man said **it***)*

would become

E o homem disse-**o**.

Despite solving the majority of the cases, these rules do not cover all variations and possible exceptions when using clitics, such as the construction of future and conditional verbal forms and some phonologic phenomena which cause slight changes in the verb and the inclusion of a consonant at the beginning of a clitic previously starting by a vowel. But these are well defined situations in the Portuguese grammar, which allowed an easy building of specific rules to accommodate exceptions such as these. For example, when applying the rules mentioned earlier to the sentence fragment

quando é necessário tratar **eles**

(when it is necessary to treat **them***)*

the result would be

*quando é necessário tratar***-os**

which is a completely wrong construction.

However, if a new rule would be able to extend the third person clitic one, it would produce the following correct result:

quando é necessário **tratá-los**

Since the verb is in its infinitive form, the new rule, which was added to BP2EP, drops the final *r*, the last syllable is explicitly accentuated and the *l* is added to the clitic for phonetic reasons.

3.3.2.2 Handling of gerunds

Since MARv also provides addition information other than the morphological class of the word, in some cases it proves to be very useful. This is one of such cases. More than simply identifying the word as a verb in its gerund form, it also returns the infinitive of the verb. Therefore, if MARv identifies a gerund, the respective infinitive is also extracted from its output and a rule similar to the following is applied:

gerund → "a " + infinitive

As an example, the following BP sentence

E nenhum de vocês parece estar **reconhecendo***.*

(And none of you seem to be **recognizing***.)*

would become

E nenhum de vocês parece estar **a reconhecer**.

Although building the rule to handle these cases was very simple, they are still very difficult to solve correctly in all situations, because gerund forms are employed in some situations but not always. The following sentence

Caso contrário, será cobrado 100% de perda, **aumentando** *substancialmente a distância.*

(Otherwise, 100% loss will be charged, significantly **increasing** *the distance.)*

fits perfectly both in EP and BP and should not be corrected, but the sentence

Porque não estamos **utilizando** *esta opção?*

(Why are we not **using** *this option?)*

although being correctly written, does not and the rule should be applied.

On the other hand

Porque não estamos **a utilizar** *esta opção?*

has exactly the same meaning but it would be the most usual EP translation.

The previous example shows that, automatically, knowing if the gerund handling rule should be applied or not, is very difficult. But since, in the majority of the cases, the gerund has to be replaced by the infinitive, the decision was to apply the rule to all of the gerunds, but keeping in mind that a minority of sentences that were correct in the first place will become incorrect.

3.3.2.3 Handling of time distances and expressions denoting existence

To handle time distances and expressions of existence, which despite being distinct differences are similar problems and solvable in the same manner, all occurrences of the verbs *ter* and *fazer* and the surrounding contexts were extracted from the CETENFolha corpus.

A preliminary analysis on a randomly selected 1000 results showed that the majority corresponded to uses of the verbs *ter* and *fazer* in situations other than time distances and existence (thus being used in EP as well), and that the ones that did correspond to the desired situations had no syntactic difference to the previous ones, as the following example shows.

In the sentence

O episódio **teve** *repercussão internacional.*

(The episode **had** *international reprecussion.)*

the verb *ter* (*teve*) has a common usage, while in the sentence

No baile **tinha** *muita gente.*

(At the ball, **there were** *many people.)*

the verb *ter* (*tinha*) is used as an expression denoting existence of something (many people, in this case).

Unfortunately, the extracted occurrences exceed 80000 results which are too much to be thoroughly analyzed by hand. On the other hand, if some semantic information would be provided in order to identify time distances and existence expressions, a rule would be very easy to apply by simply replacing the *ter* or *fazer* verb with the corresponding form of the verb *haver*.

3.3.2.4 Handling of the lack of article before the possessive

Handling the lack of an article before the possessive is very simple and performed by rules that detect possessive pronouns and place the corresponding article before them.

possessive pronoun → article + possessive pronoun

As an example, the following BP sentence

E o melhor de tudo, **sua** *motivação.*

(And best of all, **your** *motivation.)*

Would become

E o melhor de tudo, **a sua** *motivação.*

3.3.2.5 Handling of way of addressing

When the word *você* followed by a verb is detected, the sentence is transformed in order to match an EP construction using the pronoun *tu* instead. So, *você* is replaced by *tu* and the verb is changed from the third person singular to the second person form with the help of the verb list. The following rule is applied:

"você" + verb (3rd person) → "tu" + verb (2nd person)

As an example, the following BP sentence

A cada maré que vem e vai, **você encontra** *mais conchas.*

(For each tide that comes and goes, **you find** *more shells.)*

would become

A cada maré que vem e vai, **tu encontras** *mais conchas.*

This case also allows a number of variations, namely an adverb between *você* and the verb (which is handled by incorporating the detection of an adverb in the previous example rule) or the word *você* appearing after a verb or after a verb and a preposition (which leads to a similar but reversed rule where grammatical contraction rules will be applied to the verb, the pronoun *tu* and the preposition), as the following examples, respectively, show.

Você já fez *algo parecido*

You have already done *something similar*

becomes

Tu já fizeste *algo parecido*

and

Eles deixaram **você***.*

(They let **you***.)*

becomes

Eles deixaram-**te***.*

Just like the handling of gerunds, these situations are also very difficult to solve and forced a difficult decision, because in Portugal the use of *tu* (*you*) and *você* (*you*, more formal) to address someone happens in distinct situations and many times can be omitted from the sentence without affecting its meaning. In Brazil, *você* is used in the majority of situations and is almost never omitted.

For example, the sentence

Você deseja *receber pacotes úteis.*

*(***You wish** *to receive useful packages.)*

probably would not be written in Portugal.

On the other hand, the sentence

Tu desejas *receber pacotes úteis.*

would be perfectly legal in EP, but addressing the listener/reader as *Tu* would be disrespectful. Omitting the *Tu* or *Você* would not ruin the meaning of the sentence, but it would make it sound strange and unusual.

So, the decision was to change *você* to *tu* in all situations where it is followed by a verb, and to perform the corresponding correction to the verb, which allowed to obtain a correct EP sentence, even if in context or given the scope of the text, a Portuguese person would not have written it that way.

3.3.2.6 Handling of lexical differences

To solve the issue of lexical differences, the first resource used is the contrastive pairs list. However, it has been constructed in the EP-BP direction (the opposite of this work's objective), and many of the contrasts are not valid in both directions, introducing many errors.

For example, the contrast *sítio::lugar* (*place*) is useless, as in EP they are synonyms that are both used in written and spoken language. Therefore, transforming every occurrence of *lugar* into *sítio* would make no sense. As another example, let us consider the contrast *meia de leite::média* (*a cup of milk*). Probably, in Brazil, a *meia-de-leite* is always called a *média*, but reversing the contrast would imply that in Portugal, the word *média* always refers to a cup of milk, which does not happen. In fact, in EP, *média* never refers to a cup of milk, it means medium, or average.

A cleaning has been performed, by hand and by removing such occurrences, before the development of this system and while it was tested, but some errors at this level are still expected. This filtering reduced the original list (see Section 3.3.1.4) to about 2200 entries.

This list contains lexical differences of both kinds described in Section 3.2.8, but to complement it with more spelling contrasts, all of the words contained in CETEMPublico which may have a spelling difference to their BP counterparts were extracted. These words, which are possible spelling contrasts, were extracted based on (Wittmann et al., 1995), resulting in a further 10000 extra results to be used as a complement to the contrast list.

Spelling diff.	EP	BP	EN
cc::c	accionar	acionar	to activate
cç::ç	acção	ação	action
ct::t	recto	reto	straight
pc::c	excepcional	excecional	exceptional
pç::ç	adopção	adoção	adoption
pt::t	excepto	exceto	except
bd::d	súbdito	súdito	subject
bt::t	subtil	sutil	subtle
mn::n	amnistia	anistia	amnesty
ém::êm	prémio	prêmio	prize
én::ên	génio	gênio	genius
óm::ôm	económico	econômico	economic
ón::ôn	sónico	sônico	sonic
ei::éi	ideia	idéia	idea

Table 3.1: Common spelling differences between EP and BP

Table 3.1 shows examples of each type of common spelling differences between EP and BP.

Therefore, for each word in the BP text, if there is any word in the contrast list which matches any possible spelling difference between BP and EP, a substitution is performed.

For example:

idéia becomes *ideia*,

eletricidade becomes *electricidade*,

econômico becomes *económico* or

adoptar becomes *adotar*

3.4 Evaluation

Initially, the resources used to perform the evaluation of this work were the Brazilian translations of the three TED talks mentioned in Section 3.3.1.6.

These Brazilian texts were given as input to the adaptation system and the resulting texts were evaluated using two methods: one objective and one subjective.

Furthermore, a final experiment was performed with the goal of making a comparison of translation scores using original BP resources and the same resources adapted to EP to train a system. In this case, the training of the system was done with the remainder of the available TED talks' transcriptions.

3.4.1 Objective evaluation

For the objective evaluation, the standard translation evaluation measure, BLEU, was used. At first, and since both BP and EP translations were already available, the approach was to use BLEU to evaluate the ”translation” quality of all of the output text by ranking it against the EP translation as a reference.

Table 3.2 shows that the results were very poor, considering that the EP reference and the adapted text were translations of the same talk, and much better results were to be expected. However, in this case, it may have no relation to the quality of the adaptation. If fact, after inspecting both EP and BP translations it became obvious that these results were so poor due to the fact that, despite being Portuguese translations of the same TED talk, these texts were created by different people who used very different words, sometimes synonyms, expressions and even syntactic discrepancies, which BLEU is unable to capture.

TED talk	BLEU
1	35.99
2	16.52
7	17.41

Table 3.2: Results of the first evaluation using BLEU

For instance, for the original English transcription

It addresses one of the biggest health issues on the planet.

the BR translation is

Ele enfoca uma das maiores questões na área de saúde no planeta.

and the Portuguese translation is

E aborda um dos maiores problemas de saúde pública no planeta.

The translations are very different, but still, they are both correct in each variety.

As the previous approach proved to be inadequate to perform a consistent evaluation, the preferred approach was to ask an independent human translator to transform the original BP translations into EP, by making the least amount of changes possible, in the hope of obtaining an EP reference which would be closer to the original BP text. Finally, the original and the adapted texts were evaluated against the new EP references in order to assess how much quality was gained (or lost) by performing the adaptation.

TED talk	BLEU (original)	BLEU (after adaptation)
1	60.70	89.78
2	57.42	86.26
7	84.09	90.60

Table 3.3: Results of the final objective evaluation using BLEU

Table 3.3 shows the results of the objective evaluation method using BLEU.

These results are much closer to what was expected in the beginning, in part because both the adapted texts and the references were created from the same source.

The results obtained with the original texts (prior to adaptation), also show that, despite being the same language and relatively similar varieties, BP and EP differences are significant to the point that to a Portuguese speaking person, only slightly over half of the text would be common for both varieties. TED talk number 7, would be an exception to these results since many of its sentences are very short, in some cases with the length of only one word, thus being correct in both varieties.

Also the high gain in BLEU from the original to the adapted texts shows that the rules that the adaptation system uses, are in its majority, the same that a human translator would implicitly use while adapting a BP text to EP. Even with TED talk number 7, the gain is significant, since adding to the short sentences that were already common for EP, the adaptation system was able to transform many differences in the longer sentences, leading to a 6.5 point gain and the higher of the three results.

However, despite the fact that the adaptation system was able to obtain BLEU scores of almost 90, which shows that it correctly adapted the majority of the three test texts, the results shows that it still fails in some situations. These cases, which are not covered by the rules, depend on the context of the sentence or even of the whole text and its subject.

For instance, one of these cases may be the difficulty of deciding on whether to keep a gerund or to perform the adaptation and risk spoiling the sentence, as described in Section 3.3.2.2. Other situation may be the fact that some words which are correct in both varieties are only correct in certain contexts and with certain meanings. For example, in the following sentence fragment

tu terás uma questão **legal**

(you will have a **legal** *matter)*

the word *legal* has the same meaning in both varieties (related to law). But in the sentence

Um dispositivo muito inteligente, **legal** *e elegante que saiu recentemente.*

(A very intelligent, **cool** *and elegant device that came out recently.)*

the word *legal* is used as an adjective that denotes a likeable quality of the device, and it is only used in BP.

Without any context, there is no way for the adaptation system to know if a given occurrence of the word *legal* is used in a context of law (which is valid for both varieties) or in a context of a likeable characteristic of something or someone (which should be adapted to EP, and in this particular case, was adapted by the human translator).

3.4.2 Subjective Evaluation

For the subjective method, a set of a hundred sentences was selected from the input BR texts and the original and adapted EP texts. Fifty were randomly extracted from both the original BP and EP translations of the TED talks, while the other fifty were extracted from the adapted after being the target of one or more of the changes described in Section 3.3. Table 3.4 summarizes the composition of this set of sentences, which does not intend to cover every single adaptation that was perfomed, since this evaluation method was designed to be applied with human evaluators.

Type of sentence	Frequency
Original BP	25
Original EP	25
3rd person clitic	1
Pronominal clitic	7
Gerund	5
Article before possessive	4
Way of addressing	6
Lexical difference	6
Addressing and lexical difference	2
Addressing and pronominal clitic	3
Addressing and gerund	2
Gerund and lexical difference	3
Gerund and pronominal clitic	2
Article and 3rd person clitic	1
Article and lexical difference	1
Pronominal clitic and lexical difference	4
Addressing, article and 3rd person clitic	1
Gerund, article and lexical difference	2

Table 3.4: Composition of the set of evaluation sentences

This set of sentences was then given to eight people who were asked to categorize each sentence as either BP or EP. With this type of evaluation it was hoped to assess how well the changes performed by the adaptation system reflect EP orthographic, syntactic and semantic constructions, i.e., if someone would correctly identify sentences from the original EP translation as EP and sentences from the adapted text as BP, then the adaptation had not been successful. With this method, it was also hoped to point out which type of changes performed better or worse, and in which cases the adaptation rules were insufficient.

Table 3.5 shows the average percentage of correctly identified sentences over all eight evaluators.

