Universidade de Aveiro
2008
Departamento de Electrónica, Telecomunicações e Informática

CIRO ALEXANDRE DOMINGUES MARTINS

MODELOS DE LINGUAGEM DINÂMICOS PARA O PORTUGUÊS EUROPEU

Dynamic Language Modeling for European Portuguese

Universidade de Aveiro
2008
Departamento de Electrónica, Telecomunicações e Informática

CIRO ALEXANDRE DOMINGUES MARTINS

MODELOS DE LINGUAGEM DINÂMICOS PARA O PORTUGUÊS EUROPEU

Dynamic Language Modeling for European Portuguese

dissertação apresentada à Universidade de Aveiro para cumprimento dos requisitos necessários à obtenção do grau de Doutor em Engenharia Informática, realizada sob a orientação científica do Prof. Doutor João Paulo da Silva Neto, Professor Auxiliar do Departamento de Engenharia Electrotécnica e de Computadores do Instituto Superior Técnico, e do Prof. Doutor António Joaquim da Silva Teixeira, Professor Auxiliar do Departamento de Electrónica, Telecomunicações e Informática da Universidade de Aveiro

Apoio financeiro da FCT no âmbito do III Quadro Comunitário de Apoio.

Dedico este trabalho à minha família.

o júri

presidente

Doutor Manuel Augusto Marques da Silva
Professor Catedrático da Universidade de Aveiro

Doutora Tanja Schultz
Full Professor da Universidade de Karlsruhe, Alemanha e Assistant Research Professor na Language Technologies Institute (LTI) da School of Computer Science de CMU

Doutora Isabel Maria Martins Trancoso
Professora Catedrática do Departamento de Engenharia Electrotécnica e de Computadores do Instituto Superior Técnico da Universidade Técnica Lisboa

Doutor Francisco António Cardoso Vaz
Professor Catedrático Aposentado da Universidade de Aveiro

Doutor João Paulo da Silva Neto (orientador)
Professor Auxiliar do Departamento de Engenharia Electrotécnica e de Computadores do Instituto Superior Técnico da Universidade Técnica Lisboa

Doutor António Joaquim da Silva Teixeira (co-orientador)
Professor Auxiliar da Universidade de Aveiro

agradecimentos

This work was partially funded by PRIME National Project TECNOVOZ number 03/165 and by the FCT project POSC/PLP/58697/2004. The author was sponsored by a FCT scholarship (SFRH/BD/23360/2005).

This thesis would not have been possible without the motivation, support and encouragement that I have received since I decided to start my PhD degree. So, I would like to take this opportunity to thank and acknowledge to those who have helped me in this work.

First of all, I am grateful to my advisors Prof. António Teixeira and Prof. João Neto whose flexible advisor allowed me to pursue my own paths in this thesis, and with whom it has been a pleasure to work. I thank them for their help and patient. I special thank to Prof. Franscisco Vaz for all his help and support.

I would also like to thank all the researchers with whom I had directly or indirectly worked, and who have given me precious support. A special thanks to my colleague Hugo Meinedo for all his support, patient and efforts to make available all the manual transcriptions for the evaluation datasets used in this thesis.

I would also like to thank Department of Electronics, Telecommunications & Informatics/IEETA and L2F laboratory, which gave me all the technical resources and work environment to develop this work.

I express much gratitude and love for my family (my father, mother and sister) who always supported my decision to pursue this academic degree.

Thanks to all.

palavras-chave

Modelos de Linguagem, Informação Morfo-sintáctica, Reconhecimento Automático de Fala, Transmissões Noticiosas, Selecção de Vocabulários Adaptação de Modelos de Linguagem, Técnicas de Extracção de Informação.

resumo

Actualmente muitas das metodologias utilizadas para transcrição e indexação de transmissões noticiosas são baseadas em processos manuais. Com o processamento e transcrição deste tipo de dados os prestadores de serviços noticiosos procuram extrair informação semântica que permita a sua interpretação, sumarização, indexação e posterior disseminação selectiva. Pelo que, o desenvolvimento e implementação de técnicas automáticas para suporte deste tipo de tarefas têm suscitado ao longo dos últimos anos o interesse pela utilização de sistemas de reconhecimento automático de fala. Contudo, as especificidades que caracterizam este tipo de tarefas, nomeadamente a diversidade de tópicos presentes nos blocos de notícias, originam um elevado número de ocorrência de novas palavras não incluídas no vocabulário finito do sistema de reconhecimento, o que se traduz negativamente na qualidade das transcrições automáticas produzidas pelo mesmo. Para línguas altamente flexivas, como é o caso do Português Europeu, este problema torna-se ainda mais relevante.

Para colmatar este tipo de problemas no sistema de reconhecimento, várias abordagens podem ser exploradas: a utilização de informações específicas de cada um dos blocos noticiosos a ser transcrito, como por exemplo os *scripts* previamente produzidos pelo *pivot* e restantes jornalistas, e outro tipo de fontes como notícias escritas diariamente disponibilizadas na Internet.

Este trabalho engloba essencialmente três contribuições: um novo algoritmo para selecção e optimização do vocabulário, utilizando informação morfo-sintáctica de forma a compensar as diferenças linguísticas existentes entre os diferentes conjuntos de dados; uma metodologia diária para adaptação dinâmica e não supervisionada do modelo de linguagem, utilizando múltiplos passos de reconhecimento; metodologia para inclusão de novas palavras no vocabulário do sistema, mesmo em situações de não existência de dados de adaptação e sem necessidade re-estimação global do modelo de linguagem.

keywords

Vocabulary Selection, Language Model Adaptation, Morpho-syntactic Knowledge, Information Retrieval Techniques, Automatic Speech Recognition, Broadcast News.

abstract

Most of today methods for transcription and indexation of broadcast audio data are manual. Broadcasters process thousands hours of audio and video data on a daily basis, in order to transcribe that data, to extract semantic information, and to interpret and summarize the content of those documents. The development of automatic and efficient support for these manual tasks has been a great challenge and over the last decade there has been a growing interest in the usage of automatic speech recognition as a tool to provide automatic transcription and indexation of broadcast news and random and relevant access to large broadcast news databases. However, due to the common topic changing over time which characterizes this kind of tasks, the appearance of new events leads to high out-of-vocabulary (OOV) word rates and consequently to degradation of recognition performance. This is especially true for highly inflected languages like the European Portuguese language.

Several innovative techniques can be exploited to reduce those errors. The use of news shows specific information, such as topic-based lexicons, pivot working script, and other sources such as the online written news daily available in the Internet can be added to the information sources employed by the automatic speech recognizer. In this thesis we are exploring the use of additional sources of information for vocabulary optimization and language model adaptation of a European Portuguese broadcast news transcription system.

Hence, this thesis has 3 different main contributions: a novel approach for vocabulary selection using Part-Of-Speech (POS) tags to compensate for word usage differences across the various training corpora; language model adaptation frameworks performed on a daily basis for single-stage and multi-stage recognition approaches; a new method for inclusion of new words in the system vocabulary without the need of additional data or language model retraining.

Contents

List of Figures

List of Tables

1

Introduction

Ever since humans started to interact with computers, research efforts have been done to allow communication between the two to occur in a more natural way. Over the last thirty years, devices like keyboards and mice have been used as the means of entering data and commands into computers. In the late 1990s it has become realistic to expect to be able to interact with machines in a more human-like way. With significant developments in Human Language Technologies users would like computers to be able to recognize their speech and to understand their language.

Automatic Speech Recognition (ASR) technology has been experiencing large advances over the last two decades. ASR technology has moved from speaker dependent and isolated digit recognition applications to speaker independent large vocabulary continuous speech recognition systems. It has been applied to various different practical applications, such as dictation systems, spoken dialog systems, broadcast news transcription systems, etc, allowing users to transcribe audio data automatically and in a natural way. The automatic recognition of spoken data reduces the computational time and the cost to transcribe audio data when compared with a manual transcription process. Over the last years, investigation and development of integrated speech recognition systems has been carried out by many research institutions and commercial companies like BBN, Cambridge, Carnegie Mellon University (CMU), IBM, LIMSI, Nuance, etc.

Since the middle of the 1990s, and especially with the rapid expansion and importance of the Internet, there has been a growing need for fast and automatic processing of many different multimedia contents (audio and video sources). There are many examples in

which speech recognition would be useful - for instance to provide transcripts of meetings, lectures, Broadcast News (BN) and Broadcast Conversations (BC) streams, etc. Every day thousands of TV and radio stations broadcast many hours of information (news, interviews, documentaries, etc.). In emerging applications such as News on Demand and Internet News Services, users would expect to actively explore the information by finding sections of content relevant to their targeted search, rather than by following someone else path through the data stream or by viewing a large chunk of pre-produced material. Using such applications, large audio and video databases can be searched with very little effort, reducing the time spent reading or listening to large amounts of data. Technologies that make the management and the access of multimedia archives easier are receiving more and more attention due to the increasing availability of large multimedia digital libraries. Technologies for audio transcription and indexing of multimedia archives are among the emerging technologies.

Most of today's methods for transcription and indexation of broadcast audio data are manual. Broadcasters process thousands hours of audio and video data on a daily basis, in order to transcribe that data, to extract semantic information, and to interpret and summarize the content of those documents. The development of automatic and efficient support for these manual tasks has been a great challenge and over the last decade there has been a growing interest in the usage of automatic speech recognition as a tool to provide automatic transcription and indexation of broadcast news and random and relevant access to large BN databases [Gauvain et al., 2001][Neto et al., 2003][Nguyen et al., 2005][Gales et al., 2006]. Moreover, different companies and institutions are pursuing research on "human-friendly broadcasting services" to ensure that elderly viewers and people with visual or hearing impairments can enjoy those services.

Transcribing audio data is a necessary step in order to provide access to BN content and large vocabulary continuous speech recognition is a key technology for automatic processing. Commonly, broadcast news transcription systems have two main components (figure 1.1): an audio partitioner and a speech recognizer. The goal of audio pre-processing is to divide the acoustic signal into homogeneous segments, labeling and structuring the acoustic content of the data, and identifying and removing non-speech segments. For each speech segment, the ASR component determines the sequence of words in the segment, associating start and end times with each one of them.

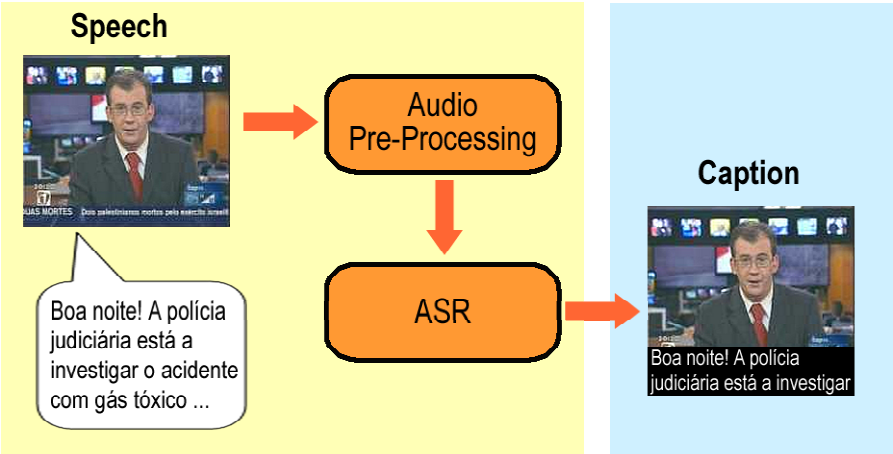

Figure 1.1: General architecture of a broadcast news transcription system.

This chapter gives some insight into automatic speech recognition and language modeling, briefly describing the Broadcast News transcription system used in this thesis. Next, we explain our motivation for the thesis work. Finally, the chapter concludes with an outline of our main contributions and a brief description of how this document is organized.

1.1 Automatic Speech Recognition

Speech recognition is concerned with the process of converting an acoustic signal containing speech data into the appropriate text transcription (figure 1.2).

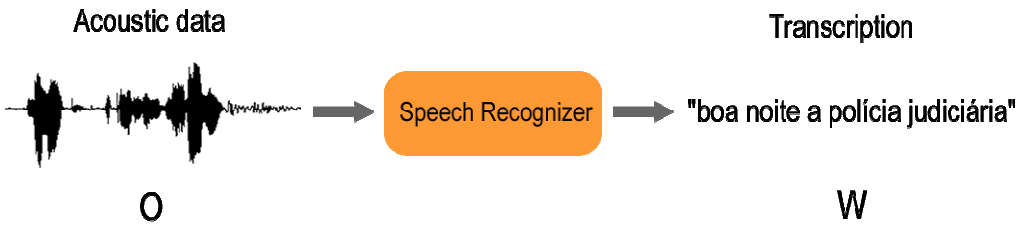

Figure 1.2: ASR process.

In the mainstream statistical formulation of the speech recognition problem [Bahl et al., 1983] the recognizer seeks to find the most probable word string, W , given an acoustic

observation, O, by computing the probability $P(W|O)$ for all possible sentences and choosing the sentence, $\hat{W}$, which produces the highest probability

$$\hat{W} = \arg\max_{W} P(W|O) \tag{1.1}$$

To model $P(W|O)$ in this probabilistic framework, a variety of assumptions must be made. In a first assumption, words are typically decomposed into sequences of phonetic units (or phones) representing the specific sounds used to distinguish between different words. For example, the Portuguese word *dia* (day) contains the phones /d/, /i/, and /A/. By applying Bayes' theorem and decomposing the sequence of words W into a sequence of small phonetic units U, the problem is reduced to finding $\hat{W}$ such that

$$\hat{W} = \arg\max_{W} P(W|O) = \arg\max_{W,U} P(O|U)P(U|W)P(W) \tag{1.2}$$

Hence, in the speech recognition problem there are four broad sub-problems to be solved:

- decide on a feature extraction algorithm and model the channel probability $P(O|U)$ - commonly referred to as *acoustic modeling*;
- model the source probability $P(U|W)$ referred to as the *lexical pronunciation model*
- model the source probability $P(W)$ commonly referred to as *language modeling*;
- *search* over all possible word strings W that could have given rise to O, finding out the most likely one $\hat{W}$.

The aim of this work is concerned with statistical language models and their use in a system for transcription of broadcast news data. In the next section, we will give an overview of the basic language modeling problem.

1.2 Language Modeling Basics

The task of language modeling is to assign a probability value to every possible word in a text stream based on its likelihood of occurrence in the context in which it finds itself. Let W be a sequence of n words, i.e. $W = w_1, \ldots, w_n$. Hence, the source probability $P(W)$ is approximated by

$$P(W) = \prod_{i=1}^{n} P\left(w_i \middle| w_1, \ldots, w_{i-1}\right) \tag{1.3}$$

where $P\left(w_i \middle| w_1, \ldots, w_{i-1}\right)$ is the probability that the word w_i has been spoken immediately following the preceding word sequence $(w_1, \ldots, w_{i-1})$. In this case, the word string $(w_1, \ldots, w_{i-1})$ is usually referred to as the *history* of the word w_i. In general, this probability $P(W)$ is estimated by examining large corpora of text for patterns and regularities, in a process known as *training*.

A first choice to face when constructing language models is the vocabulary V in which the w_i symbols take value. For practical purposes one has to limit the size of the vocabulary. A common choice is to use a finite set of words V. A second choice is the type of source model to be used. In fact, it is not feasible to compute the probability of a word given a long history of words. It should be noted that for a vocabulary of size $|V|$ there are $|V|^{i-1}$ possible distinct histories and $|V|^{i}$ values are needed for complete specification of probabilities $P\left(w_i \middle| w_1, \ldots, w_{i-1}\right)$. Even for practical vocabulary sizes such an astronomical number of estimates can neither be stored nor accessed in an efficient way. For this reason, word histories $(w_1, \ldots, w_{i-1})$ are partitioned into *equivalence classes* $K_1, \ldots, K_m$ such that each possible word history belongs to one and only one equivalence class. Hence, if $(w_1, \ldots, w_{i-1}) \in K_t$, we have

$$P\left(w_i \middle| w_1, \ldots, w_{i-1}\right) = P\left(w_i \middle| K_t\right) \tag{1.4}$$

The most common method of partitioning the word histories is by the use of *n*-grams, where the histories are partitioned according to their final $N-1$ words. Currently the most successful model assumes a Markov source of a given order N leading to the called *n-gram language model* [Rabiner and Juang, 1993][Jelinek, 1997]:

$$P\left(w_i \middle| w_1,\ldots,w_{i-1}\right) = P\left(w_i \middle| w_{i-N+1},\ldots,w_{i-1}\right) \tag{1.5}$$

The *n*-gram model attempts to capture the syntactic and semantic linguistic constraints by estimating the probability of a word in a sentence given its preceding $N-1$ words. The word string $w_{i-N+1},\ldots,w_{i-1}$ is usually referred to as history (h) of word w_i. The *n*-gram probability estimates can then be computed during the training process using the relative word frequencies, estimated according to the *maximum likelihood* method [Ney et al., 1997] as follows:

$$P\left(w_i \middle| w_{i-N+1},\ldots,w_{i-1}\right) = \frac{C\left(w_{i-N+1},\ldots,w_i\right)}{C\left(w_{i-N+1},\ldots,w_{i-1}\right)} \tag{1.6}$$

where $C\left(w_{i-N+1},\ldots,w_i\right)$ is the frequency at which $\left(w_{i-N+1},\ldots,w_i\right)$ occurs in the training corpora. The most widely used *n*-gram models are obtained for $N = 2$ (bi-grams), $N = 3$ (tri-grams) and $N = 4$ (four-grams).

However, sometimes there are some mismatches between training and testing data, for example different tasks. For some tasks it can be very hard and expansive, and time consuming to obtain a sufficient amount of task related training data. The task of broadcast news transcription is a typical example. In fact, there are some significant differences in language modeling between the broadcast news speech transcripts and written texts collected for instance from newspapers. In these cases, language model adaptation provides a mean to deal with these mismatches. Using a certain amount of task specific data to adapt the ASR language model component, two or more language models trained on different datasets can be mixed to produce a task adapted language model. In chapter two we give an overview of the state of the art in terms of language model adaptation

strategies, high-lighting the special case of broadcast news transcription systems and their evaluation measures.

In the next section, we briefly describe the overall system we used and which we improved with the work done in this thesis.

1.3 BN Transcription System for European Portuguese

This thesis presents part of the work done in the update and improvement of a fully functional prototype system for the selective dissemination of multimedia information [Meinedo et al., 2003][Meinedo, 2008]. This system was developed for BN data, specifically for European Portuguese TV news shows, being currently deployed in a real-life application. It is daily running since May 2002, successfully processing the 8 o'clock evening news of the Portuguese public TV broadcast company (RTP), and sending alert messages for registered users (http://ssnt.l2f.inesc-id.pt). The system automatically recognizes an announcer's speech, allowing closed-captioning to be created live and in real-time for that TV broadcaster (RTP).

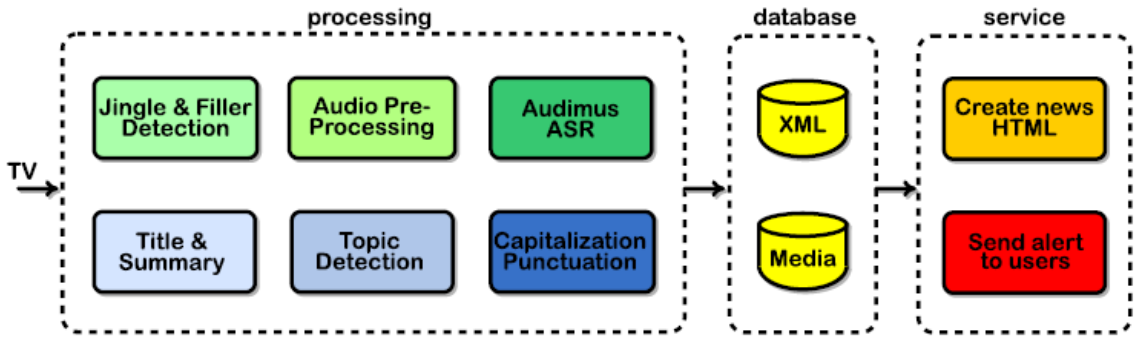

Figure 1.3: Media monitoring system.
(extracted from [Meinedo, 2008])

This media monitoring system is composed by several modules that allow to index and catalogue BN data (see figure 1.3). The system has a set of news shows to monitor from

different BN TV stations and a set of registered users each one with a profile regarding the news topics that are of his/her interest. After processing a news show, the media monitoring system compares the topics automatically detected for each story identified in that news show against the users profiles. If matches are detected then it sends alert emails to the corresponding users with the title, short summary and video link to the relevant news stories detected.

The most important modules of this system are the Audio Pre-Processing (APP), Automatic Speech Recognition (ASR) and Topic Segmentation and Indexation (Topic Detection, TD). This thesis describes the work done to specifically update and improve the ASR module. Hence, in chapter three we give a short overview of its baseline system. Further details about the overall system and other modules can be found in the PhD work presented in [Meinedo, 2008].

1.4 Motivation

The daily and real-time transcription of broadcast news is a challenging task both in acoustic and in language modeling. To achieve optimal performance in news transcription, several problems have to be overcome: variety of acoustic conditions (signal quality, environmental noise, music), variety of speakers (news anchors, interviews with a variety of speakers, outside studio reporters, etc.), many different speaking styles (from spontaneous conversation to prepared speech close in style to written texts), and topic changing over time leading to unlimited vocabulary and many new topics appearing everyday.

Even though the linguistic properties of broadcast news data change over time, most ASR components use static language models with the vocabulary selected from a large and fixed training corpus, which was the case of the BN transcription system described in section 1.3 and used in our work. This means there is usually a significant gap between the epoch of the LM training process and the use of that system for audio data processing. However, the broadcast news domain is characterized by rapid changes in topic, which means changes in vocabulary items and linguistic styles to be recognized and transcript. Particularly, when transcribing broadcast news data in highly inflected languages, like the

European Portuguese one, the vocabulary growth leads to high out-of-vocabulary (OOV) word rates (here, OOV means words that are not included in the recognizer lexicon, with the OOV word rate defined as the ratio between the number of words not covered by the vocabulary and the total number of words in the recognized texts). Hence, the structure of language indirectly influences speech recognition efficiency.

The European Portuguese language shares its characteristics with many other inflectional languages, especially those of the Romance family. European Portuguese words often exhibit clearer morphological patterns in comparison to English words. *Morpheme* is the smallest part of a word with its own meaning. In order to form different morphological patterns (derivations, conjugations, gender, number inflections, etc.), two parts of a word are distinguished: *stem* and *ending*. *Stem* is the part of the inflected word that carries its meaning, while the *ending* specifically denotes categories of person, gender and number, or the final part of a word, regardless of its morphemic structure. To outline the European Portuguese language characteristics, we show in table 1.1 an example of two semantically equal sentences differing in subject gender, but being identical in English.

European Portuguese	**masculine**	**feminine**
	O meu amigo é professor	A minha amiga é professora

English	**undefined**
	My friend is a teacher

Table 1.1: An example of two semantically equal sentences differing in subject gender, but being identical in English.

European Portuguese language distinguishes between three types of gender: masculine, feminine and neuter, while English only has one form. All nouns, adjectives and verbs in European Portuguese have a gender. They present far more variant forms than their English counterparts. Words have augmentative, diminutive and superlative forms (e.g. "small house" = *casinha*, where *–inha* is the suffix that indicates a diminutive). Moreover, European Portuguese is a very rich language in terms of verbal forms. While the regular verbs in English have just 4 variations (e.g. talk, talks, talked, talking), the European Portuguese regular verbs have over 50 different forms, with each one having its specific

suffix [Orengo and Huyck, 2001]. The verbs can vary according to gender, person, number, tense and mood. Three types for the grammatical category of person (1st, 2nd, 3rd person) reflect the relationship between communication participants. There are five tenses: present, past, past perfect, past imperfect, past pluperfect and future. Another grammatical category, mood, denotes the feeling of the speaker towards the act, which is defined by the verb. There are eight different types of mood in European Portuguese: indicative, subjunctive, imperative, conditional, infinitive, inflected infinitive, participle, and gerund.

The rich morphology of the European Portuguese language causes a large number of possible words, which in turn decreases the quality of language models (higher OOV rates). To illustrate this peculiarity of the European Portuguese, we plot in figure 1.4 the vocabulary growth for two BN corpora: an European Portuguese BN corpus used in this thesis (ALERT-SR corpus) and consisting of about 500K word tokens and a subset of the 1997 English Broadcast News Speech corpus (HUB4) with the same size. As one can observe, for the European Portuguese corpus the vocabulary growth is faster than for the English one. For a corpus size of about 500K word tokens, HUB4 subset has a vocabulary size of 19K words, while the vocabulary size for the ALERT-SR corpus is 26K, i.e. about 37% more.

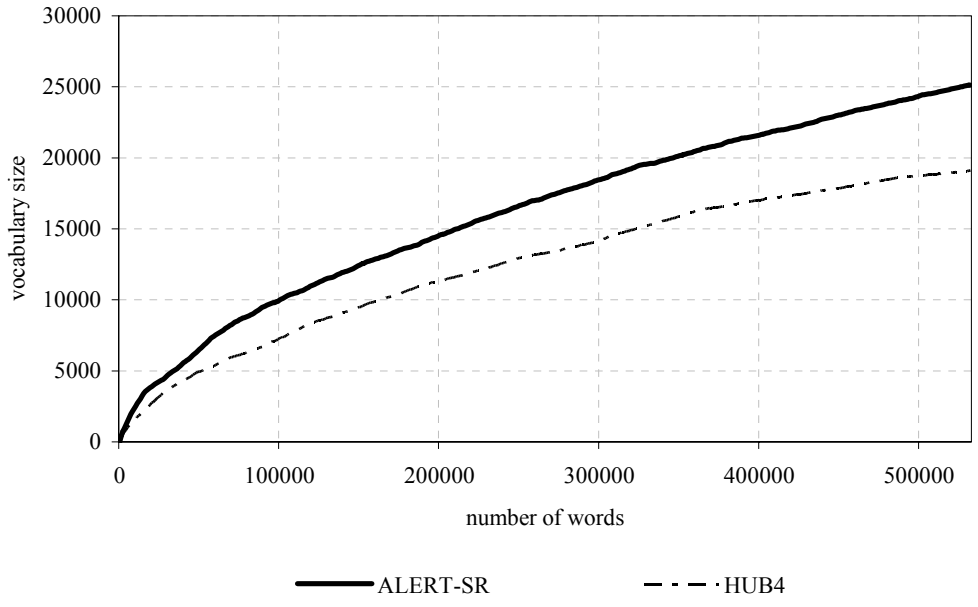

Figure 1.4: Vocabulary growth comparison between two Broadcast News corpora: the 1997 English BN Speech corpus (HUB4) and the European Portuguese BN corpus (ALERT-SR).