Type of sentence	Correct (%)
Original BP	83.5
Original EP	88
3rd person clitic	87.5
Pronominal clitic	76.5
Gerund	92.5
Article before possessive	87.5
Way of addressing	87.5
Lexical difference	62.5
Addressing and lexical difference	100
Addressing and pronominal clitic	87.5
Addressing and gerund	81.3
Gerund and lexical difference	91.7
Gerund and pronominal clitic	43.8
Article and 3rd person clitic	87.5
Article and lexical difference	12.5
Pronominal clitic and lexical difference	75
Addressing, article and 3rd person clitic	100
Gerund, article and lexical difference	62.5

Table 3.5: Results of the evaluation using the subjective method

The results show that both the sets of original EP and BP sentences, which had not been adapted, and were chosen deliberately to be easily identified, were not correctly labeled in 100% of the cases as expected. This may mean that the sentences were not so distinguishable as previously thought, or that, perhaps, one can expect that about 15% of the adaptations will not be noticed in the first place.

By analyzing Table 3.5, it is also possible to see that some types of adaptations performed much better than others. When the system adapted addressing and gerunds, gerunds and pronominal clitics or articles and lexical differences in the same sentence, less than half of the evaluators, in average, identified the sentence as still being written in BP, which is an indication that in these cases, the adaptation system may need further reviewing.

On the other hand, when handling gerunds, addressing and lexical differences, or gerunds and lexical differences, more than 90% of the times, the evaluators identified the sentences as being EP, which shows that in such situations, the adaptation system produces an output with enough quality to be included in a training set for a EP-English MT system.

When calculating a weighed average over all types of adaptations performed, the system produces 78.96% of correctly adapted EP sentences, and for the majority of these types of adaptations, the results show that the rules used by the system can cover over 75% of the tested sentences for each type of adaptation performed.

After inspecting the adapted sentences that were incorrectly labeled by the majority of the evaluators, it becomes clear that the reason for the failure in identifying them as EP is not the fact that the adaptation system had done something wrong. Instead, the failure is due to the same type of errors already discussed in the objective evaluation method, which causes any other correct adaptation present in the sentence to be overlooked by the evaluators.

For instance, the sentence

Essa filosofia de fazer as coisas **do jeito certo** *está a começar a espalhar-se.*

(This philosophy of doing things **the right way** *is beginning to spread.)*

is the result of a change in a pronominal clitic (*se espalhar* to *espalhar-se*) and a change in a gerund (*começando* to *a começar*).

However, 87.5% of the evaluators labeled it as being BP, simply because the expression *do jeito certo* (*the right way*) is very typical for BP, and in EP, it would not be used. Instead, the expression could be replaced by *da forma certa* or *da maneira certa*, and then it would probably be identified as EP.

Just like with the objective evaluation, the adaptation system, without context, cannot know if it should replace the expression containing the word *jeito*, or if it should keep it, as in other contexts, *jeito* may mean *aptness* both in BP and EP.

There were five other sentences in which half or more of the evaluators still labeled them as BP after being adapted, which is a strong indication that, in fact, there is something wrong with them.

But, again, after analyzing each of these five sentences, it was possible to identify what caused evaluators to classify them as BP after having been adapted. The reason was the same as in the previous example: in all of the sentences, the performed adaptations had been successful but the system could not identify that there was an expression that, in that context, also had to be changed in order for the sentence to be written in EP.

This also happens in many other adapted sentences which were labeled as BP, although not as frequently, showing how commonly this type of error occurs and how important it is to be able to solve it, in order to produce better adaptations. If these situations were correctly handled, many of these sentences would be correctly identified as EP and many correct adaptations would not be rendered useless, thus increasing system performance.

These situations, which with this evaluation method were easily pointed out, seem to confirm what was already stated in Section 3.4.1, i.e., that the system fails to adapt the expression whenever the change depends on the context. This evaluation also allows to show that not performing these changes may easily prevent a correctly adapted sentence (in several other BP-EP differences) to be identified as EP, thus not being suitable to train EP translation systems.

3.4.3 Final adaptation experiment

This final experiment had the goal of trying to validate the results of the adaptation achieved by BP2EP in a translation scenario. This was perfomed by using the original BP texts to train a translation system and comparing the results with the ones produced by the same system, only with the same BP texts adapted to EP.

The translation systems were trained with the statistical phrase-based decoder Moses and the TMs were trained on the English-Portuguese part of the Europarl corpus. The LMs were trained on the remaining available TED talk transcriptions. Word alignments have been extracted by GIZA++ and the LMs have been generated by SRILM. Translation results have been evaluated, for the three TED talks mentioned in Section 3.3.1.6, with BLEU against two references.

Table 3.6 shows the results achieved by both system setups, tested on three talks, the first system using the original BP texts to train the LM and the second using a LM trained after running the adaptation process over all the original texts.

TED talk	BLEU (original)	BLEU (after adaptation)
1	19.04	19.12
2	17.96	18.04
7	21.34	21.46

Table 3.6: Results of the final adaptation experiment

These results show a slight, but probably insignificant, improvement when using the adapted LM. This is caused by the fact that the nature of the training corpus prevented its use for the TM, since the original EN transcriptions and the BP translation were unaligned, and performing so manually would be unmanageable. Thus, using only a small adapted LM does not provide sufficient information to state that the adaptation was successful or if the improvement was mere coincidence. Had the TM been also adapted, although much smaller than the one trained with Europarl and used in this experiment, and probably the difference would have been more significant, and would have allowed the drawing of some conclusions regarding the quality of the adaptation in a real translation scenario.

4 Error analysis for speech-to-speech machine translation

Evaluation of Machine Translation is not a trivial task. Generally, an automatic evaluation approach is preferred over human evaluations, which are much more costly both in resources and time.

Several automatic evaluation metrics have been proposed and, today, are widely used to score the quality of a translation output based on a comparison with a set of translation references. These methods such as WER, PER, BLEU, NIST and METEOR, although very easy and quick to use, only produce a numeric value, which can be very useful to monitor the progression of a MT system throughout time. However, such automatic metrics cannot produce a comprehensive evaluation of an MT system and namely, cannot capture the relation between the score and the errors present in the MT output that caused the score not to be perfect.

In the specific case of a SLT system, besides the MT engine, there is another component, the ASR, which may also be the source of translation errors and indirectly influence automatic scores. Due to the highly disfluent nature of speech, different speech styles, different voices or accents, these components, by not being able to perfectly adapt to these speech conditions, will easily recognize some sentences incorrectly, and propagate these errors to the MT engine.

Thus, exploring the nature of MT errors may play a very important role in understanding how they influence translation quality and how a MT system can be improved.

This chapter describes the work carried out by analyzing the errors produced by a Portuguese-English SLT system in the BN domain and by an English-Portuguese system for a talk domain, and identifying their influence in the overall translation performance.

Section 4.1 presents a quick overview of related work on the subject of translation error analysis, Section 4.2 summarizes several experiments in translation under different conditions and their respective results, Section 4.3 describes the types of errors encountered when performing the early experiments and Section 4.4 presents and discusses the results when studying how each type of errors influences BLEU scores.

4.1 Background in translation error analysis

Some work has been developed with the goal of identifying how translation errors influence automatic evaluation results. Within these works, there are automatic error analysis methods, human error analysis frameworks and some more linguistically oriented approaches.

(Popovic et al., 2006) proposes an automatic error analysis method, where morpho-syntactic information is used to perform such evaluation. This approach explores the combination of POS-tagging with automatic metrics such as WER and PER to propose a new, error oriented, evaluation metric.

It explores the main differences between Spanish and English, mainly syntactic differences considering nouns and adjectives and the highly inflectional nature of Spanish.

(Elliott et al., 2004) also states that building an automated error recognizer would allow a quicker evaluation process by eliminating the expense of creating translation references. This method relies on a POS-based error classification scheme to detect which classes of words are more likely to insert errors in the MT output, and how these errors are represented in terms of fluency and adequacy. This evaluation system also incorporates error annotations in order to allow post-editing for error detection and correction.

On the other hand, (Vilar et al., 2006) proposes an error analysis and classification scheme for human evaluation. Following a hierarchical classification, where top-level error categories are missing words, word order, incorrect words, unknown words and punctuation, human evaluators identified errors present in the MT output, by comparing them with translation references, and categorized them following the proposed classification scheme.

Error statistics were collected for English-Spanish, Spanish-English and Chinese-English translation, under different conditions such as clean and corrected texts, verbatim texts and ASR generated texts.

(Condon et al., 2010) developed similar work, but with translations to and from English and Iraqi Arabic. Errors were annotated both as deletions, insertions and substitutions for morphological classes and types of errors following a similar taxonomy as the one proposed in (Vilar et al., 2006).

A statistical study of the annotated errors was performed in two distinct points within the development of the translation system and showed that as the system matured, and BLEU scores became better, the number of errors decreased, while still maintaining relative proportions.

(Secara, 2005) also presents a survey on state-of-the-art translation evaluation methods but on a much more linguistically oriented, where the focus of most of the analyzed frameworks is on annotation schemes and error weighing for assessing the quality of a translated text, and on including post-editing feedback from human experts in error reductions and translation improvements.

4.2 Early experiments

As a starting point for error analysis, baseline systems for both the BN and the talk tasks have been setup and evaluated under several different conditions.

4.2.1 Corpora and resources

One of the corpus used in this work consists in the widely popular Europarl multilingual corpus, which is employed to train the baseline TM and LM, as well as to tune feature weights. LMs have also been trained based on the interpolation of texts from the newspaper Público and WPT[1].

Other resources that have also been used are TED talks and their respective English transcriptions and Portuguese translations. The TED website provides text transcriptions of all the available talks and translations in several languages. There are, currently, 172 European Portuguese translations.

In order to perform translation experiments and evaluate them, it was necessary to create translation references. Thus, starting from a manually corrected and annotated transcription of the June 1st 2008 RTP1 Telejornal, two reference translations were produced. One translation was manually created, while the other one was automatically produced using Google's translation tools[2] and manually corrected in order to speed up the reference creation process and still obtaining a reference relatively close to a human standard.

A similar process was performed in order to obtain references for the talk task, but in this case it was not necessary to manually create a reference, since a Portuguese translation was already available at the TED website. In this case, the experiments were performed over the talk from Barry Schwartz called *Barry Schwartz on our loss of wisdom*.

4.2.2 Experimental settings

In this work, as previously stated, two different applications were performed: a) a BN translation setup and b) an oral presentation (talk) translation setup.

The baseline setups consisted in the ASR engine AUDIMUS (Meinedo et al., 2003) and the statistical phrase-based decoder Moses in a traditional uncoupled architecture. The TM and LM have been trained on the English-Portuguese part of the Europarl corpus and the respective feature weights have been optimized on a held-out development set. Word alignments have been extracted by GIZA++ and the

[1] WPT (http://xldb.fc.ul.pt/wiki/WPT_05)

[2] Google Translate (http://translate.google.com/)

LM has been generated by SRILM. Translation results have been evaluated with the BLEU metric against the two available references. Figure 4.1 illustrates the baseline system architecture.

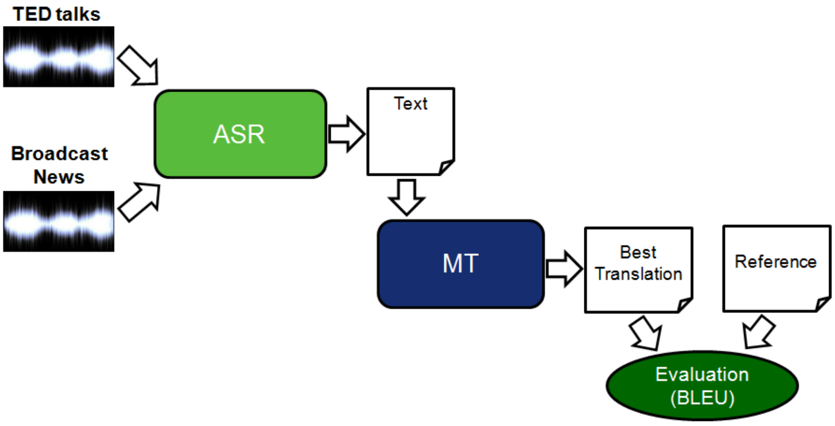


Figure 4.1: Baseline system architecture

4.2.2.1 Broadcast News setup

For the BN setup, Moses' output has been evaluated in different situations, such as when performing translation of the ASR output and when performing translation using the manually corrected transcription. In both situations, and since the original transcription files were annotated in a way that easily allowed it, the references were split into planned and spontaneous speech in order to draw conclusions about the influence of spontaneous speech and its inherent lack of structure and correctness in the overall score. As expected, when no ASR was involved, scores were much higher, with differences up to 16 points, and spontaneous speech scores were lower than the overall ones, while planned speech obtained higher scores. Table 4.1, Table 4.2 and Table 4.3 summarize the results obtained in these experiments.

Whole speech	BLEU
With ASR	26.64
Without ASR	38.38

Table 4.1: BLEU Results of BN experiments, with and without ASR, for the whole test text

Spontaneous speech	BLEU
With ASR	18.20
Without ASR	36.55

Table 4.2: BLEU Results of BN experiments, with and without ASR, only for the spontaneous part of the test text

Taking into account the fact that the test text was the transcription of the BN from Children's Day, thus featuring several small children whose voice features have not been used to train ASR acoustic

Planned speech	BLEU
With ASR	30.01
Without ASR	46.92

Table 4.3: BLEU Results of BN experiments, with and without ASR, only for the planned part of the test text

models and whose speech is even more unstructured than an adult's, it may have contributed to more severe drops in the overall score than the ones observed if the test text had been from another date.

This specific test corpus also contained more live reports than usual which may have been another decisive factor in lowering overall scores. This seems to be confirmed by the fact that overall scores are more influenced by spontaneous speech than planned speech, since spontaneous speech scores are much closer to the overall ones.

A posterior step saw the introduction of a new module in the processing chain: an automatic punctuation and capitalization recovery module for ASR (Batista et al., 2008). This module runs through the ASR output, which has no punctuation or capitalization, and enriches it, producing a new output with punctuation marks and capitalized words prior to the translation step. Figure 4.2 illustrates how this new module interacts with the rest of the architecture.