European Portuguese demands a larger vocabulary to get the same degree of text corpus coverage than for English. It contains many different word forms, all derived from the same basis (lemma). This property has already been studied and used in language modeling for European Portuguese in the author Master's thesis [Martins, 1998].

From the work we have done in [Martins et al., 2005] we could derive a qualitative performance analysis which indicated some type of recognition errors present in the BN transcription system used in this thesis and briefly described in section 1.3:

- Errors due to *speech disfluencies*, especially common in spontaneous speech, which is frequent in our BN corpora. The frequency of disfluencies is very high for this style of speech, with values of about 20% as cited in [Shriberg, 2005]. In our BN corpora only filled pauses account for about 2% of total words.
- Errors due to *insufficient or incorrect phonetic transcriptions*. Some automatic phonetic transcriptions in the pronunciation dictionary are incomplete or were incorrectly produced. This occurs mainly in case of foreign words, whose automatic transcriptions are not reliable ("Al-qaeda" is an example of a foreign word whose automatically generated phonetic transcription is incorrect.
- Errors due to *inconsistent spelling of orthographic transcriptions* both in BN manual transcriptions and written news. The most common inconsistencies occur for foreign names (for example, "Madeleine" and "Medeleine"), or consists of writing the same word entities both as separate words and as a single word (for example, "secretário geral" and "secretário-geral").
- Errors due to *out-of-vocabulary (OOV) words*. On average, one OOV word could cause more than one error, with an average rate of 1.5 additional errors [Hetherington, 1995].

For a live running closed-captioning service for news programs, like the one currently based on the broadcast news transcription system used in this work, the ability to correctly address new words appearing on a daily basis, mainly name entities, is an important factor to take into account for its performance. Hence, the problem of OOV words is a common and important one.

For a task like broadcast news transcription, it is practically impossible to define in advance and in a static way a word vocabulary that could cover all words that may appear in a news show, which means the system is constantly faced with new words spoken by different users. Typical examples of such words are proper and common names. However, the magnitude of the problem mainly depends both on the vocabulary size and the temporal gap between the training data epoch used to construct that vocabulary and the speech it is used on.

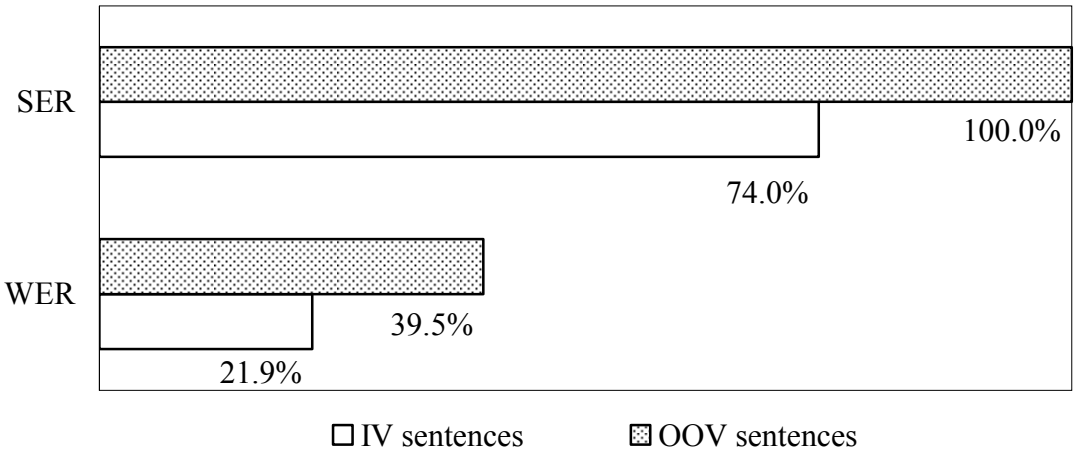

Figure 1.5: Word Error Rate (WER) and Sentence Error Rate (SER) for in-vocabulary (IV) and out-of-vocabulary (OOV) sentences.

Figure 1.5 shows the importance of the OOV problem in our BN transcription system. As one can observe, the word error rate (WER) is almost two times higher for sentences with OOV words than it is for in-vocabulary (IV) sentences. The increase in WER for OOV sentences can be related to three factors [Jurafsky and Martin, 2000]. The first one and most obvious is the fact that OOV words are wrongly recognized since they are not in the lexicon. The second one is the fact that some errors resulting from OOV words correspond to in-vocabulary words being wrongly recognized due to their proximity to unknown words. The third factor is related to the high correlation between out-of-domain sentences and sentences containing OOV words whose n-gram sequences tend to be out-of-domain and consequently harder to be correctly recognized. In the same figure, one can observe the sentence error rate (SER) too. By SER one means the percentage of sentences with at least one recognition error, and of course for OOV sentences this value is always

100%. Moreover, even though the use of a very large vocabulary in the ASR component could reduce the OOV word rate in broadcast news data, the number of sentences having at least one OOV word is still high, approximately 10-15% with a 57K vocabulary for English broadcast news data [Palmer et al., 2005]. However, for highly inflected languages like European Portuguese, that number tends to be even higher. For our BN data it is about 26%, i.e. a quarter of the sentences have at least one OOV word.

In this thesis, we address the problem of language model adaptation over time, by proposing and developing various frameworks for vocabulary selection and language model adaptation for the European Portuguese. In the next section we summarize the proposed adaptation methods.

1.5 Contributions

The aim of this thesis is to develop automatic language model adaptation frameworks for a European Portuguese broadcast news transcription system. Language model adaptation consists of selecting an appropriate word list to include in the recognizer lexicon and reducing the mismatch in language model due to possible linguistic and temporal gaps between the training corpora and the BN data to be processed.

We propose a novel approach for vocabulary selection, extensible to any number of available training corpora. This approach relies on using Part-Of-Speech (POS) tags to compensate for word usage differences across the various training corpora. Its performance is compared in terms of expected OOV word rate with the performance of the conventional and most used word frequency based approach, showing to be more efficient especially for selection of large-sized vocabularies as in case of BN transcription task.

To adapt the ASR language model component of our BN transcription system, we propose a daily and unsupervised adaptation approach which is done in a multi-stage recognition framework. The idea of vocabulary and language model adaptation is to use written news daily available on the Internet to be able to model the lexical content of the news, reducing the impact of linguistic differences over time.

In a first stage, using the proposed vocabulary selection algorithm and written news collected from the Internet, the vocabulary and language model of the ASR component are

adapted on a daily basis to be used by the live closed-captioning system. In a second stage and using the recognition results from the first stage, a new adaptation step is performed which dynamically adapts the active vocabulary and language model to the topic of the current news segment. Hence, based on the texts collected on the Web, a story-based vocabulary is selected using again the proposed vocabulary selection algorithm. Using an Information Retrieval engine [Strohman et al., 2005] and the ASR hypotheses generated on the first stage as query material, relevant documents are extracted from a dynamic and large-size dataset to generate a story-based language model. Since those hypotheses are quite small and may contain recognition errors, a relevance feedback method for automatic query expansion is used [Lavrenko et al., 2001].

Finally, the proposed LM adaptation framework is complemented with a new method that allows including new words in the system vocabulary without the need of additional adaptation data or language model retraining. This method uses morpho-syntatic information about an in-domain corpus and part-of-speech (POS) word classes to define a new language model unigram distribution associated to the updated system vocabulary.

1.5.1 Published Results

During this work we published a number of articles describing the work done in vocabulary selection algorithms and in language modeling adaptation for ASR, specifically applied to a Broadcast News Transcription system for the European Portuguese language. A complete list of articles published in international conferences and one National Journal is given bellow:

- Article describing the work done with the updating and improvement of the language model component of a continuous speech recognition system for the European Portuguese and integrated in the BN transcription system used in this thesis. Two sources of performance improvement have been studied: the inclusion of more training data to better estimate the language model parameters, and the use of different discounting and pruning techniques.

Martins, C., Teixeira, A. and Neto, J. (2005). "Language Models in Automatic Speech Recognition". In Revista do Departamento de Electrónica e Telecomunicações da Universidade de Aveiro.

- Article proposing a daily vocabulary and LM adaptation framework which directly extracts new words based on contemporary written news available on the Internet and some linguistic knowledge (lemmas-based) about the words found on those news.

  Martins, C., Teixeira, A., and Neto, J. (2006). "Dynamic Vocabulary Adaptation for a daily and real-time Broadcast News Transcription System". In Proceedings of IEEE/ACL Workshop on Spoken Language Technology, Aruba.

- Article introducing a modified vocabulary selection technique which uses part-of-speech (POS) word classification to compensate for word usage differences across the various training corpora. This approach is based on the hypothesis that the similarities between different domains can be characterized in terms of style (represented by the POS sequences).

  Martins, C., Teixeira, A., and Neto, J. (2007). "Vocabulary Selection for a Broadcast News Transcription System using a Morpho-syntatic Approach". In Proceedings of Interspeech 2007, Antwerp, Belgium.

- Article proposing a daily and unsupervised adaptation approach which dynamically adapts the active vocabulary and LM to the topic of the current news segment during a multi-pass speech recognition process. Based on texts daily available on the Web, a story-based vocabulary is selected using the morpho-syntatic technique previously introduced. Using an Information Retrieval engine, relevant documents are extracted from a large corpus to generate a story-based LM.

Martins, C., Teixeira, A., and Neto, J. (2007). "Dynamic Language Modeling for a Daily Broadcast News Transcription System". In Proceedings of ASRU 2007, Kyoto, Japan.

- Article presenting the evaluation and comparison of the two previously proposed vocabulary adaptation approaches and their integration into the multi-pass LM adaptation framework.

  Martins, C., Teixeira, A., and Neto, J. (2008). "Dynamic Language Modeling for the European Portuguese". In Proceedings of PROPOR 2008, Curia, Portugal.

- Article proposing a new method that allows including new words in the vocabulary even if no well suited training data is available, as is the case of archived documents, and without the need of LM retraining. It uses morpho-syntactic information about an in-domain corpus and part-of-speech word classes to define a new LM unigram distribution associated to the updated vocabulary.

  Martins, C., Teixeira, A., and Neto, J. (2008). "Automatic Estimation of Language Model parameters for unseen Words using Morpho-syntactic Contextual Information". In Proceedings of InterSpeech 2008, Brisbane, Australia.

1.6 Outline

The remainder of this thesis is organized in seven chapters. Following is a brief description of each chapter:

Chapter 2: State of the Art

This chapter is intended to provide the basic background needed throughout the thesis and the related state of the art. A short overview of language modeling and its use for speech recognition is given. We present a survey of approaches to the vocabulary and language model adaptation problem, with special focus on Broadcast News/Conversations tasks. The

chapter also provides a short overview of Information Retrieval Techniques (IR) and their use in the field of ASR.

Chapter 3: Resources and Baseline System

Chapter 3 provides a description of the fundamental resources used in this thesis for training and evaluation of all the proposed algorithms and adaptation approaches. The chapter briefly describes the ASR baseline system (AUDIMUS.media) used in our work. It also gives an overview of the remaining processing tools used.

Chapter 4: Vocabulary Selection

This chapter describes in detail the work done in the scope of our thesis, where we are exploring the use of additional sources of information for vocabulary adaptation of a European Portuguese broadcast news transcription system. Since the vocabulary optimization problem is mainly dependant on the specific linguistic characteristics of the target language, we present an analysis of the vocabulary growth, coverage and OOV words for the European Portuguese language using the datasets described in chapter 3. Based on that analysis and its conclusions, we devised new vocabulary selection approaches. Across this chapter a set of experiments on using these approaches are presented and their results reported and compared.

Chapter 5: Language Model Adaptation

This chapter presents the proposed dynamic vocabulary and language modeling adaptation framework, describing the experimental work carried out for this thesis. The chapter presents two approaches for language model adaptation. The first one is a single-stage recognition procedure which relies on the presented vocabulary selection algorithm. The second one is a multi-stage procedure using some Information Retrieval techniques on top of the new vocabulary selection algorithm to estimate the adapted language model. A set of experiments are presented and their results reported and compared to the baseline system. Finally, we briefly describe the integration and implementation of the proposed framework into a fully functional prototype system for the selective dissemination of multimedia information.

Chapter 6: Handling Unseen Words

To complement the approaches proposed in chapter 5, we describe in this chapter a new method that allows including new words in the vocabulary even if no well suited adaptation data is available, as is the case of archived documents. We conclude this chapter with an experimental evaluation of the proposed approach, drawing some conclusions at the end.

Chapter 7: Conclusions and Future Directions

Finally, chapter 7 presents the contributions and conclusions of the thesis. It concludes with a discussion of future directions.

2

State of the Art

The previous chapter introduced the basic concepts of language modeling and its application to large vocabulary speech recognition. In this chapter, we provide an overview of the theory of language modeling, outlining the traditional problems inherent in statistical language modeling and the techniques commonly used to overcome them. We present a brief summary of the state-of-art, giving an overview about the current state in terms of vocabulary and language model adaptation. Next, we describe the use of Information Retrieval (IR) techniques for speech recognition tasks, focusing on the techniques we used in the proposed adaptation framework. Finally, we summarize current approaches in terms of language models applied to the Broadcast News speech recognition task.

2.1 Language Modeling for Speech Recognition

Techniques for language modeling mainly fall into two categories. The first type of models are the traditional linguistic grammars, such as context free and unification based grammars [Cole et al., 1995], which although being rigorously defined from linguistic perspective, suffer from the typical limitations of rule-based systems: coverage, predictive power and computational requirements. Complex hand-built grammars often lack coverage of sentences structures going beyond its given linguistic theory, being difficult to adapt to new domains and languages. Moreover, their computation complexity is too high to be

efficiently employed in time critical applications, such as large vocabulary continuous speech recognition.

The second language model category (data-oriented grammars) is based on a statistical representation of the natural language and has gained common usage. A statistical language model probabilistically describes the constraints on word order found in language: typical word sequences are assigned high probabilities, while atypical ones are assigned low probabilities. In the next sub-sections we give an overview about the vocabulary selection approaches, describing the mostly used statistical LM, the *n*-gram model, its major advantages and drawbacks, the *discounting and smoothing* techniques used to better estimate probabilities when there is insufficient data, and some of the alternative extensions to the *n*-gram language model which have been proposed by the research community. Finally, we give an overview about the current state in terms of language model adaptation, especially in case of the BN transcription task.

2.1.1 Vocabulary Selection/Adaptation

The first step in language model construction is the selection of the *vocabulary*, i.e. the set of words that can be recognized by the ASR component. The size and performance of a language model or speech recognition system are often strongly influenced by the size of its vocabulary. The most common approaches to vocabulary selection and optimization are typically based on word frequency, including words from each training corpus that exceed some empirically defined threshold, which mainly depends on the relevance of the corpus to the target task [Gauvain et al., 2002].

Although the vocabularies of ASR systems are designed to achieve high coverage for the expected domain, *out-of-vocabulary* (OOV) words cannot be avoided. Hence, depending on the way that the language model handles the occurrence of such OOV words, its vocabulary can be called to be either *open* or *closed*. A closed vocabulary model makes no provision for OOV words. So, those words will not be recognized and an error will be reported. On the other side, an open vocabulary model allows for OOV words to occur with those words being mapped to the same symbol, typically denoted as <*unk*>. In each case, every OOV word in the input utterance is guaranteed to result in one or more output errors. However, some of those recognition errors will affect applications performance, as

the case of incorrectly recognized spoken names in tasks like broadcast news transcription and/or indexation, where this type of content words play an important role for the overall systems performance. To deal with the OOV words problem different approaches have been suggested by the research community. These approaches can be classified into different categories [Bazzi, 2002]:

Vocabulary Optimization

One approach to the OOV word problem might be to choose the vocabulary in such a way to reduce the OOV word rate by as much as possible - *Vocabulary Optimization*. This optimization could either involve increasing the vocabulary size of the ASR component, or it could also be selecting those words most representative for the target domain/task. Speech recognition systems can have a range of vocabulary sizes, depending on the target domain, the generality required, as well as the availability of computational resources. For instance, current systems for unconstrained tasks such as the transcription/indexation of broadcast news programs frequently have vocabularies between 25,000 and 64,000 words or even more as in case of highly inflected languages. Increasing the vocabulary size of a speech recognition system can result in lower error rates, in part by decreasing the percentage of OOV words in the input utterance. However, systems with larger vocabularies require more memory and run slower than those with smaller vocabularies. In addition to increased computational cost, adding words to a vocabulary increases the potential confusability with other vocabulary words [Rosenfeld, 1995].

The need for unlimited language vocabulary despite a limited ASR vocabulary suggests an alternative in which rather than trying to include all possible words in the ASR vocabulary we instead develop techniques for dynamically adapting the overall system vocabulary – *dynamic vocabularies* - using lexical resources, without requiring a larger ASR vocabulary and the problems this entails. In [Geutner et al., 1998] an approach targeted at reducing the OOV word rates for heavily inflected languages is suggested. Their work uses a multi-pass recognition strategy to generate morphological variations of the list of all words in the lattice generated in a first-pass, thus dynamically adapting the recognition vocabulary for a second-pass of the speech recognition process. The basic idea of this so-called adaptation algorithm (HDLA – Hypothesis Driven Lexical Adaptation) is that a large number of words in the hypotheses generated with a baseline vocabulary are

wrongly recognized because only the inflectional ending is wrong whereas the stem was recognized correctly. Hence, all words with the same stem are then incorporated into the adapted vocabulary for a second recognition pass, with the new words replacing the least frequent ones that did not appeared in the first recognition pass. By applying this adaptation algorithm both on Serbo-Croatian and German news data, OOV word rates were reduced by 35-45%.

In [Venkataraman and Wang, 2003] the authors propose and evaluate three different methods for selecting a single vocabulary from many corpora of varying origins, sizes and recencies such that the vocabulary is optimized for both size and OOV word rate in the target domain. They concluded that a maximum-likelihood-based approach is a robust way to select a domain's vocabulary especially when reasonable amounts of in-domain texts are available. In this approach, the normalized unigram counts of each word in each of the available training corpora are linearly interpolated, choosing the mixture coefficients which maximize the probability of the in-domain corpus. This technique scalable and extensible to any number of corpora showed to be robust especially for selection of small vocabularies.

More recently, researchers are using the Word Wide Web (WWW) as an additional resource of training data for dynamic vocabulary and language modeling adaptation procedures [Schwarm et al., 2004]. In [Federico and Bertoldi, 2004] the problem of updating over time the LM component of an Italian broadcast news transcription system is addressed. In particular, vocabulary adaptation, done on a daily basis, is carried out by adding words to the active vocabulary according to frequency and recency in contemporary written news, which allowed achieving significantly lower OOV word rates. Reported experiments showed a relative reduction of 58% in OOV word rate. The work presented in [Oger et al., 2008] suggests that the local context of the OOV words contains relevant information about them. Using that information and the Web, different methods were proposed to build locally-augmented lexicons which are used in a final local decoding pass. This technique allowed recovering 7.6% of the significant OOV words and the accuracy of the system was improved.

In [Allauzen and Gauvain, 2005a] an automatic adaptation method which makes use of contemporaneous texts available on the Internet to model the lexical and linguistic content of the news on a daily basis was proposed. A vectorial algorithm for vocabulary adaptation

is used which combines word frequencies vectors estimated on adaptation corpora to directly maximize lexical coverage on a development corpus, thus eliminating the need for human intervention during the vocabulary selection process. Authors' experiments showed a significant reduction of the OOV word rate compared with the baseline vocabulary: a relative decrease of 61 % in French and 56% in English.

A similar framework to the one presented in [Geutner et al., 1998] is proposed by Palmer and Ostendorf [Palmer and Ostendorf, 2005], but focusing on names rather morphological word differences. They proposed an approach for generating targeted name lists for candidate OOV words, which can be used in a second pass of recognition. The approach involves offline generation of a large list of names and online pruning of that list by using a phonetic distance to rank the items in a vocabulary list according to their similarity to the hypothesized word. Their reported experiments showed that OOV word coverage could be improved by nearly a factor or two with only 10% increase in the vocabulary size.

Finally, other approaches have been used for dynamic adaptation of vocabulary and/or language model to the topics present in a BN show using different Information Retrieval (IR) techniques to extract relevant documents from a large general corpus or from the Web for adaptation proposes [Bigi et al., 2004] [Chen et al., 2004] [Boulianne et al., 2006]. These multi-pass speech recognition approaches use the ASR hypotheses as queries to an IR system in order to select additional on-topic adaptation data.

Confidence Scoring

The use of confidence scoring measures to predict whether a recognized word is actually a substitution for an OOV word is another strategy to deal with the OOV words problem. The ability to estimate the confidence of the recognized hypothesis allows the ASR system to either reject all or part of the sentence if the confidence value is bellow some pre-defined threshold. However, this approach has some drawbacks too. Confidence measures can be good at predicting whether a hypothesized word is correct or not, but unable to differentiate between errors due to OOV words and those errors due to other problems such as degraded acoustical conditions. Moreover, confidence measures are only used to detect possible sentence segments with OOV words. To identify those OOV words other techniques must be used.

Different methods on how to estimate confidence scores to detect possible OOV words have been proposed (e.g. [Carpenter et al., 2001]) but the most interesting part is how to use these confidence scores to correct input segments detected as OOV words. In [Palmer and Ostendorf, 2005] an approach that involves the integration of word confidences into a probabilistic model, which can jointly identify names and errors, is used to improve OOV name resolution for applications in language processing, particularly speech recognition.

Multi-Stage Sub-Word Recognition

This strategy uses two or more stages during the recognition process [Rotovnik, 2004] [Creutz et al., 2007]. In a first stage, a sub-word recognition is performed and phonetic sequences are obtained. This way, the ASR system is able to hypothesize novel phonetic sequences which could potentially belong to OOV words. In the second stage, those sub-word sequences are mapped to word sequences using word-level information. This kind of strategies can involve many variations. For instance, the output from the first stage (which is passed to the second recognition stage) can be the single best hypothesis, the N best hypotheses, or a graph representing the search space. Moreover, the sub-word recognition can be done at different levels: phonetic level, syllable level, morpheme level, or even using automatically derived sub-word units. A major weakness of this approach derives from the fact that an important source of information, the word level constraints, is removed from the first stage, causing significant degradation in terms of WER for words in the vocabulary.

English has relatively little inflectional morphology (endings expressing case number and gender agreement) and prefers to express complex concepts as a phrase, or a hyphenated compound, rather than as a closed compound. However, other European languages exhibit a greater degree of compounding/inflection than English. A lexicon for this kind of languages needs to contain considerably more words in the ASR vocabulary than the English language in order to attain the same coverage [Gauvain et al., 2005]. To overcome this problem some ASR systems have been proposed which eschew the orthographic word as the basic unit of the language model, and instead choose morphemes or other sub-word units created through data driven processes. Most research on morpheme-based systems has been developed for inflected/compounded languages such as Turkish and Finish [Kurimo et al., 2006], German [Geutner et al., 1998], Slovenian

[Rotovnik et al., 2003], Check [Ircing and Psutka, 2002], Arabic [Choueiter et al., 2006] [Xiang et al., 2006], Korean [Kwon and Park, 2003], etc. These works report some improvements in terms of OOV word rates and/or WER, especially when combining these models with the standard ones based on words. The problem of acoustic confusability arises when larger sub-word vocabularies are used. In fact, sub-word language model units alleviate the problem of rapid vocabulary growth, especially for this kind of highly inflectional languages. But as the base units of the language model become fewer and smaller, the language model becomes less constrained and the acoustic confusability increases.

Filler Models

One of the most commonly used approaches for handling OOV words is the addition of a generic unknown word both in acoustic and language models – the so called *filler model*, sometimes known as *garbage model*. This generic word model competes during the recognition process with the remaining models of in-vocabulary words, with its presence in the hypothesis signaling the presence of an OOV word. As in the case of confidence scoring approach, filler models can potentially classify segments of the input signal corresponding to in-vocabulary words as OOV words.

In [Bazzi, 2002] the author proposes a novel approach for handling OOV words in a single-stage framework, which uses an explicit and detailed model of OOV words to augment the closed-vocabulary recognizer search space. The author explore several research issues related to designing the sub-word lexicon, language model and OOV word model topology in order to ensure that the OOV word model does not degrade system performance for in-vocabulary words. In [Hazen and Bazzi, 2001] a combination of this model with confidence scoring is given. In order to improve the transcription readability, an approach to transcribe OOV input segments identified as OOV words based on phoneme-to-grapheme conversion is presented in [Decadt et al., 2002].

OOV Word Class

In some approaches an open vocabulary language model is introduced by assuming a special OOV word class, where the addition of new words is done by extending that OOV

class and re-estimating its unigram distribution $P_{oov}(w)$. This unigram distribution takes into account the word frequencies of the adaptation texts.

In [Federico and Bertoldi, 2004], an approach of this kind was introduced. The baseline vocabulary of 62K words is extended by adding to that special OOV word class 60K new words selected from the contemporary written news on a daily basis.

In [Allauzen and Gauvain, 2005] another open vocabulary approach was reported. Special forms called back-off word classes (BOW) were used to introduce a word in the vocabulary without retraining the language model. Thus, during language model training one of these forms replaces one or more words which are not yet known, by discounting a mass of probability from the OOV words. Then, prior to decoding, new words can be added as alternate orthographic forms of these special classes. Words are linked with their lexical BOW according to their POS tag. An oracle experiment was performed to estimate an upper-bound on the gain that could be obtained with that method. This experiment was carried out by adding all the OOV words in the manual transcripts to the baseline vocabulary via their associated BOW. Reported results showed that about 80% of all new words introduced by this method were correctly recognized.