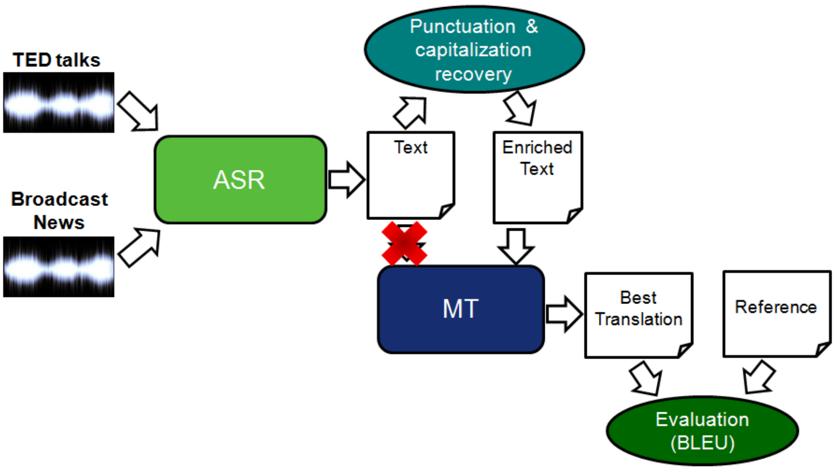


Figure 4.2: System architecture with punctuation and capitalization module

Experiments were performed with a MT component in which the lowercasing of the training corpus was skipped, resulting in capitalized models, where this module was used to insert capitalization and punctuation prior to translation. With these settings, the overall system was evaluated under two different scenarios: the first where the text segmentation of the ASR output, after being enriched, was kept (roughly one sentence per line, but not necessarily so), simulating that the translation was being performed offline, and the second where the enriched ASR output was segmented in every punctuation mark, simulating a real-time situation where a segment is immediately translated whenever the speaker makes a pause.

Taking advantage of this tool's flexibility, the system was yet again tested under different conditions, but in this case, using the standard training procedure of lowercasing the training corpus, such as inserting only punctuation, both punctuation and capitalization and punctuation and capitalization of only potential named entities (NE's), i.e. capitalized words that do not follow a full stop, while the remaining capitalizations are recovered after translation.

Table 4.4 and Table 4.5 show the BLEU results achieved in these experiments.

Experimental conditions	BLEU
Baseline	26.64
Offline segmentation	17.89
Online segmentation	19.95
Punctuation and capitalization of named entities	25.43

Table 4.4: BLEU Results of BN experiments using capitalized models and the punctuation and capitalization module

Experimental conditions	BLEU
Baseline	26.64
Only punctuation	28.5
Punctuation and capitalization of named entities	29.18

Table 4.5: BLEU Results of BN experiments using lowercased models, offline segmentation and the punctuation and capitalization module

Using capitalized models resulted in a severe drop in translation quality, since although there are alignments that contain capitalized words, for almost every word in the beginning of a test sentence, Moses cannot find a translation candidate, and the word is copied to the output as if it was an unknown word, when in reality, if lowercased it would be translated.

These results show that, despite having a capitalization recovery module prior to translation, capitalization should not be performed before translation since the TM will not be able to handle capitalized words, not even if it has been trained with such words.

Results were largely improved by keeping only the capitalized words that do not start a sentence, as those, probably are named entities and do not require translation. These results show that the words that probably require capitalization before translation are only the ones that do not start a sentence, as those, since punctuation is kept after translation, are easily recased simply because they will follow a full stop. Combining them with the previous results, it is possible to draw the conclusion that before translation, only punctuation and capitalized words that do not start a sentence are useful.

By simulating an online scenario, results are increased by 2 points, which gives a good indication of how smaller segments with local contexts and low reordering costs may help in improving overall translation quality. Translation time was also reduced relatively to the longer, original sentences, but this experiment was only a simulation and translation times were not actually measured in order to understand if the system would indeed perform in a real-time manner.

With the standard lowercased models and recasing after translation, results were improved over the original baseline scores, now showing that without the losses caused by the capitalized models, punctuation helps the translation in being closer to the capitalized and punctuated references and that adding only the capitalization in potential named entities, the BLEU score is even better. It is expectable that if the text had a larger number of possible named entities, the gain would have been larger.

4.2.2.2 Talk setup

This setup was similar to the previous one, only in the opposite direction and without a clear distinction between planned and spontaneous speech. Thus, baseline experiments were performed with standard lowercase models and consisted only in evaluating the system when the ASR output is translated and when the translation is performed over a corrected transcription.

For this talk, in particular, speaker adaptation has been performed with a decrease of WER from 14.835 to 14.086, and the language model was created from an interpolation of the Público corpus and WPT instead of the usual target side of Europarl.

Again, and as expected, a large improvement is visible when no ASR is involved in the process as Table 4.6 shows.

Whole speech	BLEU
With ASR	18.33
Without ASR	33.66

Table 4.6: BLEU Results of TED talks experiments, with and without ASR, for the whole test text

These results are lower than the ones produced by the BN system and are closer to the ones obtained with spontaneous speech rather than planned speech. This may be caused, in part, by the fact that in oral presentations, while not being completely spontaneous because, probably, the speaker has practiced a lot before the presentation and is following a written script, speech is still subject to hesitations, corrections and a certain degree of improvisation and is far from being similar to speech being read in news reports and by news anchors.

This particular TED talk is about a very specific domain: practical wisdom and morality, which is not very close to the subjects present in Europarl and used to train the system. On the other hand, the BN domain is very broad and encompasses several other smaller ones, but the majority of the stories are related to politics or economy, which makes them much closer to the topics addressed in Europarl than a very specific talk about a very specific subject. Therefore, the distance between the test domain and the training domain may also explain why BN results, in absolute terms, are higher than the ones from the talk setup.

Drawing upon the experience from the BN system, in this case, the evaluation jumped directly to the setting in which the automatic punctuation and capitalization module is used to insert punctuation and capitalization only in words that do not start sentences, with the latter words being recased after translation. Table 4.7 shows that, again, this setting results in a relevant improvement in BLEU.

Experimental conditions	BLEU
Baseline	18.33
Punctuation and capitalization of named entities	20.95

Table 4.7: BLEU Results of TED talks experiments using lowercased models, offline segmentation and the punctuation and capitalization module

4.3 Types of errors

In order to understand what kind of errors were present in a translation, the translations produced by the early experiments already described, were analyzed and searched for errors, for both the EN-PT and PT-EN directions and for TED talks and BN, respectively.

The errors found in these translations were then compared to the taxonomy proposed by (Vilar et al., 2006), with the goal of finding which types of errors were in fact present in the analyzed translations. This led to a simplified version of this taxonomy, which is presented on Figure 4.3.

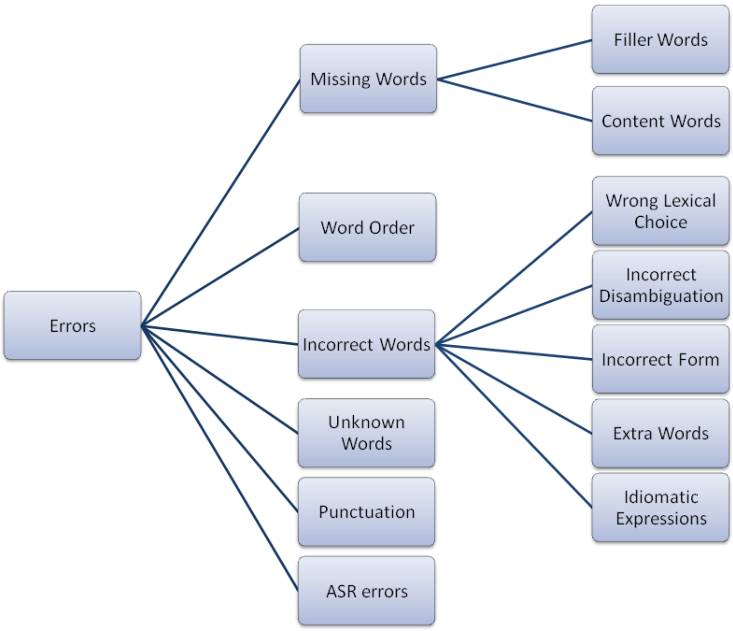


Figure 4.3: Simplified error taxonomy for error classification used in this work

4.3.1 Missing words

When one or more words are missing in the translation, they can either be classified as missing filler words or missing content words.

4.3.1.1 Missing filler words

This type of error occurs when the translation is missing a word, usually a preposition, a conjunction or an article, that does not affect the meaning of the sentence.

Example:

PT-EN

PT: *também tem direito a alguns minutos* **de** *glória*

EN: *also has the right to a few minutes glory*

Correct translation: *also has the right to a few minutes* **of** *glory*

EN-PT

EN: *the war against moral competence*

PT: *a guerra contra competência moral*

Correct translation: *a guerra contra* **a** *competência moral*

4.3.1.2 Missing content words

This type of error occurs when the translation is missing a word which is vital to the meaning of the sentence. Usually, these words are verbs, pronouns, nouns, adjectives, etc.

Example:

PT-EN

PT: *são as maiores empresas nacionais da* **indústria**

EN: *are the largest national companies of the*

Correct translation: *are the largest national companies of the* **industry**

EN-PT

EN: *virtue is an* **old-fashioned** *word*

PT: *virtude é uma palavra*

Correct translation: *virtude é uma palavra* **antiquada**

A particular case of this error, for the PT-EN language pair, is that in Portuguese, the subject of a sentence may be omitted while in English it must be always present. Therefore, in the PT-EN direction, a noun or a pronoun may be omitted from the source PT sentence (without it being incorrect), causing the EN translation to be missing the subject.

Example:

PT-EN

PT: *esperam que sejam revistas*

EN: *expect that should be reviewed*

Correct translation: *expect that* **they** *should be reviewed*

4.3.2 Word order

These errors happen when the reordering model is unable to perform a reordering, causing the translated sentence to look and sound strange and unusual. Although (Vilar et al., 2006) distinguishes word and phrase order, for both short and long ranges, the analyzed texts only show ordering errors for single words and for short ranges.

In the particular case of the PT-EN language pair, these errors are mainly due to the fact that in English, the adjective is always placed before the noun, while in Portuguese, it is most common to place the adjective after the noun (although a syntax similar to English is still possible).

Example:

PT-EN

PT: *processo completamente certificado*

EN: *process completely certificate*

Correct translation: *completely certificate process*

EN-PT

EN: *have the moral ability*

PT: *tem a moral habilidade*

Correct translation: *tem a habilidade moral*

4.3.3 Incorrect words

This type of error occurs when the translation engine is unable to correctly translate a word or expression, producing instead, a wrong translation that severely affects the intended meaning of the sentence.

4.3.3.1 Lexical choice

In this case, the translation engine chooses the wrong translation candidate and produces a sentence that probably makes no sense.

Example:

PT-EN

PT: *em situação presencial*

EN: *the situation is expected in health*

Correct translation: *in a face-to-face situation*

EN-PT

EN: *this would clearly be an absurd*

PT: *é absurdo sobre o seu rosto* (*it is absurd over its face*)

Correct translation: *isto seria claramente um absurdo*

4.3.3.2 Disambiguation

In some situations, when the source language word may have several meanings on the target side, the chosen translation candidate may not be the correct one among all alternative meanings.

For instance, in Portuguese, the verb *to be* has two meanings: *ser* or *estar*, where *ser* refers to a more permanent characteristic of the subject, as in *to be smart* or *to be red*, while *estar* has a more temporary meaning as in *to be at home*. As another example, there is the verb *to play*, that in Portuguese also has several meanings, as in *to play a game* (*jogar*) or *to play guitar* (*tocar*).

Example:

PT-EN

PT: *o jogador foi* **emprestado**

EN: *the player was* **borrowed**

Correct translation: *the player was* **loaned**

EN-PT

EN: *here* **is** *an example*

PT: *aqui* **é** *um exemplo*

Correct translation: *aqui* **está** *um exemplo*

4.3.3.3 Incorrect forms

These errors occur when the translation engine, despite producing words with the correct lemma, fails to produce the correct form of the word, usually by incorrectly translating a verb form or by being unable to maintain gender or number consistency in nouns, adjectives, pronouns, etc throughout the sentence.

This is one of the most common errors when translating from EN to PT due to the highly inflectional nature of Portuguese. For instance, the verb *to* do in the past simple tense, only has one form (*did*), while in Portuguese, the corresponding verb *fazer*, has six different forms. There is also the case of the article *the*, which in English can be used for both genders and both numbers, while in Portuguese there are four different forms that, if incorrectly translated, may cause an inconsistency between the article and the corresponding noun.

For the PT-EN direction, these errors usually happen when translating a gerund form, since in Portuguese, the gerund is often replaced by a construction with the word *a* (which can be translated to *to* or *the*) followed by the infinitive. Thus, the English translation will have the infinitive form of the verb, occasionally preceeded by a filler word corresponding to the word *a*, instead of the gerund.

Example:

PT-EN

PT: *a crise está* **a afectar** *os rendimentos*

EN: *the crisis is* **affect** *the income*

Correct translation: *the crisis is* **affecting** *the income*

EN-PT

EN: *to say* **the same** *words*

PT: *dizer* **o mesmo** *palavras*

Correct translation: *dizer* **as mesmas** *palavras*

EN: *some psychologists* **interviewed** *janitors*

PT: *alguns psicólogos* **entrevistados** *auxiliares*

Correct translation: *alguns psicólogos* **entrevistaram** *auxiliares*

4.3.3.4 Extra words

These errors happen when the generated translation contains words, usually fillers such as prepositions or conjunctions, that should not be present, and must be simply removed to obtain a correct sentence in the target language.

Example:

PT-EN

PT: *também dentro de alguns anos*

EN: *also in* **that** *a few years*

Correct translation: *also in a few years*

EN-PT

EN: *I would like to tell a little story*

PT: *eu gostaria de contar* **que** *uma pequena história*

Correct translation: *eu gostaria de contar que uma pequena história*

4.3.3.5 Idiomatic expressions

This category of errors encompasses any kind of expression that does not fit in the above mentioned categories, and that intends to express a meaning other than its literal one. Thus, in these case, if the expression is translated as any other, the generated translation will express its literal meaning rather than its common understanding.

Example:

PT-EN

PT: *é claro*

EN: *it is clear*

Correct translation: *of course*

EN-PT

EN: *what the hell does it mean*

PT: *o que o inferno significa*

Correct translation: *o que diabo significa*

There are also some particular situations where a literal translation cannot be applied such as with hours, years, ages or nationalities.

In English, hours range from 1 to 12 AM and 1 to 12 PM, while in Portuguese, hours may be expressed from 1 to 24 (or 0 to 23).

Example:

PT: *começou às* **vinte**

EN: *it started at* **twenty**

Correct translation: *it started at* **eight PM**

In Portuguese, years, with four digits, are expressed as the literal number, while in English, a year is read in two groups of two digits.

Example:

EN: *nineteen eighty-four*

PT: *dezanove oitenta e quatro*

Correct translation: *mil novecentos e oitenta e quatro*

In English, when expressing an age, the verb *to be* is used, while in Portuguese, the verb *to have* (*ter*) is used instead.

Example:

PT: **tem** *cinco anos*

EN: **has** *five years*

Correct translation: **is** *five (years old)*

Finally, in English, nationalities must be capitalized, while in Portuguese they do not.

Example:

PT: *americana*

EN: *american*

Correct translation: *American*

4.3.4 Unknown words

Unknown words or expressions are the ones for which the translation engine cannot find any translation candidate and therefore are simply copied to the translation output.

This may be caused either by the fact that the word was not present in the training corpus, probably by being part of a very specific domain, or that although present, it is only in the extracted alignments as part of a phrase and not the word itself.

Example:

PT-EN

PT: *em outra* **varanda**

EN: *on another* **varanda**

Correct translation: *on another* **balcony**

EN-PT

EN: scrolling *on another screen*

PT: scrolling *noutro ecrã*

Correct translation: a rolar *noutro ecrã*

4.3.5 ASR errors

These errors are predominantly present when recognizing spontaneous speech, and mainly consist of substitution errors, when a word is incorrectly recognized as another one, probably phonetically similar, and insertions errors when extra words and garbage is present in the recognized output.

When words are incorrectly recognized as something phonetically similar, the word is translated in isolation which causes the sentence to lose its meaning, having a similar effect to missing content words. Particularly, if the word is a verb and the ASR recognizes a different inflected form, it originates errors as if the MT had made an inflection error (although the error comes from the ASR).

In any case, the translation is performed almost word by word, which means that the ungrammaticality of the input sentence is propagated to the translation sentence.

Example:

PT-EN

PT: *que ele devesse ser vinte comprimento aterragem*

EN: *who he should be twenty length landing*

4.3.6 Punctuation and capitalization

These errors may be caused when the capitalization and punctuation module fails, or in certain specific situations that derive from the experiment settings and may require additional attention.

Example:

PT: *longos períodos. em tratamento*

EN: *long periods. in treatment*

Correct translation: *long periods in treatment*

The previous example is a case where the capitalization and punctuation module failed, by inserting a punctuation mark that will cause the following word to be capitalized, in a post processing step, when it should not.

Another situation where capitalization may insert some errors in the translation is with the experimental settings described in Section 4.2.2.1 and Section 4.2.2.2, where whenever a word is capitalized but does not follow a full stop, the capitalization is kept and the word is treated as if it was a named entity. Since the system is trained with a lowercased corpus, this word is treated as an unknown word and copied to the output.

This solution works well for names of people and some institutions, but for nationalities and names of places that must be translated, it fails. For instance, the name *Manuela Ferreira Leite* will be kept in the translation but, *Espanha* will be translated to *Espanha* instead of *Spain*.

On the other hand, if recasing was to be performed for the entire text only after translation, *espanha* would be translated to *spain* (and possibly recognized as a named entity and recased to *Spain* afterwards), but *manuela ferreira leite* would be *manuela ferreira milk*, and probably the word *milk* would not be a named entity. The same would happen for every name that, when lowercased could be translated as common noun, such as *porto* (*Oporto*) which would be *port* or *madeira* (the island) that would be translated to *wood*.

Punctuation can also help in delimiting local contexts, such as in the experiments described in Section 4.2.2.1 and Section 4.2.2.2, which results in an increase in BLEU but may prevent long range reordering if necessary (although not present in the translations that were analyzed).

4.4 Influence of errors on system performance

This experiment was designed in order to assess how each type of error influences the BLEU score of a SLT system.

It was performed on the TED talk system, thus in the EN-PT direction, with the conditions that yielded the best results: lowercased models, punctuation recovery and capitalization recovery only on capitalized words that do not follow a full stop. The first step was to perform, manually, a statistical frequency analysis of each type of error, with the goal of understanding which ones are the most common, while in the second step, each category of errors was manually corrected, one at a time, as if the MT engine had produced correct translations, accordingly to the references, for each error category. Each corrected output was then evaluated in order to understand how higher the BLEU score would be if a given error type was not present in the translation output.

Table 4.8 presents some general statistics about the test corpus and Table 4.9 shows the error frequency for each type of errors.

Test TED talk translation	
Number of lines	215
Number of words	3264

Table 4.8: Statistics of the test TED talk translation

Type of errors	Frequency
Missing filler words	36
Missing content words	42
Word order	54
Wrong lexical choices	20
Wrong disambiguation	132
Incorrect forms	149
Extra words	45
Idiomatic expressions	5
Named entities	1
Nationalities	1
Unknown words	49
Total	535

Table 4.9: Frequency for each type of errors present in the TED talk test text

With these results, it becomes obvious that the most common errors in EN-PT translation are the ones that concern disambiguation and incorrect forms, followed by word reordering.

The disambiguation problem is easily explained by the fact that, depending on the context, a given source language word may have several different translations in the target language, so choosing the correct translation for every different situation is not as easy task for the MT engine.

The incorrect forms are also easily explained by the fact that, as already presented in Section 4.3.3.3, Portuguese is a highly inflective language, as opposed to English, thus this type of errors is very common, especially for verbs, but also, although in a lesser extent, for words that require gender and number concordance such as adjectives or pronouns. For the opposite direction, it would be expected that these errors would be much less frequent, since the source-target word relation would be many-to-one instead of one-to-many.

The word reordering problem also comes from a common and well known difference between English and Portuguese, which is the ordering of a noun and a respective adjective.

For the BLEU evaluation after correcting each type of errors, Table 4.10 summarizes the results.

Types of corrected errors	BLEU	Gain
No corrected errors	20.95	0
ASR	33.66	12.71
Missing filler words	34.65	0.99
Missing content words	34.73	1.07
Word order	36.06	2.40
Wrong lexical choices	36.21	2.65
Wrong disambiguation	42.68	9.03
Incorrect forms	39.57	5.91
Extra words	33.84	0.18
Idiomatics + NE's + nationalities	34.88	1.22
Unknown words	36.52	2.86
All errors	59.97	25.31

Table 4.10: Partial and total BLEU results for the test TED talk after correcting each type of errors

Results show that there is a rough proportionality relation between the error frequency and the loss of BLEU, which means that, as expected, common errors have a larger responsibility in BLEU scores. But despite this fact, if error categories were ordered from most frequent to less frequent, and then ordered from higher loss of BLEU to lower loss of BLEU, the two orderings would not be the same. For instance, incorrect forms errors are the most common but the ones that have a larger influence in BLEU are disambiguation errors.

While both error types are similarly common, correcting disambiguation errors yields an increase of BLEU of almost the double relatively to incorrect forms.

This may be caused by the fact that incorrect forms, usually, do not influence the surrounding words, i.e. only one word should be changed into its correct form. When correcting disambiguation errors, one must also correct some of the surrounding context, since a disambiguation error will often cause ordering errors and missing words. Correcting only the disambiguation problem, in these cases, would not help understanding its role in the BLEU score, since the sentence would still be very different from the ones from the references. In these cases, the extra errors caused by the incorrect disambiguation were also corrected, which led to a gain of 9.03, where the more common incorrect form errors only produced a gain of 5.91.

On the bottom are extra and missing words, whose little influence is explained by the fact that in the first case, all the required words are already present in the translation, therefore, BLEU has already taken them into account when scoring the translation, and in the second case, the surrounding context is already correct and has also been correctly scored. Of course, when adding the missing words, the score increase is higher than when removing extra ones because in the latter, scores cannot rise much further.

There is also little difference between missing filler words and missing content words. Despite the fact that content words are much more important to the sentence meaning, BLEU does not take it into account, so every missing word affects the score equally, without distinction between filler and content.

Correcting unknown words and word order, also have a similar effect, as while a word was wrong or two words were switched, the context was correct, and correcting these errors only provided higher gains because they were more frequent. Also, unknown words are, sometimes, unknown expressions and word reordering usually takes place over two words, which may have helped in further increasing BLEU, relatively to single missing and extra words.

Incorrect lexical choices, while relatively uncommon, are also influential since, usually, when the MT engine selects a wrong translation candidate, the majority of the sentence is also incorrect, such as with the more common, but similar, disambiguation errors. Therefore, in these cases, one lexical choice error was corrected, not only on a single word, but on the entire expression that was affected by the error.

Idiomatic expressions, named entities and nationalities, which are only a marginal part of the entirety of the errors, end up having a larger influence since, usually, they are not isolated words, but instead, are larger expressions.

ASR errors clearly show how a talk recognized with a 14.086 WER, can produce errors that amount to 12.71 BLEU points, and how ASR failures seriously propagate to the MT component under the form of incorrect forms, incorrect lexical choices, or simply extra words that do not make any sense. Still, one could improve these results either by improving the ASR module itself, or by coupling ASR and MT more tightly with n-best lists or lattices.

Still, after correcting all errors, one might expect the score to be 100, but this is not the case. The final score of 59.97 comes from the fact that BLEU is based on the idea that the same sentence can be translated in *n* different ways, and therefore there is not just only one correct translation. The remaining 40 points represent the difference between this corrected output and the reference translations, which use synonyms of certain words and different, but valid, syntactic and semantic constructions.

So, from the references' point of view, some sections of the corrected text are wrong, when in reality they are not, and as such were not considered to be errors in this experiment. Thus, the corrected translation can be viewed as a third, equally valid and correct, reference. In fact, if repeating the whole process with this corrected text as a third reference, BLEU scores are much higher and reach a perfect 100 as show in Table 4.11.

Types of corrected errors	BLEU	Gain
No corrected errors	36.54	0
ASR	63.45	26.91
Missing filler words	64.79	1.34
Missing content words	65.31	1.86
Word order	66.91	3.46
Wrong lexical choices	65.87	2.42
Wrong disambiguation	75.09	11.64
Incorrect forms	72.76	9.31
Extra words	64.54	1.09
Idiomatics + NE's + nationalities	64.88	1.43
Unknown words	67.45	4.00
All errors	100.00	36.55

Table 4.11: Partial and total BLEU results for the test TED talk after using the corrected text as an extra reference

5 Conclusions and future work

The field of SLT has seen great advances in recent years. SLT has evolved from speech and domain-constrained systems to completely open and unlimited situations. Evolution has been accompanied by the development of several MT paradigms, from which the most flexible and most used one is phrase-based SMT, and the development of comprehensive toolkits and shared evaluations aimed at lowering entry barriers to new researchers.

One of the objectives of this work was to develop baseline experiments in order to perform automatic translation of speech between English and Portuguese, each presenting distinct and common challenges. These tasks are the translation of BN and the translation of spoken presentations. This work has been carried out by using available and widely used tools and resources such as Europarl, Moses, GIZA++ and SRILM, and the in-house ASR engine, AUDIMUS, from L2F. Publicly available resources, such as TED talks transcriptions were used to perform experiments in dialect adaptation from BP to EP with the goal of improving the quality of the training process. The two developed baseline systems have been tested under several different conditions and have been the subject of an error analysis with the objective of understanding how automatic evaluation metrics are influenced by translation errors.

Results show that the developed BP-EP adaptation system performs reasonably well, achieving BLEU scores around 90 when evaluated against a human adapted BP text. This produces high gains in BLEU (up to 30 points) in comparison to the original BP text evaluated with the same human adapted reference. This shows that after being adapted, the original BP texts, become 30% closer to being EP.

When evaluated by humans, in average, 78.96% of the adapted sentences were identified as being EP, which is also a very good indicator of the quality of the adaptation.

However, when using BP2EP to adapt resources to be used in the training of a translation system, results are inconclusive.

Early baseline experiments show that errors derived from the ASR engine have contributed to great decreases in BLEU scores and that the spontaneous speech in the test text has a large influence in the overall BLEU score. They also show that using capitalized models for translation, is not a suitable solution to handle capitalized words, since the number of unknown words quickly rises due to the fact that the majority of words that start a sentence will not be translated, causing large decreases in BLEU.

On the other hand, using an external capitalization and punctuation recovery module for ASR, helps in increasing BLEU scores when used to insert punctuation prior to translation and capitalization for potential named entities, which can be used to recover capitalizations of sentence starting words after translation, to delimit local translation contexts instead of larger segments and to prevent named entities of being translated as normal words, as if they were lowercased.

The error analysis shows that for a EN-PT talk translation task, involving ASR, the whole speech, lowercased models, punctuation and capitalization of only potential named entities, the most common errors are incorrect disambiguation and incorrect forms, that together represent 14.94 BLEU points when corrected and are the two most influential error types for this task, along with errors produced by the ASR component.

Instead of trying to correct errors, these results aim at providing an idea as to how and where to direct efforts in trying to solve particular error situations which are the most common or influential in the overall translation quality.

In summary, this work:

- Provides an overview of different state-of-the art approaches, resources and technologies in the field of SLT, and presents how modern challenges in this field have been tackled in recent advances.
- Presents a method of adapting BP texts into EP by applying transformation rules specific to each type of differences between the two dialects. Unfortunately, testing how translation quality is improved when using training material adapted by this method, has, so far, been inconclusive.
- Describes several baseline experiments in translating BN and TED talks.
- Shows how each type of errors present in the translation output from the mentioned experiments influences BLEU scores. On the other hand, this error analysis would be more useful if the errors could be automatically corrected after translation.

5.1 Future work in BP2EP

BP2EP results also show that the system lacks the handling of particular cases, such as contrasts and variations not present in the contrast list, time distances, expressions denoting existence, and the cases where a word or expression, which may be used in BP and EP, depends on its context to be specific to either one of the dialects, or common to both. Not handling these cases results in several sentences not being usable as training material in MT systems.

In order to achieve a better adaptation quality and usability as training material, these cases must be solved, either by extracting specific rules from larger and better linguistic resources, such as books translated in both dialects, or by applying a different technique in the adaptation process, such as statistical one.

This adaptation approach contains several decisions in cases that could have been handled in different ways. These decisions were weighed but any different way of handling these situations would have had its advantages and its drawbacks. Specifically, when handling gerunds, instead of applying a rule, the verb could have been left untouched and when handling the way of addressing, the word *você* could have be left in place or simply removed intead of being changed to *tu*.