2.1.2 Word-based *n*-gram Models

In chapter 1 it was shown that the role of the language model in ASR is to provide an estimate of $P(W)$. Currently the most successful model assumes a Markov source of a given order N leading to the called n-gram model [Rabiner and Juang, 1993], in which an estimate of the likelihood of a word w_i is made solely on the identity of the preceding $N-1$ words in the sentence, with the choice of N being based on a trade-off between detail and reliability. Hence, this choice will be dependent on the quantity of training data available. For the quantities of training data typically available from newspapers written texts and BN data, N=4 (4-gram models) seems to be the best balance between precision and robustness for task like broadcast news transcription [Goodman, 2001].

The strengths of the *n*-gram model come from its success at capturing local syntactic and semantic constraints, from the simplicity of its training process, and from its computational efficiency within the recognition framework. One drawback, however, is

that the current word w_i is clearly dependent on much more words than the previous two or three words. In fact, it is easy to construct a nonsense sentence or at least an ungrammatical one, but which has a high probability according to a 3-gram or 4-gram language model. Table 2.1 shows an example of a recognized sentence which is ungrammatical, but consists of very plausible 4-grams. This drawback clearly shows the inability of *n*-gram language models to take into account the long-range syntactic and semantic dependencies.

Type	Transcription
Reference	a lei que regulamenta o código do trabalho foi APROVADA no parlamento
Hypothesis	a lei que regulamenta o código do trabalho foi APROVADO no parlamento

Table 2.1: Example showing the ASR output for a BN sentence and its correct transcription.

Of course, the most obvious extension to 4-gram models would be to simply move to higher order n-grams, such as 5-grams and so on. In [Goodman, 2001] it is shown that in fact, significant improvements can be gotten from moving to *n*-grams of higher order if sufficient training data and computational power is available. However, to overcome this limitation of *n*-gram models other extensions have been proposed. In the next sub-section we give an overview of them.

2.1.3 Extensions to Word-based *n*-gram Models

Class-based *n*-gram models

An extension to *n*-gram models are *Class-based n*-grams (also called *Clustering* models). Clustering models differ from standard *n*-gram models since they attempt to make use of the similarities between words to define a mapping of the vocabulary words into a smaller number of classes. For instance, if we have seen occurrences of sentences blocks like "novas eleições no sábado" and "novas eleições no domingo", then we might think that the

word “sexta-feira”, being semantically similar to both “sábado” and “domingo”, is also likely to follow the sentence block “novas eleições no”. Thus, the *n*-grams are then based on classes rather than on words.

There are different approaches concerning the problem of how to get the best classes (clusters). Some of those approaches are linguistically motivated, and correspond to the word's part-of-speech (POS). On other approaches classes are automatically derived from the language model training data using data-driven techniques. Many automatic clustering approaches have been investigated, for example in [Ney et al., 1994].

Another question related to class-based *n*-gram models is how to use those classes. For example, when dealing with a trigram, the class-based model could be defined in any of the following ways [Goodman, 2001]:

$$P\left(w_i \left| w_{i-2} w_{i-1}\right.\right) = P\left(w_i \left| c\left(w_{i-2}\right) c\left(w_{i-1}\right)\right.\right) \tag{2.1}$$

$$P\left(w_i \left| w_{i-2} w_{i-1}\right.\right) = P\left(w_i \left| c\left(w_{i-2}\right) w_{i-1}\right.\right) \tag{2.2}$$

$$P\left(w_i \left| w_{i-2} w_{i-1}\right.\right) = P\left(c\left(w_i\right) \left| c\left(w_{i-2}\right) c\left(w_{i-1}\right)\right.\right) P\left(w_i \left| c\left(w_i\right)\right.\right) \tag{2.3}$$

where $c\left(w_i\right)$ represents the cluster for word w_i.

Class-based *n*-gram models have several advantages over word-based n-gram models: much more compact models due to the reduction in the number of contexts; reduction in the problem of data sparsity since the number of potential *n*-grams in greatly reduced; and more reliable probability estimates for events which were not seen in the training data. A clear disadvantage of this type of models is that they lose some of the semantic information that makes the word-based model more powerful. Thus, this latter model is probably not as good as the word-based one, but mixing both may lead to some improvements.

Different research works have reported some improvements when class-based models were mixed with the standard word-based models. In [Moore and Young, 2000] a statistically significant improvement in word recognition accuracy was obtained using a topic-dependent class-based language model interpolated with a word-based n-gram language model. In [Yokoyama et al., 2003] a class-based LM was built based on recognition hypotheses obtained using a general word-based LM, and linearly interpolated

with that general LM. The proposed method was applied to the recognition of spontaneous presentations and was found to be effective in improving the recognition accuracy.

Skipping or Intermediate-distance *n*-gram models

Another simple extension to the word-based *n*-gram models are the *Skipping or Intermediate-distance* models [Rosenfeld, 1994], in which we condition the probability on a different context than the previous $N-1$ words according to a pre-defined distance d. For instance, considering $d=2$ and $N=3$, then $P\left(w_i \left| w_{i-2} w_{i-1}\right.\right) = P\left(w_i \left| w_{i-3} w_{i-2}\right.\right)$. These models attempt to capture directly the dependence of the predicted word on (*N-1*)–grams which are some distance back.

However, intermediate-distance *n*-grams alone do not perform well. Although they capture word sequence correlations even when the sequences are separated by distance d, they fail to appropriately merge training instances that are based on different values of d [Rosenfeld, 1994].

Caching *n*-gram models

Caching models make use of the observation that if you use a word, you are likely to use it again. In [Kuhn et al., 1992] it was shown that words that have occurred recently have a higher probability of occurring in the immediate future than would be predicted by a standard word-based language model. These models tend to be easy to implement and to lead to relatively large perplexity improvements, but relatively small word error rate improvements [Goodman, 2001].

Sentence Mixture *n*-gram models

Sentence Mixture models make use of the observation that there are many different sentence types. Thus, making models for each type of sentence would be better than using one global model. Traditionally, only 4 to 8 types of sentences are used, but in [Goodman, 2001] it was shown that improvements can be obtained by going to 64 mixtures and more. An insightful review of mixture models can be found in [Iyer et al., 1999].

Trigger Pairs *n*-gram models

An interesting enhancement, facilitated by the maximum entropy estimation methodology, has been the use of *triggers* [Rosenfeld, 1994] as the basic information elements for extracting information from the long-distance document history. In this approach, words in the context outside the range of the *n*-gram model are identified as "triggers" and retained together with the "target" word in the predicted position. These pairs (trigger, target) are then treated as complementary sources of information and combined with the standard *n*-gram probability estimates using the maximum entropy methodology. This approach has proven successful, however computationally very demanding.

Factored language models

More recently, a new enhancement to the conventional *n*-gram models was introduced – the so called *factored language models* (FLM) and the *generalized parallel backoff* (GPB) technique which generalizes backoff to arbitrary and even multiple parallel conditional probability paths [Kirchhoff, 2002] [Bilmes and Kirchhoff, 2003].

In a factored language model, each word is viewed as a vector of k factors: $w_i = \left\{ f_i^1, \ldots, f_i^k \right\}$, where those factors can be anything, including linguistic knowledge such as morphological information (POS classes, stems, roots, etc.). Hence, an FLM provides the probabilistic model $P\left(f \middle| f_1, \ldots, f_N \right)$ where the prediction of factor f is based on N parent factors $\left\{ f_1, \ldots, f_N \right\}$. For instance, if w_i is a word and c_i its POS tag, then the probabilistic model $P\left(w_i \middle| w_{i-2}, w_{i-1}, c_{i-1} \right)$ predicts the current word w_i based on the conventional 3-gram model as well as the POS tag of the previous word.

Two features make an FLM distinct from a conventional *n*-gram model. First, the factors $\left\{ f_1, \ldots, f_N \right\}$ can include heterogeneous information other than single words (e.g., word classes, morphological features, etc.). Second, there is no temporal backoff order as in conventional models. In fact, many of the parent factors $\left\{ f_1, \ldots, f_N \right\}$ might be the same age, and the GPB does not necessarily drop the oldest factors first.

In [Vergyri et al., 2004] the authors explore the use of factored language models in a first-pass recognition system for conversational Arabic, obtaining perplexity and word error rate reductions on a large vocabulary recognition task.

Other Models

There are other types of language models which are not based on the conventional *n*-grams. They hardly are being used in large vocabulary speech recognition tasks. A number of alternative models have been developed over the past decade, e.g. application of *decision trees* for clustering of the word histories [Jelinek, 2000], and connexionist models based on neural networks [Bengio et al., 2003].

2.1.4 Discounting and Smoothing Techniques

One of the problems inherent to statistical prediction of natural language is the problem of *data sparseness*. For *n*-gram language models, most of the possible *n*-gram sequences are never encountered in the training data, regardless of the corpora size. Thus, in order for such models to be reliable, one ensure that the probabilities they assign to word sequences are nonzero. Otherwise the "unseen" word sequences would not be recognized at all. Hence, when the *n*-gram estimates are poor, a technique called *smoothing* is applied to adjust those estimates, hoping to produce more accurate models.

Generally, these techniques can be divided in two steps: in a first step some probability mass is removed (*discounting*) from the observed events, being assigned (*smoothing*) in a second step to events which were "unseen" during the training phase. This section will discuss the techniques which are mostly used to smooth the data to correct the bias of the maximum likelihood estimate, and to ensure that no word strings are assigned zero probabilities.

2.1.4.1 Discounting Techniques

The basic idea behind the following discounting approaches is to remove some probability mass from the observed events and assign it to events which were "unseen" during the training phase. In the proposed techniques the count $C(\mathrm{E})$ of an n-gram event $\mathrm{E} = (h, w_i)$ is *discounted* by multiplying it by a discount coefficient $d_{C(\mathrm{E})}$, where $0 \le d_{C(\mathrm{E})} \le 1, \quad \forall\, C(E) \ge 1$:

$$C^*(E) = d_{C(E)} C(E) \tag{2.4}$$

with the remaining probability mass being distributed among "unseen" events.

One of these discounting schemes is based on *Good-Turing discounting*, and was first applied to language modeling by Katz [Katz, 1987]. *Absolute discounting* [Ney et al., 1994] is another technique, which involves subtracting a constant from each of the counts, a simple technique with modest performance that formed the basis for the *Kneser-Ney discounting* [Kneser and Ney, 1995], which showed to perform very well.

In [Chen and Goodman, 1998] and [Chen and Goodman, 1999] the authors carried out an extensive empirical comparison of the most widely-used discounting techniques for $d_{C(E)}$ estimation. On those experiments Kneser-Ney discouting and some variants proposed by the authors (*modified Kneser-Ney discouting*) were found to consistently outperform all other approaches.

2.1.4.2 Smoothing Techniques

Smoothing techniques may be mainly divided into the following categories: *backing-off* and *interpolation*. In the first case, given a certain context, the best *n*-gram model is selected, whereas in the second case all the *n*-gram models of eventual different specificities are mixed together to form a better model for language modeling prediction.

Backing-off

In the principle behind backing-off [Katz, 1987] the most detailed model that is able to provide sufficiently reliable information about the current context is used and models are defined recursively in terms of lower order models. One could, for example, backing-off from 2-gram to *1*-gram models. Normally, the condition for backing-off is that the *2*-gram event does not occur in the model. Thus,

$$P(w_i | w_{i-1}) = \begin{cases} \dfrac{C^*(w_{i-1}, w_i)}{C(w_{i-1})} & \text{if } C(w_{i-1}, w_i) \geq 1 \\ \alpha(w_{i-1}) P(w_i) & \text{otherwise} \end{cases} \tag{2.5}$$

where $C^*(w_{i-1}, w_i)$ is the discounted count, $\alpha(w_{i-1})$ is the back-off weight, being chosen so that $\sum_{w \varepsilon V} P(w|w_{i-1}) = 1$, and V the vocabulary set.

Deleted Interpolation Method

A common alternative to back-off models previously described is the *deleted interpolation* technique [Jelinek and Mercer, 1980][Jelinek, 1990], in which higher-order models are *mixed* with lower-order models. Considering the same 2-gram example, that probability is a linear combination of the unigram and bigram probabilities estimates as follows:

$$P(w_i|w_{i-1}) = \lambda_1 \frac{C(w_i)}{T} + \lambda_2 \frac{C(w_{i-1}, w_i)}{C(w_{i-1})} \tag{2.6}$$

where $\sum_j \lambda_j = 1$, T is the number of words in the language model training dataset, and the λ_j are chosen to maximize the likelihood of some held-out dataset.

2.1.5 Combining Language Models

This sub-section describes various methods of combining different information sources (in this case *n*-gram probability estimates), discussing their advantages and drawbacks.

2.1.5.1 Mixture Models

One of the most widely used techniques for combining language models is the so called *linear interpolation* or *mixture models* [Kneser et al., 1993]. Its simplicity comes from the fact that it is easy to use and any kind of model can be included in the combination process. The most basic way to linearly combine a set of m language models $M_1, M_2, \ldots, M_m$ is to take for each word w_i and context h,

$$P_{LI}(w_i|h) = \sum_{j=1}^{m} \lambda_j P_{M_j}(w_i|h) \tag{2.7}$$

where $\sum_{j=1}^{m} \lambda_j = 1$ and $0 \le \lambda_j \le 1$ for all $j = 1,\ldots,m$.

To obtain the interpolation coefficients λ_j which maximize the likelihood of some *held-out data* the expectation maximization (EM) algorithm [Dempster et al., 1977] can be used. If the held-out data is large enough and representative, these coefficients will be close to optimal for the test data.

This general technique has been frequently applied to combine statistical models of different types. Some examples include combination of back-off and maximum entropy models [Martin et al., 1999] and interpolation of cache, high-order *n*-grams, skipping and sentence-based models [Goodman, 2000].

2.1.5.2 Log-Linear Interpolation

In [Klakow, 1998] a method for combining information sources called *log-linear interpolation* has been introduced. This method and can be viewed as linear interpolation in the log domain, where in contrast with regular linear interpolation described in the preceding sub-section, no explicit constraints appear on the interpolation coefficients.

This method exploits the constrained conditional relative entropy approach. Given the m language models M_j to be combined and their corresponding probability distributions $P_{M_j}\left(w_i|h\right)$, $j = 1,\ldots,m$, the conditional relative entropy of the unknown target model $P_{LLI}\left(w_i|h\right)$ with respect to each of the given language models is defined by the following Kullback-Leibler distance measure:

$$D\left(P_{LLI}\left(w_i|h\right)\middle\|P_{M_j}\left(w_i|h\right)\right) = \sum_{h} P_{LLI}\left(h\right)\sum_{i=1}^{|V|} P_{LLI}\left(w_i|h\right)\log\left(\frac{P_{LLI}\left(w_i|h\right)}{P_{M_j}\left(w_i|h\right)}\right) = d_j \qquad (2.8)$$

where $D(.)$ is the relative entropy between conditional probability distributions $P_{LLI}\left(w_i|h\right)$ and $P_{M_j}\left(w_i|h\right)$, d_j are the constraints on the system, and V the vocabulary set. The target probability distribution should be minimized in terms of its conditional

relative entropy with respect to some additional model. The combined language model probability estimates introduced by Klakow is then defined as

$$P_{LLI}\left(w_i|h\right)=\frac{1}{Z\left(h\right)}\prod_{j=1}^{m}P_{M_j}\left(w_i|h\right)^{\lambda_j} \tag{2.9}$$

where $Z\left(h\right)$ is a normalization factor chosen to ensure that $\sum_{i=1}^{|V|}P_{LLI}\left(w_i|h\right)=1$, being V the vocabulary set. The computation of $Z\left(h\right)$ is very expensive and can usually be dropped without significant loss in performance [Martin et al., 2000]. To estimate the optimal values for the interpolation coefficients λ_j, the generalized iterative scaling algorithm or the simplex method can be employed [Malouf, 2002]. In [Gutkin, 2006] a theoretical framework for smoothing the *n*-gram probability estimates obtained by log-linear interpolation was formulated and has shown to outperform the conventional linear interpolation and back-off techniques when applied to *n*-gram smoothing tasks.

This approach has been used to combine language models of different type, outperforming linear interpolation as reported in some works where standard *n*-gram models were combined with more specific models like distance-based models [Peters and Klakow, 2000] [Beyerlein et al., 2002] [He and Young, 2004].

2.1.5.3 Maximum Entropy

In the methods described in the previous sub-sections, each information source is used separately to construct a model, and the models are then combined. In the *Maximum Entropy* approach [Rosenfeld, 1994], one does not construct those separate models. Instead, a single model, which attempts to capture all the information provided by the various information sources (*features*), is constructed. Each such feature gives rise to a set of *constraints* to be imposed on the combined model. These constraints are typically expressed in terms of marginal distributions. Maximum entropy modeling produces a probability model that is as uniform as possible while matching empirical feature expectations exactly. It combines multiple overlapping information sources as follows:

$$P(o|h) = \frac{\exp\left(\sum_i \lambda_i f_i(o|h)\right)}{\sum_{o'} \exp\left(\sum_j \lambda_i f_j(o'|h)\right)} \tag{2.10}$$

which describes the probability of a particular outcome o given the history or context h. The denominator includes a sum over all possible outcomes, o', which is essentially a normalization factor for probabilities to sum to 1. The indicator functions f_i (features) are "activated" when certain outcomes are generated for certain context:

$$f_i(o|h) = \begin{cases} 1 & if\ o = o_i\ and\ q_i(h) = 1 \\ 0 & otherwise \end{cases} \tag{2.11}$$

where o_i is the outcome associated with feature f_i and $q_i(h)$ is an indicator function on histories. As an example, a bigram feature f_i representing the word sequence "BOA NOITE" in the Maximum Entropy approach would have $o_i = \text{"NOITE"}$ and $q_i(h)$ would be the question "Does the context h contains the word "BOA" as the previous word of the current word ?".

The next step in the Maximum Entropy approach is to choose, from among the models in that set, the one which has the highest entropy, i.e. the maximum entropy model. This can be achieved using the Generalized Iterative Scaling (GIS) algorithm [Malouf, 2002].

The Maximum Entropy formalism allows to fully integrate complementary statistical properties of limited training data. However, the training algorithm for this approach is computationally very demanding, which explains the lack of widespread use of this language modeling technique. Further issues involved in maximum entropy language modeling can be found in [Rosenfeld, 1994].

2.1.6 Language Models Adaptation

Training an *n*-gram language model requires large quantities of text matching the target recognition task both in terms of style and topic. In tasks involving conversational speech

like broadcast news, the ideal training material, i.e. transcripts of spoken speech, is costly to obtain, which limits the amount of training data currently available. Methods have been developed for the purpose of language model adaptation, i.e. the adaptation of an existing model to new topics, domains, or tasks for which little or no training material may be available. A general LM adaptation framework is depicted in figure 2.1. Two datasets are considered: an adaptation dataset (*in-domain data*), relevant to the target recognition task, and a background dataset (*out-of-domain data*), associated with a related but perhaps somewhat different task and/or out-dated data. The general idea is to dynamically modify the background model probability estimates on the basis of what can be extracted from the adaptation data (task/domain specific knowledge).

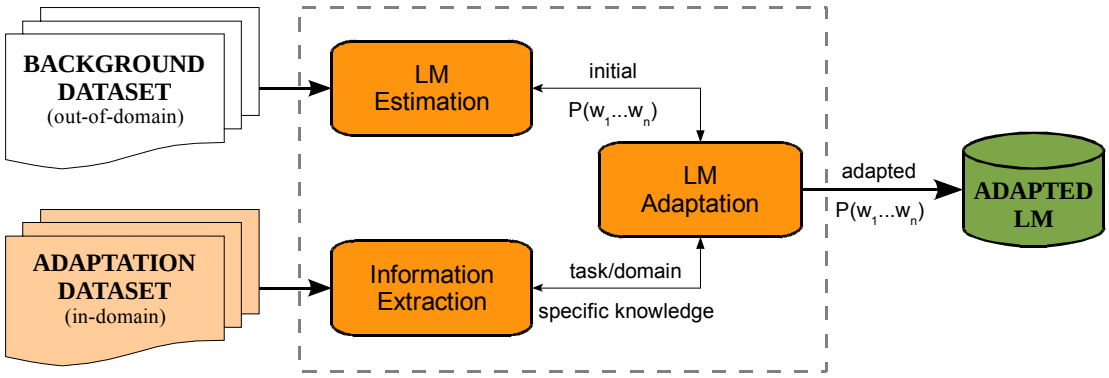

Figure 2.1: A general framework for LM adaptation.
(adapted from [Bellegarda, 2004])

Language model adaptation can take several forms. Adaptation can be performed *offline*, in which the language models are adapted in advance to their use, in opposed to an *online* approach where the language models change at run-time. Moreover, adaptation can be classified as *supervised adaptation* (a priori) when using data chosen for a particular and known domain. This is different from the *unsupervised adaptation* approach, where the language model is adapted in some form based on the sentences that have been recognized already.

The various adaptation techniques that have been proposed along the time can be generically classified into three major categories [Bellegarda, 2004]:

Model Interpolation

A common approach for LM adaptation is the *mixture model* (see section 2.1.5), i.e., the interpolation of two or more component models considered at the *n*-gram level. The simplest and more used way to do so is by means of linear interpolation, with a large number of variants depending on the type of models interpolated. It is common to use a background model based on words, interpolated with class-based models, or distance models, or cache models, etc.

For instance, in [Zhu and Rosenfeld, 2001], the *n*-gram counts estimated from the Web are interpolated with traditional corpus-based estimates, resulting in a significant reduction in ASR word error rate. In [Beyerlein et al., 2002], the Philips/RWTH system for transcription of broadcast news was improved by means of logarithmic interpolation. A standard n-gram model was combined with various "distance" language models, reporting better results. In [Federico and Bertoldi, 2004], a rolling language model with an updated vocabulary was implemented for an Italian broadcast news transcription system using a single-step adaptation framework. The baseline vocabulary of 62K words is extended by adding 60K new words selected from the contemporary written news and the baseline language model is interpolated with a new language model estimated from those written news on a daily basis. This approach allowed an average relative reduction of 58% in terms of OOV word rate and 3.4% in terms of WER. In [Lavecchia et al., 2006] the authors introduce an original cache model called Features-Cache (FC) to estimate the gender and the number of the word to predict. Henceforth, in their model a word depends not only on its left context, but also on the gender and number present in left contexts. This model is linearly interpolated with a classical *n*-gram model, with the new model outperforming the baseline one, in terms of word error, by 3%.

Constraint Specification

For this approach, the in-domain dataset is used to extract some pre-defined features, with the background language model being constrained to satisfy them. Historically, constrained based techniques have been associated with exponential models trained using the maximum entropy (ME) method, being commonly referred as *minimum discrimination information* (MDI) estimation. This adaptation technique has been investigated in [Federico, 1999], and can be expressed as follows: given a background model $P_{Back}\left(w|h\right)$

and an adaptive corpus $Adapt$, we aim to find a model $P(w|h)$ satisfying a set of linear constraints extracted from $Adapt$ and minimizing the Kullback-Leibler distance between $P(w|h)$ and $P_{Back}(w|h)$. The MDI model can be trained by using the GIS (Generalized Iterative Scaling) algorithm.

A special case of this approach is the MDI adaptation with unigram constraints. The basic approach is to choose the adaptive model as close as possible to the background model estimates while constraining them to respect the locally estimated unigram probabilities of the adaptation dataset. In fact, given the typically small amount of adaptation data available, it is often the case that only unigram features can be reliable extracted from the adaptation dataset. In [Kneser et al., 1997] the authors describe and evaluate a new method to quickly modify a given static *n*-gram model such that the local unigram properties are correctly modeled without destroying the full context dependency of the original model.

In [Chen et al., 2004] the performance of 5 widely used LM adaptation methods using the same broadcast news adaptation data are compared, with the experimental results showing that MDI method yields the best performance. Experiments carried out for BN transcription in English and Mandarin showed a relative word error rate reduction of 4.7% in English and 5.6% relative character error rate reduction in Mandarin with MDI adaptation.

In [Boulianne et al., 2006] an adaptation approach was implemented for closed-captioning of live TV broadcast in French, which uses texts from a number of websites, newswire feeds and broadcaster's internal archives to adapt its language model component. To generate the baseline model, texts are automatically classified into 8 pre-defined topics using a Naïve Bayes classifier. Using those topic-specific vocabularies of 20K words, a 3-gram language model is estimated for each topic partition. Each day, newly collected texts are compared against the current topic-specific vocabularies and potential new words associated to each one of them. This association has a limited lifetime, so words become inactive in a topic after 60 days, with the exception that words from the baseline vocabularies never become inactive. This vocabulary adaptation procedure showed to be effective, allowing the dynamic vocabulary to produce only around half the static vocabulary OOV word rate. The topic-specific language models are adapted in an

unsupervised way with texts classified into topics automatically. The unigrams of the baseline language model are interpolated with a unigram language model estimated from the adaptation data, and then higher-order n-gram probabilities in the baseline language model are re-estimated according to the MDI method, which attempts to adapt the baseline LM by minimizing the Kullback-Leibler divergence between the adapted LM and the baseline LM subject to a constraint that the marginal unigram distribution of the adapted LM is equal to the adaptation unigram distribution. This procedure showed to provide good recognition results for words added with very small quantity of adaptation data.

In [Tam and Schultz, 2006] the same MDI method was used into a similar adaptation approach but using the Latent Dirichlet Allocation (LDA) model [Blei and Jordan, 2003], a Bayesian latent semantic analysis approach, to estimate the marginal unigram distribution based on the ASR hypotheses. Results computed on a Mandarin Broadcast News test set showed a relative character error rate reduction of 2% when compared to the un-adapted baseline language model.

In [Wang and Stolcke, 2007] the integration of various language model adaptation approaches are investigated for a cross-genre adaptation task to improve the performance of Mandarin ASR system performance on a recently introduced new genre, broadcast conversation (BC). Various language model adaptation strategies were investigated, including unsupervised language model adaptation from ASR hypotheses and ways to integrate supervised *Maximum a Posteriori* (MAP) and marginal adaptation within an unsupervised adaptation framework. By combining these adaptation approaches on a multi-phase ASR system, a relative gain of 1.3% on the final recognition error rate in the BC genre was achieved.