The different outcomes of these decisions could also be tested in order to assess if the one adopted in this work, was indeed the best.

If the adaptation system could be extended to handle all of these situations, it would probably be able to produce texts very close to EP, which could prove to be very useful in increasing translation quality of MT systems for EP.

Finally, in order to assess how BP2EP performs when adapting training resources for a translation task, the adaptation should be tested also on the TM and larger texts for both the TM and LM would be best suited. However, since the only current available BP resources are TED talks transcriptions, which compared to the length of the Europarl corpus total only about 10% of it, such an experiment would probably produce weak BLEU scores and would not be useful for a real scenario. On the other hand, and considering the limitations imposed by the small relative size of the adaptation corpus, this work seems to be much more useful for experiments such as domain adaptation, where a small domain-specific corpus could be adapted and added or interpolated with the larger out-domain corpus. The gain or loss in the translation quality could then be evaluated relatively to the system prior to domain adaptation and between domain adaptation performed with the BP texts and with the adapted EP texts.

Aditionally, and also due to the difficulty of finding aligned EN-BP texts, the option of exploring comparable corpora instead of aligned corpora, such as translations of books, could be useful.

5.2 Future work in error analysis

Obviously, in a real situation, the described analysis approach would not be feasible, since manual correction is too expensive and time consuming. Also, these BLEU gains would also be near impossible, since automatically detecting and correcting each type of error is very difficult.

Still, it would be very important to try to develop a way of, as much as possible, automatically detecting errors with the goal of using this knowledge to improve the translation process and understand how to avoid producing them. It would also be very important if these errors could be corrected automatically, so that this correction module could also be integrated in the processing chain as an additional step, or coupled with the translation component.

In the future, in order to achieve this, it would be desirable to develop a tool or an error analysis pipeline that could detect patterns in the translation output that correspond to errors. For instance, by comparing the translation input with its output, detecting unknown words is very easy by simply searching for words that are common in the same sentence, the same applying for detecting missing and extra words by comparing the translation output and the references, or by using a POS-tagger to detect number and gender concordance problems for incorrect forms.

Bibliography

Allauzen, A., Crego, J., Max, A., & Yvon, F. (2009, March). LIMSI's statistical translation systems for WMT'09. In *Proceedings of the fourth workshop on statistical machine translation* (pp. 100–104). Athens, Greece: Association for Computational Linguistics.

Alshawi, H., Bangalore, S., & Douglas, S. (1998). Automatic acquisition of hierarchical transduction models for machine translation. In *Coling-acl* (p. 41-47).

Alshawi, H., Buchsbaum, A. L., & Xia, F. (1997). A comparison of head transducers and transfer for a limited domain translation application. In *Acl* (p. 360-365).

Altintas, K. (2002). A machine translation system between a pair of closely related languages. In *Seventeenth international symposium on computer and information sciences.*

Axelrod, A., Yang, M., Duh, K., & Kirchhoff, K. (2008, June). The University of Washington machine translation system for ACL WMT 2008. In *Proceedings of the third workshop on statistical machine translation* (pp. 123–126). Columbus, Ohio: Association for Computational Linguistics.

Bach, N., Eck, M., Charoenpornsawat, P., Kohler, T., Stuker, S., Nguyen, T., et al. (2007). The CMU TransTac 2007 Eyes-free and Hands-free Two-way Speech-to-Speech Translation System. In *Proc. of the international workshop on spoken language translation.* Trento, Italy.

Banerjee, S., & Lavie, A. (2005, June). METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization* (pp. 65–72). Ann Arbor, Michigan: Association for Computational Linguistics.

Batista, F., Caseiro, D. A., Mamede, N. J., & Trancoso, I. (2008, October). Recovering capitalization and punctuation marks for automatic speech recognition: Case study for the portuguese broadcast news. *Speech Communication*, *50*(10), 847-862.

Bertoldi, N., Cettolo, M., Cattoni, R., & Federico, M. (2007). Fbk @ iwslt 2007. In *International workshop on spoken language translation (iwslt)* (pp. 76–83).

Bertoldi, N., & Federico, M. (2009, March). Domain adaptation for statistical machine translation with monolingual resources. In *Proceedings of the fourth workshop on statistical machine translation* (pp. 182–189). Athens, Greece: Association for Computational Linguistics.

Brown, P. F., Pietra, V. J., Pietra, S. A. D., & Mercer, R. L. (1993). The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, *19*, 263–311.

Casacuberta, F., Vidal, E., & Vilar, J. M. (2002). Architectures for speech-to-speech translation using finite-state models. In *Proceedings of the acl-02 workshop on speech-to-speech translation: algorithms and systems* (pp. 39–44). Morristown, NJ, USA: Association for Computational Linguistics.

Charniak, E., Knight, K., & Yamada, K. (2003). Syntax-based language models for statistical machine translation. In *In mt summit ix. intl. assoc. for machine translation.*

Civera, J., & Juan, A. (2007, June). Domain adaptation in statistical machine translation with mixture modelling. In *Proceedings of the second workshop on statistical machine translation* (pp. 177–180). Prague, Czech Republic: Association for Computational Linguistics.

Condon, S., Parvaz, D., Aberdeen, J., Doran, C., Freeman, A., & Awad, M. (2010, may). Evaluation of machine translation errors in english and iraqi arabic. In N. C. C. Chair) et al. (Eds.), *Proceedings of the seventh conference on international language resources and evaluation (lrec'10).* Valletta, Malta: European Language Resources Association (ELRA).

Déchelotte, D., Adda, G., Allauzen, A., Bonneau-Maynard, H., Galibert, O., Gauvain, J.-L., et al. (2008, June). Limsi's statistical translation systems for WMT'08. In *Proceedings of the third workshop on statistical machine translation* (pp. 107–110). Columbus, Ohio: Association for Computational Linguistics.

Elliott, D., Hartley, A., & Atwell, E. (2004). *A fluency error categorization scheme to guide automated machine translation evaluation. amta: Machine translation: From real users to research.*

Federico, M., Bertoldi, N., & Cettolo, M. (2008). Irstlm: an open source toolkit for handling large scale language models. In *Interspeech 2008* (pp. 1618–1621). ISCA.

Forcada, M., Garrido, A., Canals, R., Iturraspe, A., Montserrat-Buendia, S., Esteve, A., et al. (2001). The spanish-catalan machine translation system internostrum. *0922-6567 - Machine Translation*, *VIII*, 73-76.

Fügen, C., Kolss, M., Paulik, M., & Waibel, A. (2006, June). Open domain speech translation: From seminars and speeches to lectures. In *Tc-star workshop on speech-to-speech translation* (pp. 81–86). Barcelona, Spain.

Germann, U. (2003). Greedy decoding for statistical machine translation in almost linear time. In *Hlt-naacl.*

Hajič, J., Hric, J., & Kuboň, V. (2000). Machine translation of very close languages. In *Proceedings of the sixth conference on applied natural language processing* (pp. 7–12). Morristown, NJ, USA: Association for Computational Linguistics.

He, Y., Zhang, J., Li, M., Fang, L., Chen, Y., Zhou, Y., et al. (2008). The CASIA Statistical Machine Translation System for IWSLT 2008. In *Proc. of the international workshop on spoken language translation* (p. 85-91). Hawaii, USA.

Hildebrand, A. S., & Vogel, S. (2009, March). CMU system combination for WMT'09. In *Proceedings of the fourth workshop on statistical machine translation* (pp. 47–50). Athens, Greece: Association for Computational Linguistics.

Knight, K. (1999). *A statistical mt tutorial workbook.* Unpublished, August.

Knight, K., & Al-Onaizan, Y. (1998). Translation with finite-state devices. In *Amta* (p. 421-437).

Knight, K., & Koehn, P. (2003). What's new in statistical machine translation. In *Hlt-naacl.*

Koehn, P. (2005). Europarl: A parallel corpus for statistical machine translation. In *Mt summit 2005.*

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., et al. (2007). Moses: Open source toolkit for statistical machine translation. In *Acl.*

Koehn, P., Och, F. J., & Marcu, D. (2003). Statistical phrase-based translation. In *Naacl '03: Proceedings of the 2003 conference of the north american chapter of the association for computational linguistics on human language technology* (pp. 48–54). Morristown, NJ, USA: Association for Computational Linguistics.

Koehn, P., & Schroeder, J. (2007, June). Experiments in domain adaptation for statistical machine translation. In *Proceedings of the second workshop on statistical machine translation* (pp. 224–227). Prague, Czech Republic: Association for Computational Linguistics.

Lane, I., Zollmann, A., Nguyen, T. L., Bach, N., Venugopal, A., Vogel, S., et al. (2007). The UKA-CMU Statistical Machine Translation Systems for IWSLT 2007. In *Proc. of the international workshop on spoken language translation.* Trento, Italy.

Lavie, A., , Levin, L., , Gates, D., , et al. (1997). Janus iii: Speech-to-speech translation in multiple languages. In *Proceedings of icassp 97.*

Lavie, A. (1996). *Glr*: a robust grammar-focused parser for spontaneously spoken language*. Unpublished doctoral dissertation, Pittsburgh, PA, USA. (Chair-Tomita, Masaru)

Lavie, A., Langley, C., Waibel, A., Pianesi, F., Lazzari, G., Coletti, P., et al. (2001). Architecture and design considerations in nespole!: a speech translation system for e-commerce applications. In *Hlt '01: Proceedings of the first international conference on human language technology research* (pp. 1–4). Morristown, NJ, USA: Association for Computational Linguistics.

Leusch, G., Matusov, E., & Ney, H. (2009, March). The RWTH system combination system for WMT 2009. In *Proceedings of the fourth workshop on statistical machine translation* (pp. 51–55). Athens, Greece: Association for Computational Linguistics.

Levin, L., Gates, D., Wallace, D., Peterson, K., Lavie, A., Pianesi, F., et al. (2002). Balancing expressiveness and simplicity in an interlingua for task based dialogue. In *Proceedings of the acl-02 workshop on speech-to-speech translation: algorithms and systems* (pp. 53–60). Morristown, NJ, USA: Association for Computational Linguistics.

Li, Z., Callison-Burch, C., Dyer, C., Khudanpur, S., Schwartz, L., Thornton, W., et al. (2009, March). Joshua: An open source toolkit for parsing-based machine translation. In *Proceedings of the fourth workshop on statistical machine translation* (pp. 135–139). Athens, Greece: Association for Computational Linguistics.

Liu, Y., He, Z., Mi, H., Huang, Y., Feng, Y., Jiang, W., et al. (2008). The ICT System Description for IWSLT 2008. In *Proc. of the international workshop on spoken language translation* (p. 52-57). Hawaii, USA.

Lonsdale, D. W., Franz, A., & Leavitt, J. R. R. (1994). Large-scale machine translation: An interlingua approach. In *Iea/aie* (p. 525-530).

Lopez, A. (2007, Apr). *A survey of statistical machine translation* (Tech. Rep. No. 2006-47). University of Maryland Institute for Advanced Computer Studies.

Marcu, D., & Wong, W. (2002). A phrase-based, joint probability model for statistical machine translation. In *Emnlp '02: Proceedings of the acl-02 conference on empirical methods in natural language processing* (pp. 133–139). Morristown, NJ, USA: Association for Computational Linguistics.

Mateus, M. H., Brito, A. M., Duarte, I., Faria, I. H., Frota, S., Matos, G., et al. (2003). *Gramática da língua portuguesa* (7 ed.). Lisboa: Editorial Caminho.

Mathias, L., & Byrne, W. (2006). Statistical phrase-based speech translation. In *Icassp 2006.*

Matusov, E., Kanthak, S., & Ney, H. (2005, September). On the integration of speech recognition and statistical machine translation. In *Interspeech* (p. 3177-3180). (ISCA Best Student Paper Award)

Matusov, E., Leusch, G., Bender, O., & Ney, H. (2005). Evaluating Machine Translation Output with Automatic Sentence Segmentation. In *Proceedings of the International Workshop on Spoken Language Translation (IWSLT) 2005.*

Mauser, A., Vilar, D., Leusch, G., Zhang, Y., & Ney, H. (2007, October). The rwth machine translation system for iwslt 2007. In *International workshop on spoken language translation* (p. 161-168). Trento, Italy.

McLachlan, G. J., & Peel, D. (2000). *Finite mixture models / geoffrey mclachlan, david peel* [Book]. Wiley, New York ; Chichester :.

Meinedo, H., Caseiro, D., Neto, J. a. P., & Trancoso, I. (2003). Audimus.media: A broadcast news speech recognition system for the european portuguese language. In N. J. Mamede et al. (Eds.), *Proceedings of the 6th international workshop on computational processing of the portuguese language (propor '03)* (Vol. 2721, pp. 9–17). Springer.

Munteanu, D. S., & Marcu, D. (2002). Processing comparable corpora with bilingual suffix trees. In *Emnlp '02: Proceedings of the acl-02 conference on empirical methods in natural language processing* (pp. 289–295). Morristown, NJ, USA: Association for Computational Linguistics.

Munteanu, D. S., & Marcu, D. (2005). Improving machine translation performance by exploiting non-parallel corpora. *Comput. Linguist.*, *31*(4), 477–504.

Nakov, P. (2008, June). Improving English-Spanish statistical machine translation: Experiments in domain adaptation, sentence paraphrasing, tokenization, and recasing. In *Proceedings of the third workshop on statistical machine translation* (pp. 147–150). Columbus, Ohio: Association for Computational Linguistics.

Nakov, P., & Hearst, M. (2007, June). UCB system description for the WMT 2007 shared task. In *Proceedings of the second workshop on statistical machine translation* (pp. 212–215). Prague, Czech Republic: Association for Computational Linguistics.

Nakov, P., & Ng, H. T. (2009a). Improved statistical machine translation for resource-poor languages using related resource-rich languages. In *Emnlp '09: Proceedings of the 2009 conference on empirical methods in natural language processing* (pp. 1358–1367). Morristown, NJ, USA: Association for Computational Linguistics.

Nakov, P., & Ng, H. T. (2009b, March). NUS at WMT09: Domain adaptation experiments for English-Spanish machine translation of news commentary text. In *Proceedings of the fourth workshop on statistical machine translation* (pp. 75–79). Athens, Greece: Association for Computational Linguistics.

Ney, H. (1999). Speech translation: coupling of recognition and translation. *Acoustics, Speech, and Signal Processing, IEEE International Conference on*, *1*, 517-520.