Meta-Information Extraction

One common approach is exploiting the underlying topic of the discourse. Thus, the adaptation dataset is used to extract information about the related subject matter, being that information used to improve the background model. Considering a set of T topics, manually defined or data-driven obtained, a smaller model for each one those topics is trained on the relevant portion of the background dataset. The simplest approach is based on linear interpolation. Hence, those T *n*-gram models are linearly interpolated in such a way that the resulting mixture bet matches the adaptation dataset. In [Iyer et al., 1999] a

review of topic dependent mixture models is given, showing improvements both in terms of perplexity and WER comparing with the conventional *n*-gram models.

Approaches taking advantage of semantic knowledge try to exploit not just the topic information, but the entire semantics present in the adaptation dataset. As seen before, the word trigger pairs approach exploit correlations between the current word and features of the long-range context [Rosenfeld, 1994]. *Latent Semantic Analysis* (LSA) [Deerwester et al., 1990], a paradigm formulated in the field of Information Retrieval, has extended the word trigger concept by defining a more general framework to handle the trigger pair selection. In this paradigm, co-occurrence analysis still takes place across the span of an entire document, but every combination of words from the vocabulary is viewed as a potential trigger combination. This leads to the systematic integration of long-term semantic dependencies into the analysis. For this approach it is assumed that the available training data is tagged at the document level, i.e., there is a way to identify article boundaries.

Hybrid *n*-gram+LSA language models, constructed by embedding PLSA (*Probabilistic Latent Semantic Analysis*) into the standard *n*-gram formulation, showed to result in a substantial reduction in both perplexity and average word error rate [Bellegarda, 2000]. In this work, the author proposes the following integrated language model probability:

$$P\left(w\middle|h,\hat{h}\right)=\frac{P\left(w\middle|h\right)\beta\left(w,\hat{h}\right)}{Z\left(h,\hat{h}\right)} \tag{2.12}$$

where $\hat{h}$ represents the global document history which in most cases can have a much larger span than the n-gram history h, $\beta\left(w,\hat{h}\right)$ is a measure of the correlation between the current word w and the global history $\hat{h}$ defined by the PLSA paradigm, and $Z\left(h,\hat{h}\right)$ being a normalization factor. Experiments conducted on the Wall Street Journal domain showed a relative reduction in average word error rate of over 20%.

In [Mrva and Woodland, 2006] a PLSA-based approach was investigated for unsupervised language model adaptation of a broadcast conversation transcription system.

Results showed a relative improvement in the system performance when the PLSA framework was embedded within the conventional n-gram language model. For a task of Mandarin broadcast conversation transcription, this language model adaptation done with PLSA and LDA brought 1.3% absolute character error rate gain.

Generally, the adaptation dataset may already be available. However, if it is too small, or some form of dynamic data updating is necessary, like in tasks as broadcast news transcription/indexation or spoken dialogues, it is possible to use document retrieval techniques to collect up-to-date data by dynamically searching online databases and/or the Word Wide Web [Schwarm et al., 2004]. In section 2.2 we give a brief overview about Information Retrieval methods and their use for language model adaptation.

2.2 Information Retrieval and LM Adaptation

In our work we propose a multi-pass adaptation approach using Information Retrieval techniques. Hence, for a general overview, we will briefly introduce the standard techniques developed for Information Retrieval systems and its use for LM adaptation.

2.2.1 Brief Introduction to IR

Primarily Information Retrieval is concerned with the identification of information sources that are related to a user's request: automatically retrieving *documents* that are most likely relevant to a user's *query* by selecting those that contain *terms* (for example, words) that identify such *relevance*. Figure 2.2 illustrates the procedure for retrieving documents (d) according to a user's query (q), presenting them in order of decreasing *rank*.

Text normalization is a very common and important procedure to normalize the documents for IR, allowing to reduce the set of terms that could represent the content of documents. *Stopping* and *stemming* are two common ways of doing this. The first operation is a standard processing, which removes common terms and irrelevant terms (for example, most of the functional words are often discarded). Stemming is another common procedure, which transforms each word stem (*word stemming*). However, contradictory

results have been reported concerning the use of stemming in IR related tasks [Lo and Gauvain, 2005].

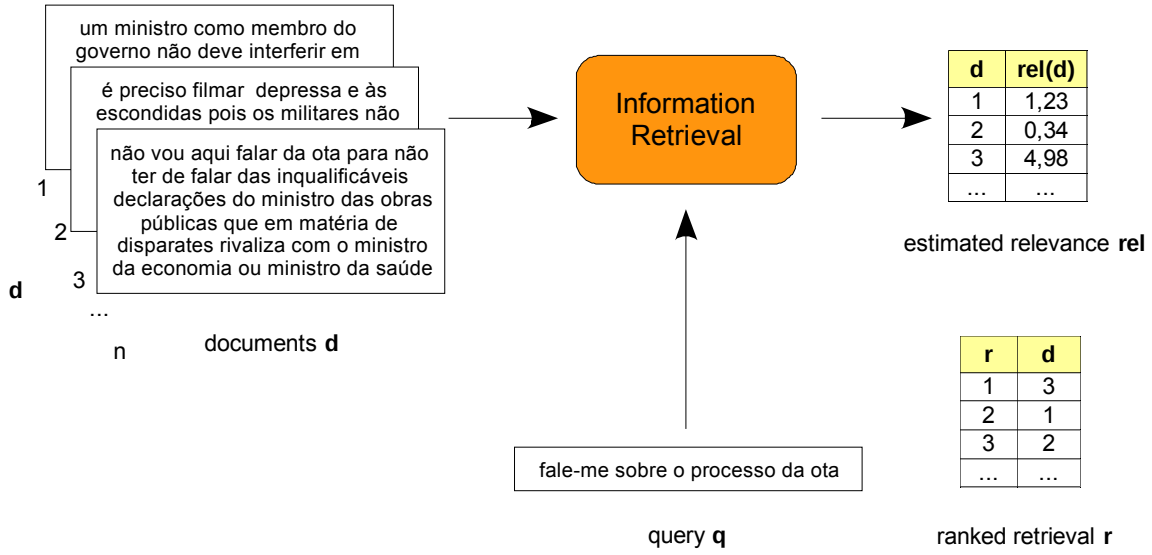

Figure 2.2: A Typical Information Retrieval (IR) System.

Early IR systems were Boolean systems which allowed users to specify their information need using a complex combination of Boolean operators: ANDs, ORs and NOTs. However, Boolean systems have several shortcomings, e.g., there is no inherent notion of document ranking, and it is very hard for a user to form a good search query. Hence, nowadays most IR systems assign a numeric *score* to every document and rank them by this score. Several models have been proposed for this process. The three most used models in IR research are the *vector space model*, the *probabilistic models*, and the *inference network model* [Singhal, 2001], that we briefly describe in the next sub-sections.

2.2.2 IR Model Types

In the vector space model each document and query are represented as a vector with each dimension corresponding to a separate term. Thus, if a term occurs in the document, its value in the vector is non-zero. Different approaches for computing those vector values (known as term weights) have been proposed [Jurafsky and Martin, 2000]. In the classical model the term weights in the document vectors are products of local and global

parameters. The model is known as *term frequency-inverse document frequency model* (tf-idf) [Salton et al., 1975]. To assign a numeric score to a document for a query, the model measures the similarity between the query vector and the document vector. Typically, the angle between two vectors is used as a measure of divergence between the vectors, and *cosine* of the angle is used as the similarity measure. However, most of the current IR systems use other state-of-the-art scoring methods like the *Okapi weighting* or the *pivoted normalization weighting* [Singhal, 2001], which showed to be especially powerful.

Another classic retrieval method is the probabilistic models, where the probability that a specific document will be judged relevant to a specific query, is based on the assumption that the terms are distributed differently in relevant and non relevant documents. This is often called the *probabilistic ranking principle* [Robertson, 1977]. Since true probabilities are not available to an IR system, probabilistic IR models are used to estimate the probability of relevance of documents for a query. This estimation is the key part of the model, and this is where most probabilistic models differ from one another. More recently, statistical language model was suggested for Information Retrieval [Ponte, 1998]. When used in IR, a language model is associated with a document in a collection. Thus, given a query q as input, retrieved documents are ranked based on the probability $P\left(q\middle|M_d\right)$, the probability that the document's language model would generate the terms of the query. As in speech recognition, various smoothing techniques have been proposed to deal with the sparseness data problem [Song and Croft, 1999].

The inference network model approach to Information Retrieval, first introduced in [Turtle and Croft, 1991], provides a theoretical framework to combine many sources of evidence of document relevance. In this approach, a query is viewed as a series of concepts, which can be terms, phrases, or more complex entities. Hence, one document is relevant precisely when it contains the concepts present in the query.

2.2.3 Query Expansion

Another technique that has been shown to be effective in improving document ranking is query modification via *relevance feedback*. The idea behind relevance feedback is to take the results that are initially returned from a given query and to use information about whether or not those results are relevant to perform a new expanded query. In general, the

user is asked to judge the relevance of the top few documents retrieved by the system. Based on these judgments, the system modifies the query and issues the new query for finding more relevant documents from the collection. This is the called *explicit relevance feedback*.

However, for the majority of practical applications those judgments do not exist, and new techniques to do meaningful query expansion in absence of any user feedback were developed. One of these techniques, the so called *pseudo relevance feedback*, has been shown to be a very effective technique, especially for short user queries [Buckley et al., 1995] [Lavrenko et al., 2001]. With pseudo relevance feedback the IR system retrieves and ranks documents according to the initial user's query, extracts relevant terms from the top T documents, which are then added to the initial query. With this new expanded query, the IR system assigns similarity scores to each one of the documents ranking them.

2.2.4 LM Adaptation using IR

As seen in the previous section, language model adaptation is recognized to be an important research area in speech recognition. However, despite the numerous efforts to improve upon the commonly used *n*-gram language models, adaptation approaches have been only moderately successful for complex tasks such as broadcast news transcription, with different reasons accounting for this observation [Chen et al., 2001]: the wide variety of BN data, with a given audio segment almost always related to more than one topic; the content of BN data is open, i.e., new stories appearing on a daily basis, which makes it impossible to obtain adaptation data in advance; and large linguistic differences in style between the training and input data to be recognized.

To solve the linguistic problem in adaptive language modeling for BN transcription, various approaches using Information Retrieval (IR) technology have been proposed, with various Information Retrieval techniques being applied to both vocabulary and language model adaptation problems. These multi-phase speech recognition approaches use the ASR hypotheses as queries to an IR system in order to select additional on-topic adaptation data.

In [Kemp and Waibel, 1998] and [Yu et al., 2000] IR techniques were applied to the OOV problem by dynamically adapting the active vocabulary to the current news topic. A similar approach was implemented in [Chen et al., 2004]. In this case the vocabulary

remains static, with only the language model being updated. To address the changing property of broadcast news data, static and dynamic language models for language model adaptation were investigated. In static modeling the language model is updated once for the all BN show. Dynamic modeling updates the language model at each automatically detected story of the BN show, which means estimating multiple story-based language models for each BN show. Experiments were carried out for broadcast news transcription in English and Mandarin Chinese. A relative WER reduction of 4.7% was obtained in English and a 5.6% relative character error rate reduction in Mandarin with story-based language model update and using the MDI adaptation technique.

In [Bigi et al., 2004] an approach of this type using the Kullback-Leibler symmetric distance to retrieve documents was implemented to select a dynamic vocabulary instead of a static one, obtaining an OOV word rate reduction of about 28% with the same vocabulary size as the baseline vocabulary. Moreover, a new topic language model was trained on the retrieved data, and interpolated with the baseline language model, allowing for a relative improvement of 1.8% in terms of WER.

2.3 Evaluating Language Models Quality

The ultimate measure of the quality of a language model is its impact on the performance of the application it was designed for. Thus, in speech recognition, we would evaluate a language model based on its effect on the recognition word error rate (WER). However, all attempts to derive an algorithm that would directly estimate the model parameters so as to minimize WER have failed. As an alternative, a statistical language model is evaluated by how well it predicts a string of words W (commonly referred to as test dataset) generated by the information source to be modeled. Next we present a brief review of the common measure of ASR accuracy – the word error rate (WER), and the perplexity (PP).

2.3.1 Word Error Rate (WER)

The performance of an ASR system is commonly evaluated by the word error rate (WER). This measure is based on the comparison of a *reference transcription* of the test dataset,

with the corresponding output of the ASR system which is referred to as the *hypothesis transcription*. Thus, a scoring algorithm searches for the *minimum edit distance* (in words) between the reference and the hypothesis transcriptions, returning the number of word mismatches: substitutions (Sub), deletions (Del) and insertions (Ins). Hence, the WER is defined as:

$$WER = \frac{Sub + Del + Ins}{\text{total number of words in reference}} \tag{2.13}$$

A related measure is the rate of words correct, which measures the proportion of words that were correctly recognized, and therefore ignores insertion errors.

However, this WER measure has a number of drawbacks. For example, reliably measuring the word error rate entails the processing of large amounts of test data, which is very time consuming and forces to manually transcribe that data which is very costly. Moreover, the count of errors is done independently of the words wrongly recognized, i.e., this measure considers that all errors are equally harmful independently of the task domain. However, for some applications words play different roles, being more important to correctly recognize content words like names than generic words as the functional ones.