Och, F. J. (2003). Minimum error rate training in statistical machine translation. In *Acl* (p. 160-167).

Och, F. J., & Ney, H. (2000). Improved statistical alignment models. In *In proceedings of the 38th annual meeting of the association for computational linguistics* (pp. 440–447).

Och, F. J., & Ney, H. (2003). A systematic comparison of various statistical alignment models. *Computational Linguistics*, *29*(1), 19–51.

Och, F. J., & Ney, H. (2004). The alignment template approach to statistical machine translation. *Computational Linguistics*, *30*(4), 417-449.

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th annual meeting of the association for computational linguistics* (pp. 311–318). Philadelphia, PA.

Paulik, M., Rottmann, K., Niehues, J., Hildebrand, S., & Vogel, S. (2007, June). The ISL phrase-based MT system for the 2007 ACL workshop on statistical machine translation. In *Proceedings of the second workshop on statistical machine translation* (pp. 197–202). Prague, Czech Republic: Association for Computational Linguistics.

Popovic, M., Ney, H., Gispert, A. D., Marino, J. B., Gupta, D., Federico, M., et al. (2006). Morpho-syntactic information for automatic error analysis of statistical machine translation output. In *In proc. of naacl workshop on statistical machine translation* (pp. 1–6).

Quan, V. H., Federico, M., & Cettolo, M. (2005). Integrated n-best re-ranking for spoken language translation. In *Interspeech 2005* (pp. 3181–3184).

Ribeiro, R., Mamede, N. J., & Trancoso, I. (2003). Using Morphossyntactic Information in TTS Systems: comparing strategies for European Portuguese. In *Computational processing of the portuguese language: 6th international workshop, propor 2003, faro, portugal, june 26-27, 2003. proceedings* (Vol. 2721). Springer.

Rosti, A.-V., Zhang, B., Matsoukas, S., & Schwartz, R. (2009, March). Incremental hypothesis alignment with flexible matching for building confusion networks: BBN system description for WMT09 system combination task. In *Proceedings of the fourth workshop on statistical machine translation* (pp. 61–65). Athens, Greece: Association for Computational Linguistics.

Scannell, K. P. (2006). *Machine translation for closely related language pairs.*

Secara, A. (2005). *Translation evaluation - a state of the art survey.*

Soricut, R., Knight, K., & Marcu, D. (2002). Using a large monolingual corpus to improve translation accuracy. In *Amta* (p. 155-164).

Stolcke, A. (2002). *Srilm – an extensible language modeling toolkit.*

Tomita, M. (1987). An efficient augmented-context-free parsing algorithm. *Comput. Linguist.*, *13*(1-2), 31–46.

Ueffing, N., Haffari, G., & Sarkar, A. (2007). Transductive learning for statistical machine translation. In *Acl.*

Ueffing, N., Simard, M., Larkin, S., & Johnson, H. (2007, June). NRC's PORTAGE system for WMT 2007. In *Proceedings of the second workshop on statistical machine translation* (pp. 185–188). Prague, Czech Republic: Association for Computational Linguistics.

Vilar, D., Xu, J., D'Haro, L. F., & Ney, H. (2006, May). Error Analysis of Machine Translation Output. In *International conference on language resources and evaluation* (p. 697-702). Genoa, Italy.

Vogel, S. (2005). Pesa: Phrase pair extraction as sentence splitting. In *in proceedings: the tenth machine translation.*

Wahlster, W. (1993). Verbmobil - translation of face-to-face dialogs. In *Grundlagen und anwendungen der künstlichen intelligenz, 17. fachtagung für künstliche intelligenz, humboldt-universität zu* (pp. 393–402). London, UK: Springer-Verlag.

Waibel, A., Finke, M., Gates, D., Gavalda, M., Kemp, T., Lavie, A., et al. (1996). Janus-ii-translation of spontaneous conversational speech. In *Icassp '96: Proceedings of the acoustics, speech, and signal processing, 1996. on conference proceedings., 1996 ieee international conference* (pp. 409–412). Washington, DC, USA: IEEE Computer Society.

Waibel, A., & Fügen, C. (2008). *Spoken language translation - enabling cross-lingual human-human communication* (No. 3). IEEE Signal Processing Magazine.

Wang, H., Wu, H., Hu, X., Liu, Z., Li, J., Ren, D., et al. (2008). The TCH Machine Translation System for IWSLT 2008. In *Proc. of the international workshop on spoken language translation* (p. 124-131). Hawaii, USA.

Wittmann, L. H., Pêgo, T. R., & Santos, D. (1995). *Português brasileiro e português de portugal: algumas observações.*

Wölfel, M., Kolss, M., Kraft, F., Niehues, J., Paulik, M., & Waibel, A. (2008). Simultaneous machine translation of german lectures into english: Investigating research challenges for the future. 233-236.

Woszczyna, M., Aoki-Waibel, N., , Coccaro, N., Horiguchi, K., Kemp, T., et al. (1994). Janus 93: Towards spontaneous speech translation. In *Proceedings of the icassp 94.*

Wu, D. (1997). Stochastic inversion transduction grammars and bilingual parsing of parallel corpora. *Comput. Linguist.*, *23*(3), 377–403.

Wu, H., Wang, H., & Zong, C. (2008, August). Domain adaptation for statistical machine translation with domain dictionary and monolingual corpora. In *Proceedings of the 22nd international conference on computational linguistics (coling 2008)* (pp. 993–1000). Manchester, UK: Coling 2008 Organizing Committee.

Yamada, K., & Knight, K. (2001). A syntax-based statistical translation model. In (pp. 523–530).

Yamada, K., & Knight, K. (2002). A decoder for syntax-based statistical mt. In *Acl* (p. 303-310).

Zhang, R., Kikui, G., Yamamoto, H., Watanabe, T., Soong, F., & Lo, W. K. (2004). A unified approach in speech-to-speech translation: integrating features of speech recognition and machine translation. In *Coling '04: Proceedings of the 20th international conference on computational linguistics* (p. 1168). Morristown, NJ, USA: Association for Computational Linguistics.

I
Appendices

A Examples of speech translation systems and projects

This appendix presents an overview of some speech translation systems and projects:

A.1 JANUS

JANUS (Woszczyna et al., 1994) was an early speech-to-speech translation effort, which aimed at providing a way for two different parties to establish a spontaneous dialog in a limited domain.

Developed in cooperation between CMU and the Karlsruhe University, it was initially designed for the Scheduling domain in which two participants establish a negotiation dialog with the purpose of scheduling a meeting, with English, German and Spanish as both source and target languages.

The most recent version of JANUS was called JANUS III (Lavie et al., 1997), which was built upon several improvements on its predecessors and started to scale up into other languages such as Japanese and related domains other than Scheduling, namely travel planning. Travel planning is still limited but it is a much more complex domain since it has more types of interaction and is composed of several sub-domains (transport, hotel accommodation or events).

JANUS had several components: speech recognition, parsing, translation and speech generation. Each of these modules was designed to be language independent to allow an easy adaptation to different languages and domains.

In order to deal with lack of robustness and translation accuracy, two distinct parsers with complementary strengths are used, and its results combined: the GLR translation module and the Phoenix translation module (Waibel et al., 1996). The GLR module produces more complete and accurate translations and the Phoenix module is more robust when dealing with speech disfluencies.

A GLR parser (Tomita, 1987) is an extension of the standard LR algorithm specifically designed to deal with natural language's inherent ambiguity. JANUS's GLR* parser (Lavie, 1996) further enhances the standard GLR algorithm by dealing with noise in the input and limited grammar coverage, which seemingly suits spoken input characteristics such as disfluencies, stuttering, repetitions, false starts, ungrammaticalities and speech recognition errors. Within the GLR translation module, the parser is designed to produce grammars with feature structures that correspond to a frame-based and language-independent semantic representation of the discourse's sequential spoken utterances, which are called semantic dialog units (SDU). Thus, the parser captures the meaning of each SDU by mapping them to a detailed Interlingua format (C-STAR's Interchange Format) as exemplified in the following sentence and respective representation:

We have two hotels available.

a:give-information+availability+hotel((hotel-type=hotel,quantity=2))

The generation is done by GenKit (Tomita and Nyberg, 1988), which compiles unification grammars and produces very accurate results for well specified Interlingua texts.

The Phoenix module, uses the Phoenix parser, which unlike GLR* only attempts to identify the key semantic concepts in an utterance and its underlying structure. The parser creates grammars where in-domain concepts are represented, and compiles them in hierarchical structures called Recursive Translation Networks (RTN) before matching as much of the input to the patterns specified in the RTNs as possible. Thus, by ignoring any unknown or unnecessary utterance between each two consecutive top-level concepts, the parser can handle unexpected and out-of-domain input. The output is simply produced by sequentially generating, in the target language, each consecutive concept. The result is an awkward and almost telegraphic sentence, but still meaningful.

The JANUS systems have been the subject of several end-to-end evaluations from speech input to translation output. Bilingual graders compare the source language input SDUs (transcribed and recognized) with the target language output SDUs and assign a grade, based on the following well-defined criteria.

- Acceptable
  - Perfect - Fluent translation with all information conveyed
  - OK - All important information translated correctly but some unimportant details missing or translation is awkward
- OK tagged - The sentence or clause is out-of-domain and no translation is given
- Bad - Unacceptable translation

On a 1997 Scheduling domain, Spanish to English, evaluation for JANUS III, comprised of 3 dialogs (103 utterances) spoken by unknown voices, 83.3% of transcriptions are acceptably translated when combining both parsers and 63.3% of recognized SDUs achieve the same acceptable status, which is a large improvement over the results obtained by each of the parsers on its own.

On a 1999 Travel planning domain evaluation comprised of 132 unseen sentences, where a subject is trying to book a trip to Japan, JANUS III gets slightly lower scores using only the Phoenix module:

77% on transcribed and 62% on recognized SDUs with the English-Japanese pair, and 70% transcribed and 58% on recognized SDUs on the English-German language pair.

A.2 NESPOLE

NESPOLE (Lavie et al., 2001) was a project funded by the US NSF[1] and the European Comission and developed by a collaboration between IRST[2] in Trento Italy, ISL[3] at University of Karlsruhe in Germany, UJF[4] in Grenoble France, CMU in Pittsburgh and two industrial partners (APT- the Trentino provincial tourism bureau, and Aethra - an Italian telecommunications commercial company).

Its goal is to perform speech-to-speech translation in practical and realistic e-commerce scenarios (mainly in the tourism domain) by focusing not only on translation performance, but on system-wide usability as well. It uses a distributed client-server architecture to allow user to connect in real time to a human agent of the service provider, who speaks another language, trough a web browser, and using commercially available video conferencing technologies and standard audio and video hardware. Both parties speak and hear the synthesized translated speech in their own languages and in real time. The system tries to take advantage of multimodal interaction in order to enhance oral communication between the two parties, by using a whiteboard where users can point to shared maps, documents and websites.

NESPOLE has a highly modular architecture as it was designed with special attention on flexibility and easy development and integration of new and improved language-specific translation modules into the global system. Global NESPOLE servers and its language-specific servers are clearly separated from the user end and from the communication channels, which are managed by a dedicated module, the Mediator. The system tries to further enhance its design flexibility by using an Interlingua translation approach. NESPOLE's Interlingua approach builds upon previous work initially designed by some of its partners in the C-STAR context, and extends it. The following examples are utterances and their respective Interlingua representation.

Thank you very much.

c:thank

And we'll see you on February twelfth.

a:closing (time=(February, md12))

On the twelfth we have a single and a double available.

[1] United States National Science Foundation

[2] Center for Scientific and Technological Research

[3] Interactive Systems Laboratories

[4] Universite Joseph Fourier

a:give-information+availability+room(room-type=(single double), time=(md12))

This allows different research sites to develop their own language-specific analysis and generation modules independently, using different approaches while adhering to the system's communications standards and Interlingua representation.

An important issue is that network conditions must be addressed so that network traffic does not prevent the system from serving its purpose. This distributed architecture enables the modules to run in physical locations so that network configuration is optimal. Still, it has to deal with insufficient bandwidth conditions and in order to guarantee real time communication between client and agent, the system has to drop short segments of video and speech that were delayed, contributing to lower performances in speech recognition.

In 2001 the system's first-showcase was the subject of a multi-perspective evaluation targeting several aspects: impact of network traffic, impact of multimodality and an end-to-end evaluation. The scenario consisted in a German, French or English speaking client trying to get information about winter sports possibilities in the Alps region of Trentino using a NetMeeting connection, and in an Italian speaking agent at APT answering the questions. The end-to-end evaluation of NESPOLE first showcase system was done by using four unseen dialogs, and by grading the Interlingua SDUs following the same human evaluation criteria as in JANUS, as the Interlingua approach is essentially identical. The system produced better results with the English to Italian translation: 55% acceptable SDUs for transcribed text and 43% for SDUs resulting from speech recognition.

A.3 Verbmobil

Verbmobil (Wahlster, 1993) was a long term project, that ran from 1993 through 2000, focused on developing a portable device, such as a mobile phone, that can be carried to face-to-face situations, for instance, a meeting with people who speak foreign languages, and translate on demand what they are saying. It was designed to handle three distinct limited domains: appointment scheduling, travel planning and remote PC maintenance. The project was funded by the German Ministry for Research and Technology (BFMT) and carried out by an international industrial and academic consortium of 29 partners, including the German Research Center for AI (DFKI), the ATR Interpreting Telecommunications Research Laboratories in Kyoto, and US research groups at CMU, Stanford University and Berkeley.

It assumes that both participants in the conversation are German or Japanese passive speakers of English, but not fluent, and uses English as the pivot language (common understanding language) most of the time. Despite this, in some technical situations or uncommon and complex words and phrasings, the user may switch the device to its own language for better understanding. So, the system must follow the discourse even when the user speaks in English with an accent, hesitates, or says something wrong, must be capable of translating to English when the user speaks in its own language and must be able to deal with switches from the source languages to English and vice-versa while preserving coherence in the translated output.

Since the system focuses on a face-to-face setting, it has to ensure close to real-time performance. In order to do so, the analysis and translation should be as shallow as possible while maintain understandability in the results, and the generation should be performed as soon as possible. The system uses anytime modules, which are composed of sequential processing steps and yield better results as computation time increases. The purpose of these anytime modules, is to allow the system to interrupt processing and still generate inaccurate, yet usable translations.

The system is composed by 69 independent and highly interactive modules that combine several techniques at each processing stage. A distinguishable characteristic is that it uses prosodic information about the source utterance systematically at all processing steps, as it is passed on to the parsing and translation modules in order to improve the generation and synthesis of the target utterance. This is an important feature since prosodic differences in one language can result in syntactic or semantic differences in another language, for instance, the German sentence fragment *wir haben noch* can be translated as *we still have* or *we have another*, whether the word *noch* is stressed by the speaker.

Verbmobil uses three parallel parsing strategies: a chunk parser, a statistical LR parser and a HPSG parser[5]. Each parser processes the same input word hypothesis with the same prosodic annotations, but produces different partial results. These results are used by a semantic construction module in order to produce partial canonical semantic representations called Verbmobil Interface Terms (VITs), which are compiled in charts and reconstructed by a robust semantic module into the complete VIT corresponding to the input source utterance.

The system has five different translation modules that combine both shallow and deep approaches: case based, sub-string based, statistical, dialog based and semantic transfer. While, for instance, the statistical approach provides quick and dirty results and is able to naturally handle speech recognition problems in a robust way, semantic transfer is more expensive computation-wise but produces more accurate results. A selection module chooses which translation has the best score in each source language fragment and combines each approach's strengths to produce the most suitable result to be generated.

Verbmobil was evaluated by trying to assess the approximate correctness of each translated speech turn in light of the project's objectives. This means that the evaluation criteria must not be too tight. Rather than considering only close to perfect translations as acceptable as in JANUS and NESPOLE evaluations, Verbmobil evaluators focused on the understandability and preservation of the original meaning captured by the translations. Only two outcomes are possible: if the translated speech turn is

[5] Head-Driven Phrase Structure Grammar

both understandable and able to preserve the original meaning, it is considered acceptable, if it fails on either criterion, it is unacceptable. Neither of the two criteria have more weight than the other.

Final results, at the end of the project in 2000, showed that the average processing time was about four times the duration of the input signal, spontaneous speech recognition achieved a 75% rate, approximately 80% translations managed to maintain understandability and what the speaker originally intended to transmit, and a 90% success rate was obtained in real scenarios with real users performing dialog tasks.

A.4 Babylon/TRANSTAC

Babylon was a project funded by DARPA, which aimed at rapidly developing two-way speech-to-speech systems in several languages for military use in combat and other field situations such as force protection, refugee processing and medical triage. Its seed project, called Rapid Multilingual Support (RMS), was deployed to Afghanistan in 2002 and focused on high-risk terrorist situations and low-resource languages that were not supported by commercially available systems. Mandarin, Arabic, Pashto and Dari were the selected languages. In order to support field operations, the systems were to be delivered in handheld off-the-shelf PDAs devices and standard laptops.

Babylon evolved into TRANSTAC[6] which was another speech translation program launched by DARPA in 2006. Its goal extended Babylon's goal and was to enable two-way spontaneous speech communication between English speaking military personnel and local foreign language speaking people in tactically relevant military environments. In a force protection situation, there are not many willing or trusted local translators available, thus the program aimed at demonstrating capabilities that allowed speakers to communicate, in real-time, with each other in life-threatening contexts, where quick, accurate and robust translations are essential in order to ensure the safety of military personnel and local population. Another important objective of the program was the rapid development and deployment of new and improved versions of the systems, less than 100 days, mainly concerning new and low-resource language pairs, such as Arabic dialects (Iraqi) or Farsi, as the areas and languages of interest to the US Department of Defense may change rapidly in times of national emergency.

A.4.1 Mastor

Perhaps the most successful system under the flag of the Babylon program was IBM's Mastor project. Initially called IBM MARS S2S, it handled the air travel reservation domain, and was significantly improved to deal with broader domains such as medical interaction and force protection. Eventually, it

[6] Spoken Language Communication and Translation System for Tactical Use (http://www.darpa.mil/IPTO/programs/transtac/transtac.asp)

was also applied in Transtac.

Mastor initially used a decoupled cascaded architecture where the output of the speech recognizer was passed to a translation module, which generated the translated sentence to be synthesized by the speech synthesis component. Acoustic models for English and Mandarin were trained over 200 hours of speech for each language, while the Arabic recognizer was trained with only 50 hours of speech containing about 200k short utterances. The English language model was an linear interpolation between one component using in-domain data and another background component that consisted in a very large generic and domain independent collection of data, while the Arabic language model was an interpolation between a trigram LM, a class-based LM and a morphologically processed LM, all trained with a small corpus of words. Two statistical translation approaches were combined: a concept-based approach that takes advantage of natural language understanding (NLU) and natural language generation (NLG) models based on an annotated corpus, and a phrase-based finite state transducer approach which was trained in as un-annotated corpus. The latter approach was further investigated into a framework called Statistical Integrated Phrase Lattice (SIPL), which aimed at a tighter integration between ASR and MT components. This enhanced translation strategy proved to be much faster and memory efficient, thus more suited to the project's hardware limitations, while producing more accurate results when combined with the NLU/NLG approach, and a formal syntax component that increases vocabulary coverage and handles unseen word sequences more fluently.

Mastor obtained, in November of 2008, a WER score of under 20% in noisy environments and just over 2% in quiet environments, for English, and 25% in Farsi, in quiet conditions, while in January 2007 it reached a BLEU score close to 50%. Evaluation on low-level concept transfer was also conducted, in order to assess the meaning preservation between the source utterance and its translation, by a panel of judges that compared a source-language transcription to the target-language MT output and scored each low-level concept as correctly transferred, deleted, transformed into something else, or inserted. In June 2008, the pooled odds of successful transfer from English to Arabic and Arabic to English were of 6.46.

A.4.2 CMU TRANSTAC and IraqComm

Two other systems submitted to Transtac were the CMU TransTac 2007 Eyes-free and Hands-free Two-way Speech-to-Speech Translation System (Bach et al., 2007) and SRI International's IraqComm. Both systems were extensions to what was already done for the Babylon project, and were designed to be mobile and ran in off-the-shelf platforms such as handhelds or laptops specially prepared to sustain use in a war environment.

The CMU system focuses on the need of the military fighter to keep his hands and eyes as free as possible, as there is no visual interface, only the user's voice is necessary to operate the system, and only its sound output is necessary to provide feedback. The system can even be fully controlled only by its users' voice, since he can instruct the system to turn translation mode on or to repeat the last translation by issuing voice commands.

The system is also capable of performing automatic and continuous speech recognition and translation, and is able to operate under various conditions (indoors, outdoors, noisy combat environments), and adapt itself to the speaker. On the other hand it can also be controlled manually by push-to-talk buttons. Another important feature is that after performing the translation, the system translates the target sentence back to the source language. This allows the user to understand if the sentence was indeed correctly translated and what he wanted to say was correctly transmitted to the target language speaker.

By training the speech recognition engine with about 300 hours of American speech, 320 of Iraqi speech and 110 of Farsi speech it achieved close to 28% WER on the English-Iraqi pair and 26% on English-Farsi. On the MT component, a statistical approach was employed by using PESA (Vogel, 2005) and the Pharaoh toolkit to train and combine two phrase tables in order to cover most of the foreign vocabulary, since, for instance, more than 50% of Farsi words were not present in Pharaoh's phrase table because the sentences were very long. For the English-Farsi system, which was developed in 90 days, with both phrase tables combined and with a 6-gram suffix-array language model, it achieved a BLEU score of 16.44.

IraqComm[7] was a more commercial system that, although addressing the same issues as the CMU system, provided the user with a graphical interface in addition to the eyes-free and nearly-hands-free control mode. It also displays the translations in a textual form on the device's screen. To speed-up the interaction and due to its graphical interface, it provides a shortcut menu with the most used sentences.

Since it was a part of TRANSTAC's phase I, IraqComm was not evaluated in the same systematic way as the CMU system (phase II). Instead, it was deployed in Iraq, where a limited number of units were tested in a more task-oriented way, and produced rudimentary translations of high-level concepts in noise-free environments and when speech was made of short and easily understandable sentences.

A.5 GALE

GALE[8] is a project launched by DARPA, in 2005, with the goal of developing technologies to gather, analyze and extract relevant content from very large foreign language, and potentially threatening, speeches

[7] IraqComm (http://www.iraqcomm.com/)

[8] Global Autonomous Language Exploitation (http://www.darpa.mil/ipto/programs/gale/gale.asp)

and texts. The project is aimed at eliminating the need for linguists and translators and providing decision makers, military leaders and field personnel with an integrated and automated system that is able to transform large quantities of raw data, in different formats, from foreign sources into vital, precise, relevant and easily understandable information.

The systems are required to have three independent and self-contained engines: transcription, translation and distillation. Although independent, these modules must be tightly integrated and produce not only downstream outputs to the next component, but also English texts and references to be available to human users. While transcription (along with speech recognition) and translation are standard features in SLT, a distillation module is something entirely new. The user may query the system by using several human-computer interfaces, such as natural language, and it must apply language analysis techniques to identify and extract all of the relevant and non-redundant information regarding a user query over a collection of both foreign and English data sources. The following example shows a user query and a system response:

Query: *Describe attacks in Kuwait*

Response: *Since Jan. 10 (2005), police have clashed with Muslim fundamentalists and pursued them around the country, killing eight militants and arresting scores of others.*

Each engine must support several forms of speech and text inputs, namely, radio and television broadcast news, talk shows (in-studio and spectator calls), telephone conversations, newsgroups and weblogs, must be language and domain independent and scalable and adaptable to new languages, source data forms and speech or writing styles. In addition to English, Chinese and Arabic were the starting source languages.

Its main challenges are to achieve, by 2010, 90-95% accuracy in both transcription and translation into English and to gather knowledge and intelligence and distil information with the same quality as a human analyst, or even higher. These levels must be maintained consistently across different languages, domains and styles.

Three teams were hired to develop the solutions requested by DARPA: BBN, IBM and SRI International. Competition was increased with the possibility of at the end of each year's evaluation the worst team could be eliminated from the project and lose million-dollar contracts on which they heavily rely, and academic research was boosted as these companies sub-contracted top university lab investigators to work on the project.

Proposed approaches to reach the intended 90-95% accuracy goal, typically included combinations of strengths of different translation and linguistic paradigms: rule-based translation which produces high-quality translations but is very hard to scale and improve, statistical translation which although being context sensitive and requires large training data, is more easily extendable to other languages, NLP statistical paradigms such as parsers, morphology information, paraphrasing, ACE to perform entity identification and extraction (names) and analysis of information, and new methods to perform language modeling such as domain specific and distance-based LMs.

By 2007, The BBN team produced 75.3% accuracy and 69.4% on Arabic text and speech, respectively, and 75.2% and 67.1% on Chinese. IBM scored higher with Arabic text and SRI scored even higher in Chinese.

A.6 Lecture translator

A system aimed at solving the combined challenge of very specific domain adaptation and online speech segmentation is the Simultaneous Lecture Translation System at the University of Karlsruhe (Fügen et al., 2006), which performs translation from German to English. Since 2006, some approaches have been proposed as improvements over the ASR and MT components.

Some attention has been placed into dealing with language-specific details, in this case German, and lecture-specific problems. The fact that German has a high number of compound words poses a great difficulty for the recognition engine, since that in a standard configuration it uses a static vocabulary, and by having to deal with a possibly infinite number of words the vocabulary no longer is static. In order to reduce the vocabulary and OOV[9] rates, compound words are split into their parts, keeping in mind that splitting every compound word would result in a loss of context, mainly with in-domain expressions. Therefore, only out-of-domain compound words are split. However, the best suited splitting for the ASR component may not be the best for the MT one. Since MT does not need source language information to perform language modelling, it does not depend on compound splitting and thus if it is given an incorrectly split input it may be able to correct it by translating whole phrases. On the other hand, in an integrated speech translation system, the recognizer output must match the MT system, and in order to guarantee it, this system performs a more aggressive splitting on top of the previously described technique both in the training data and in the ASR output.

The fact that in the context of lectures speech is often composed by very technical knowledge has also received some attention. In fact, in these situations, is very common for the speaker to use English expressions in the middle of German speech. Because English and German derive from the same origins, many words also share some pronunciation and spelling characteristics and thus transcription, such as *kind* or *man*. On the other hand, some words may be spoken similarly but have completely distinct transcriptions and meanings such as *eagle* and *Igle*. Recognition results showed that by mapping the English pronunciation dictionary into the German one, and by using both dictionaries in a parallel way, the average WER over both languages decreases by 1%.

[9] Out Of Vocabulary

Speaker adaptation is performed by adapting acoustic and language models to specific speakers. This means that apart from training the ASR engine with the usual data sources such as EPPS[10] and CHIL[11] seminars, BTEC, broadcast news or TED talks, the models are extended with the use of recorded lectures and tuned offline with text presentations and publications from that individual speaker and relevant information retrieved from web sources.

Recognition is also sped up in order to run faster than real-time by restricting the search space, and by favoring short sequences of words when performing the speech segmentation. However, suitable online segmentation has to be taken into account since the MT component follows a statistical approach and, therefore, is trained on phrase-aligned text and expects phrase-aligned input as well.

A decoupled architecture is used to integrate ASR and MT in which a 1-best hypothesis is produced by the recognizer and passed to the translation engine. The traditional approach at segmenting the translation input would be to take the 1-best hypotheses and split it into smaller phrases before passing it the MT component, but it would also severely increase latency times. In order to run in real-time, the system uses silence regions and a 3-gram source language model to identify segment boundaries. But using a standard pipelined approach would create a dilemma: should one create shorter segments and thus speeding the recognition at the cost of losing some context information, or should one favor lengthier segments which would capture the entire context of the sentence and produce better translations at the cost of increasing latency times? A compromise between the two would decrease translation quality and still would not be able to maintain real-time performance.

The alternative employed by this system is to perform stream decoding while the continuous speech input stream is processed directly in real-time. Using a continuously refreshed translation lattice when input is received and output is generated, this continuous decoder decides on when to generate outputs independently from fixed segment boundaries. Simultaneously it performs word reordering within a sliding fixed-size window that defines the maximum latency time. When most words that need to be reordered are inside the window, the system achieves the same performance as the traditional segmentation approach at very low latencies.

Despite these improvements, and the fact that English and German have the same origins, word reordering is a very difficult task and has great costs for the overall performance of the system. In German, verbs, typically, appear at the end of clauses whereas in English they appear in the second position, moreover, several German verbs may consist of two parts: the auxiliary appears at the end of the sentence and the main part appears in the second position. Thus, reordering inside a small window will lead to poor translations in the specific case of German to English. The solution to this problem was to use a rule-based reordering strategy in which a lattice is built at decoding time with different word

[10] European Parliament Plenary Sessions

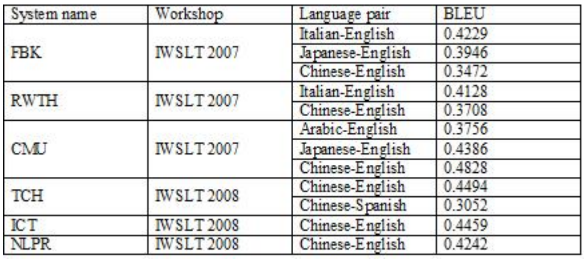
[11] Computers in the Human Interaction Loop

orders and rules learned from POS-tagged data. In the training, reordering patterns are extracted based on the POS information from word alignments and the context on which the pattern was extracted is added to the reordering model as a new feature.

End-to-end evaluations conducted in 2008 show that the system achieves a 12.5% WER on real-time and standard-hardware conditions and a 33.8 BLEU score. Human experts further evaluated the system and revealed that the English translation often maintained a rough idea of what the speaker intended to transmit. However, much work remains to be done concerning word reordering as the translation results are still awkward and difficult to read and understand. The evaluations also concluded that the system obtained an average latency of 7 words plus some extra time included by all of the automatic components of the system, resulting in delay of a one to two sentences relatively to the speaker, which is especially serious if the speaker has already changed to the next slide.

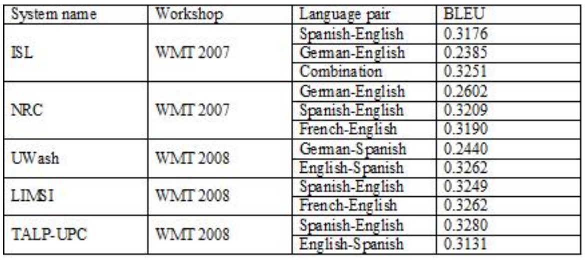
B Systems results in recent editions of IWSLT and WMT

This appendix shows the results obtained by several systems in recent editions of IWSLT and WMT:



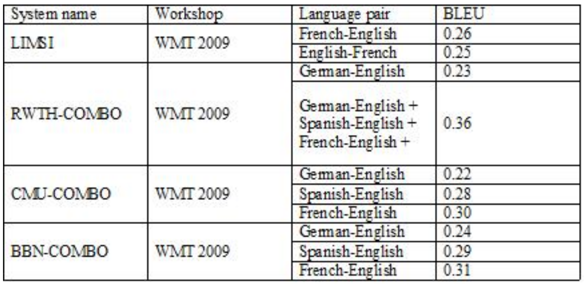
System name	Workshop	Language pair	BLEU
FBK	IWSLT 2007	Italian-English	0.4229
		Japanese-English	0.3946
		Chinese-English	0.3472
RWTH	IWSLT 2007	Italian-English	0.4128
		Chinese-English	0.3708
CMU	IWSLT 2007	Arabic-English	0.3756
		Japanese-English	0.4386
		Chinese-English	0.4828
TCH	IWSLT 2008	Chinese-English	0.4494
		Chinese-Spanish	0.3052
ICT	IWSLT 2008	Chinese-English	0.4459
NLPR	IWSLT 2008	Chinese-English	0.4242

Figure B.1: Results of several systems in IWSLT editions of 2007 and 2008



System name	Workshop	Language pair	BLEU
ISL	WMT 2007	Spanish-English	0.3176
		German-English	0.2385
		Combination	0.3251
NRC	WMT 2007	German-English	0.2602
		Spanish-English	0.3209
		French-English	0.3190
UWash	WMT 2008	German-Spanish	0.2440
		English-Spanish	0.3262
LIMSI	WMT 2008	Spanish-English	0.3249
		French-English	0.3262
TALP-UPC	WMT 2008	Spanish-English	0.3280
		English-Spanish	0.3131

Figure B.2: Results of several systems in WMT editions of 2007 and 2008



System name	Workshop	Language pair	BLEU
LIMSI	WMT 2009	French-English	0.26
		English-French	0.25
RWTH-COMBO	WMT 2009	German-English	0.23
		German-English + Spanish-English + French-English +	0.36
CMU-COMBO	WMT 2009	German-English	0.22
		Spanish-English	0.28
		French-English	0.30
BBN-COMBO	WMT 2009	German-English	0.24
		Spanish-English	0.29
		French-English	0.31

Figure B.3: Results of several systems in WMT edition of 2009

C Running BP2EP

This appendix presents the manual for running BP2EP:

Run MARv: cat *BPtextFile* | xipl2f -prexip > *POSTaggedFile*

Extract best hypothesis: cat *POSTaggedFile* | sed -n '/hmm/p' > *bestHypothesisFile*

Correct contractions: java ContractionsCorrector/Main *bestHypothesisFile* > *correctedBestHypothesisFile*

(MARv separates a contracted word into its components)

Run BP2EP: java BP2EP/Main *correctedBestHypothesisFile verbsListFile contrastListFile contrastsExtractedFromCorpusFile* > *outputFile*

(open files to see input formats)

D Subjective evaluation of BP2EP

This appendix presents the set of sentences used to perform the subjective evaluation of BP2EP:

BP Agora, o slide show. Eu o atualizo a cada apresentação.

BP É um alvo fácil e visível, e é assim que deve ser, mas há mais poluição que causa aquecimento de construções do que de carros e caminhões.

BP Isolamento, melhor design, comprar eletricidade verde onde você puder.

BP Dormi no avião, até que no meio da noite, aterrisamos nos Açores para reabastecer.

BP E comecei contando a história que havia acabado de acontecer no dia anterior em Nashville.

BP Uma rede de restaurantes familiares barateiros, para quem não conhece.

BP Invista em sustentabilidade. Majora falou isso.

BP Não apenas esta, mas conectados com as idéias que estão aqui, para fazê-las mais coerentes.

BP O bagaço fica lá empilhado perto do moinho de açúcar até que eventualmente eles o queimam.

BP Eu estou falando do mundo onde mulheres e crianças gastam 40 bilhões de horas por ano buscando água.

BP Bom, agora vou levar vocês para um continente diferente.

BP Você o veda para restringir a quantidade de oxigênio que entra na fornalha, e então você acaba com esse tipo de material carbonizado.

BP É um projeto que tem potencial para ter um grande impacto ao redor do mundo.

BP Então esse é um projeto que eu acho extremamente excitante, e eu realmente estou ansiosa para ver onde ele vai nos levar.

BP Elas os conhecem muito bem, mas não tem outra escolha.

BP Isso leva a todo tipo de problemas ambientais e problemas que afetam as pessoas por toda a nação.

BP Com a força que lhe resta, você liga para o 911 e reza por um bom médico.

BP Se você melhorar um software várias vezes, você acaba estragando ele.

BP Você compra o gabinete, eu te vendo o código.

BP Existe uma pequena empresa que está se saindo muito bem com essa história de simplicidade e elegância.

BP As pessoas gostam de se rodear de capacidades desnecessárias, correto?

BP Agora já é noite, e meu jantar está ficando frio - e está apodrecendo.

BP Você vê um novo documento em branco?

BP As coisas estão chegando em um ritmo mais rápido.

BP Sua ligação poderá ser gravada para nosso controle de qualidade?

EP Acordei, abriram a porta, saí para apanhar um pouco de ar fresco, e, de repente, vi um homem a correr na pista.

EP Estes são investimentos que se pagam a si próprios.

EP Fiz a minha palestra, voltei para o aeroporto para apanhar o avião para casa.

EP Há apenas dois dias tivemos novos recordes de temperatura em Janeiro.

EP Este movimento vai crescer muito mais rapidamente do que as projecções prevêem.

EP Coloquem-se no meu lugar!

EP Era hora de jantar e começámos a procurar um lugar para comer.

EP Vou contar-vos uma pequena história para ilustrar como tem sido a minha vida.

EP Podem encontrar-se famílias como estas que procuram uma árvore na floresta, abatem-na e fazem carvão com ela.

EP Fica apenas num grande monte perto da refinaria até que acabam por queimá-lo.

EP Isso é como se toda a mão de obra do Estado da Califórnia trabalhasse a tempo inteiro durante um ano não fazendo mais do que procurar água.

EP Umas das outras coisas em que estamos a trabalhar é na procura de maneiras para testar a qualidade da água a baixo custo.

EP Actualmente, 98% do Haiti está desflorestado.

EP Então começa-se com o bagaço e depois, com uma simples fornalha, que se pode fazer com um bidão velho de 25 l.

EP Então precisamos de criar este futuro, e precisamos de começar a fazê-lo agora.

EP Refiro-me ao mundo onde mulheres e crianças passam 40 biliões de horas por dia na busca de água.

EP Vou mostrar-vos o que quero dizer.

EP E serás pago por isto.

EP Estou a dizer-vos, está lá!

EP Certo, então esta empresa não está a contar os toques.

EP Não têm escolha, irão comprar o meu código.

EP As pessoas sentem-se como coisas.

EP Terás que experimentá-las, jogar com elas, avaliá-las até que o factor novidade se esgote, antes que tenhas que devolvê-las.

EP Com as vossas vozes nele.

EP E nenhum de vocês está a acenar em confirmação, porque morreu.

Adapted to EP Tu podes escolher em tudo o que tu compras, entre coisas que têm um efeito duro ou muito menos impacto.

Adapted to EP Eu acho que tu não entendes .

Adapted to EP A cada maré que vem e vai, tu encontras mais conchas.

Adapted to EP Pode muito bem ser o design daquilo que tu estás a usar.

Adapted to EP E a beleza disso é que tu não precisas formar pastilhas.

Adapted to EP E na verdade, se tu prestares atenção, bem aqui, dá para ver que diz, U.S. Peace Corps.

Adapted to EP E o homem disse-o: desceu bastante, não foi?

Adapted to EP E deve transformar-se numa aplicação matadora que permitir-nos-á continuar a usar os combustíveis fósseis de uma forma segura.

Adapted to EP E as pastilhas esfarelavam-se um pouco.

Adapted to EP Então isto torna-se cada vez mais um desafio.

Adapted to EP Transforme-se num catalizador de mudança.

Adapted to EP Eu vou mostrar-lhes algumas novas imagens e vou recapitular só quatro ou cinco.

Adapted to EP E a verdade é que, por anos eu senti-me um pouco deprimido.

Adapted to EP O que eu faço é, eu tiro o cartão de memória, dobro-o no meio, e a conexão da porta USB aparece.

Adapted to EP E parte do problema é que, ironicamente, a indústria passou tanto tempo a pesquisar como tornar as coisas fáceis de usar.

Adapted to EP E nenhum de vocês parece estar a reconhecer.

Adapted to EP As vezes é uma porcaria, mas a imprensa está a cobrir.

Adapted to EP As coisas estão a chegar num ritmo mais rápido.

Adapted to EP E há sinais de que a indústria está a entender o recado.

Adapted to EP Ela anotou o nosso pedido e foi atender um casal.

Adapted to EP E com certeza todos vocês já fizeram algo parecido uma vez na vida, ou então os seus filhos.

Adapted to EP E o melhor de tudo, a sua motivação: simplicidade vende.

Adapted to EP Então é aqui que estamos a seguir adiante com o nosso projecto de carvão.

Adapted to EP Bem aqui está o meu laptop de 100 dólares.

Adapted to EP Uma das coisas mais excitantes é pegar a desinfecção de água através do sol, e melhorar a habilidade de fazer isso.

Adapted to EP Use o comboio eléctrico.

Adapted to EP Mas quero colocar isso em perspectiva.

Adapted to EP E dentro de 12 horas, havia 700 mensagens de leitores na secção de comentários do site do New York Times, desde utilizadores que diziam Eu também!

Adapted to EP Crianças e mulheres são especialmente afectadas por isso, por que eles são os que ficam ao redor dos fornos e das fogueiras.

Adapted to EP Mas há mais poluição que causa aquecimento de construções do que de carros e camiões.

Adapted to EP Isolamento, melhor design, comprar electricidade verde onde tu puderes.

Adapted to EP E se tu és uma das pessoas que projectam as coisas, fácil é difícil.

Adapted to EP É mais parecido com associar-se a um clube, onde tu pagas anuidade regularmente.

Adapted to EP Isso não significa que se tu és um Republicano eu estou a convencer-te a ser um Democracta.

Adapted to EP Ele parabenizou-me, e eu respondi Obrigado, o que tu fazes aqui?

Adapted to EP Vejam bem, eu não estou a dizer que a Apple é a única empresa que adoptou a filosofia da simplicidade.

Adapted to EP Mas também há uma nova Microsoft que está realmente a fazer interfaces simples, e óptimas.

Adapted to EP A plateia foi à loucura, mas eu estava a pensar, onde será que eu já vi isso?

Adapted to EP Só o que tu vês é um adolescente que começou a trabalhar na sua garagem com apenas um amigo chamado Woz.

Adapted to EP Tu tens que estar a olhar para cinco ou dez anos adiante.

Adapted to EP Existe uma pequena empresa que está se a sair muito bem com essa história de simplicidade e elegância.

Adapted to EP Essa filosofia de fazer as coisas do jeito certo está a começar a espalhar-se.

Adapted to EP Para que as comunidades possam manter o seus próprios sistemas de água, saber quando eles estão a funcionar, quando é necessário tratá-los, etc.

Adapted to EP Então é aqui que estamos a seguir adiante com o nosso projecto de carvão.

Adapted to EP Eu adoraria ter o seu aconselhamento e ajuda em como dizer isso de forma a conectar-me com mais pessoas.

Adapted to EP Com ela, tu podes calcular precisamente as suas emissões de CO 2 e então tu recebes opções para reduzi-las.

Adapted to EP Torne-se activo politicamente.

Adapted to EP Não apenas esta, mas conectados com as ideias que estão aqui, para fazê-las mais coerentes.

Adapted to EP As pessoas gostam de rodear-se de capacidades desnecessárias, correcto?

Adapted to EP E embora fosse um óptimo uso para papelada proveniente do governo esse voluntário trazer-los junto com ele para sua vila, ela estava distante 800 quilómetros.