2.3.2 Perplexity

The most common metric for evaluating a language model is the *perplexity* (PP) concept, which measures the LM capability to predict an unseen sequence of words, i.e., a sequence of words not used for model training.

Assume we compare two models M_1 and M_2, which assign probabilities $P_{M_1}(W)$ and $P_{M_2}(W)$, respectively, to a word string $W = w_1,\ldots,w_n$ that has not been seen at the training step of either models and that was supposedly generated by the same information source that we are trying to model. As it would be "natural", we can consider M_1 being a better model than M_2 if $P_{M_1}(W) > P_{M2}(W)$. Hence, a quality measure under the name of perplexity was introduced [Bahl et al., 1977][Bahl et al., 1983], which relates the quality of

a given model M to the entropy of its underlying information source, formulating perplexity as

$$PP_M = \exp\left(-1/n \sum_{i=1}^{n} \ln\left[P_M\left(w_i \middle| w_1, ..., w_{i-1} \right) \right] \right) = P_M\left(W \right)^{-1/n} \tag{2.14}$$

To get an intuitive understanding for PP (2.14) we can state that it measures the average surprise of model M when it predicts the next word w_i in the current context $w_1, \ldots, w_{i-1}$. The goal of statistical language modeling therefore can be viewed as minimizing the perplexity so as to bring it down as close as possible to the true entropy of the language.

When comparing perplexity numbers for different texts and/or different models one fact should be taken into consideration: the perplexity measure is a function of both the model and the text. Thus, a meaningful comparison can only be made between perplexities of several models, all with respect to the same text and the same vocabulary. Vocabularies must be the same, or else a smaller vocabulary will bias the model towards a lower perplexity value. Even if vocabularies are identical, different texts will not produce a meaningful comparison, since texts could have different out-of-vocabulary word rates.

2.4 Summary

Statistical Language Models have been successfully applied to many state-of-the-art ASR systems, with the n-gram models being the dominant technology in language modeling. Usually large training corpora are used to estimate the language model parameters, with different smoothing techniques, such as discounting, backing-off and interpolation, being applied to better estimate probabilities when there is insufficient data to estimate probabilities accurately. However, the collection of those suitable training corpora is an expensive, time-consuming and sometimes unfeasible task. Moreover, it is also recognized that generic language models trained on large amounts of textual data can be advantageously adapted to more specific domains in order to improve their accuracy related to a particular domain [Schwarm et al., 2004].

Therefore, the idea of language model adaptation is to use a small amount of domain specific data (in-domain data) to adjust the LM and reduce the impact of linguistic differences between the training and testing data over time. For that propose, several techniques have been developed by the research community. In [Bellegarda, 2004] these techniques are classified into three major categories: interpolation based approaches, such as the cache model and maximum a posteriori adaptation (MAP); constraints based models, such as the maximum entropy (ME) and minimum discrimination information (MDI); and meta-information based frameworks, such as the trigger model, latent semantic analysis (LSA) and structured language models.

In terms of vocabulary selection, one approach is to choose the words in such a way to reduce the OOV word rate by as much as possible – a strategy usually called by vocabulary optimization. This optimization could either involve increasing the vocabulary size of the ASR component, or it could also be selecting those words most representative for the target domain/task. The most common approaches are typically based on word frequency, including words from each training corpus that exceed some empirically defined threshold, which mainly depends on the relevance of the corpus to the target task [Gauvain et al., 2002]. Thus, to eliminate the need for human intervention during the vocabulary selection process various approaches have been suggested.

During the last decade other strategies have been proposed to address the OOV word rate problem. To achieve a usable OOV word rate, morphemes or other sub-word units (namely stems and endings) have been used instead of words, with those units defined through data driven processes. Most research on morpheme-based systems has been developed for inflected languages. These works report some improvements in terms of OOV word rates and/or WER, especially when combining these models with the standard ones based on words. The problem of acoustic confusability arises when larger sub-word vocabularies are used. In fact, sub-word language model units alleviate the problem of rapid vocabulary growth, especially for this kind of highly inflectional languages. But as the base units of the language model become fewer and smaller, the language model becomes less constrained and the acoustic confusability increases.

For broadcast news and conversational speech applications there have been various works using data from the Web as an additional source of training data for unsupervised language modeling adaptation over time, also referred to as dynamic vocabulary and LM

adaptation. Some of those works have been using different Information Retrieval (IR) techniques for dynamic LM adaptation to the topics present in a BN show using relevant documents obtained from a large general corpus or from the Web. These multi-pass speech recognition approaches use the ASR hypotheses as queries to an IR system in order to select additional on-topic adaptation data.

3

Resources and Baseline System

In this chapter we present the experimental setup under which the work for this thesis was carried out. This chapter describes the corpora used for the task focused on this work - broadcast news transcription for the European Portuguese language. Details of the various speech and text corpora used for training, adaptation and evaluation propose are given. Next, the baseline system used for our empirical studies is briefly described in terms of it ASR module, presenting details of the vocabulary and language model used. The chapter concludes with a description of the various processing tools used on this work for language modeling, Information Retrieval extraction and morpho-syntactic tagging.

3.1 The Corpora

These corpora are constituted mainly by news reports from two sources: television Broadcast News (BN) and Web Text News (TN).

The speech and text BN corpus is composed by news shows transmitted by the Portuguese public television broadcast company (RTP). The BN resources were used for training the ASR module of the baseline system (both acoustic and language model).

The text TN corpus called WEBNEWS-PT is composed by Web journal articles from several European Portuguese daily newspapers. The TN resources were used as training material for the language model component of the baseline system. Some datasets from both BN and TN corpora were used as adaptation and evaluation data for the work

presented in this thesis. The next subsections describe in detail each of these speech and text resources.

3.1.1 Broadcast News Corpus (ALERT-SR)

The ALERT-SR BN corpus [Neto et al., 2003] was the first Broadcast News corpus collected for the European Portuguese language. The collection process started in April 2000 and lasted until the end of 2001. It is entirely constituted by European Portuguese BN shows transmissions from the main public Portuguese channel, RTP, and includes both the speech signal and its orthographic transcription which were manually provided by specialized transcribers following the LDC Hub4 [LDC-Hub4, 2000] (Broadcast Speech) transcription conventions.

Initially this corpus was divided into five different datasets. Two datasets were used for training proposes (pilot and train datasets) and three datasets were used for testing purposes (devel, eval, and jeval datasets). Table 3.1 gives an overview of the ALERT-SR corpus in terms of quantity (number of news shows), duration (speech signal) and purpose of the datasets.

datasets	**#shows**	**audio**	**Propose**
pilot	11	5 h	Training (cross-validation)
train	99	46 h	Training
devel	13	6 h	Development (parameters estimation)
eval	12	4 h	ASR evaluation
jeval	14	13 h	Media monitoring evaluation
11march	7	5 h	Evaluate daily LM adaptation
RTP-07	2	2 h	Evaluate daily LM adaptation
Total	158	81 h	

Table 3.1: ALERT-SR datasets: speech statistics.

As can be seen in Table 3.1 ALERT-SR has approximately 51 hours for training purposes, 6 hours for parameters estimation and 17 hours for evaluation purposes.

In terms of orthographic transcriptions the training datasets consist of a total of roughly 34.3 K sentences and 531.7 K tokens (see table 3.2). The word statistics are shown in terms of "tokens", in which all occurrences of a word are counted, and in terms of "types", in which only unique words are counted.

Each point in the datasets where the topic being discussed changes was labeled, allowing each dataset to be partitioned into topic-homogeneous segments. This was important for the construction of the multi-pass recognition approach described in Chapter 4. By using such partition, it resulted in approximately 1,651 news segments in the training set.

Datasets	**#segments**	**#sentences**	**#types**	**#tokens**
pilot	116	3.2 K	6.9 K	50.0 K
train	1535	31.1 K	25.0 K	481.7 K
devel	221	4.1 K	8.5 K	66.4 K
eval	169	3.1 K	7.0 K	47.4 K
jeval	387	8.0 K	12.9 K	137.7 K
11march	151	3.1 K	7.0 K	53.0 K
RTP-07	52	0.4 K	3.7 K	16.1 K

Table 3.2: ALERT-SR datasets: text statistics.

Later on, and for proposes of our work, two more evaluation datasets were added: the 11march and the RTP-07 datasets. To evaluate the adaptation approaches proposed in this thesis we started by choosing the week starting on March 8th and ending on March 14th as our target dataset. Due to the unexpected and awful events occurring on March 11th of 2004 in Madrid, we would expect to cover a typical situation of rich content and topic changing over time. Thus, we collected the "11march" dataset consisting of seven BN shows from the 8 o'clock pm (prime time) news from the main public Portuguese channel, RTP. These BN shows had a total duration of about 5 hours of speech signal (roughly 53 K tokens). Finally, and to evaluate the practical implementation of our adaptation algorithms in the current online BN transcription system, we selected another two BN shows, the "RTP-07" dataset. The selected BN shows had a total duration of two hours of speech,

consisting of about 16.1 K tokens. These BN shows were collected on May 24th and 31st of 2007.

	8th	9th	10th	11th	12th	13th	14th
#tokens	8.7 K	1.9 K	8.4 K	8.3 K	8.8 K	7.0 K	9.4 K
#types	2.4 K	0.7 K	2.4 K	2.0 K	2.1 K	1.8 K	2.1 K

Table 3.3: ALERT-SR.11march dataset: text statistics.

	May 24th	May 31st
#tokens	8.1 K	7.9 K
#types	2.3 K	2.3 K

Table 3.4: ALERT-SR.RTP-07 dataset: text statistics.

In tables 3.3 and 3.4 we present more detailed statistics related to these two last datasets ("11march" and "RTP-07" respectively). As one can observe, each BN show had an average size of about 8.3 K tokens and only 2.2 K different words (types). Related to day 9th of "11march" dataset due to a technical problem only 12 minutes of speech was collected. For that reason, this day has different statistics considering the average values.

3.1.2 Web Text News Corpus (WEBNEWS-PT)

Despite all the research done in the last two decades, n-gram language models still dominate as the technology of choice for state-of-the-art speech recognizers. Typically, n-gram language models for large vocabulary speech recognizers are trained on hundred of millions or billions of word strings for better estimation of their parameters. For training the language model component of an ASR system for tasks like BN transcription, the best approach would be to use transcriptions from BN news shows. However, due to the small quantity of transcribed BN news shows available, it is common to use other sources like newspapers written texts.

Since 1995 we had been collecting on a daily basis the online editions of all major Portuguese newspapers. This Web text news corpus called WEBNEWS-PT and collected until the end of 2005 includes texts selected from newspapers of different styles (daily newspapers covering all topics, weekly newspapers also with a broad coverage of topics, economics newspapers and sports news). The newspapers were selected for their content and reliability to better reflect the lexical and linguistic content of current news events. This corpus has over 42 million sentences and 740 million words (tokens). Table 3.5 gives a brief summary of this text corpus.

Newspaper	**#sentences**	**#tokens**	**Style**
A Bola	1.9 M	32.2 M	daily, sports
Diário de Notícias	5.2 M	88.9 M	daily, generic
Diário Económico	5.9 M	66.6 M	daily, economics
Expresso	2.0 M	39.9 M	weekly, generic
Jornal de Notícias	5.4 M	94.5 M	daily, generic
O Jogo	6.9 M	91.0 M	daily, sports
O Independente	0.2 M	2.4 M	weekly, generic
O Público	14.9 M	325.8 M	Daily, generic
Total	42.4 M	741.3 M	

Table 3.5: WEBNEWS-PT corpus: text statistics.

After collecting the daily web edition of a given newspaper, scripts were used to convert the text from html format into simple text format. At this stage the processing scripts checked to see if there were repeated news articles by comparing each article with last day articles from the same newspaper. Due to the heterogeneous variety of sources, a normalization process is applied to the collected texts in order to clean errors due to common misspellings, expanding abbreviations and acronyms, processing ambiguous punctuation marks, and converting numbers to word sequences. The resulting normalized texts are then sgml tagged and compressed to save disk space with the text now being ready for use.

After all these processing stages, information like the topic of the article is preserved. For our work this information will be used by an Information Retrieval engine to

dynamically index, store and retrieve those articles. The WEBNEWS-PT corpus is split into 1,554,156 different files, each containing text from a different news article.

Finally, in tables 3.6 and 3.7 we present more detailed statistics related to the two datasets ("11march" and "RTP-07" respectively) of the WEBNEWS-PT corpus. The WEBNEWS-PT.11march contains the written news collected during the same week as the ALERT-SR.11march dataset, with WEBNEWS-PT.RTP-07 collected on the same days as the ALERT-SR.RTP-07 dataset. For evaluation proposes, these two datasets were only used during the proposed adaptation processes. As one can observe, the WEBNEWS-PT.11march dataset had an average size of about 280 K tokens and only 25 K different words (types). The WEBNEWS-PT.RTP-07 dataset had an average size of about 80 K tokens and 11 K types.

	8th	9th	10th	11th	12th	13th	14th
#tokens	280 K	270 K	286 K	232 K	250 K	319 K	310 K
#types	24 K	23 K	25 K	24 K	25 K	27 K	26 K

Table 3.6: WEBNEWS-PT.11march dataset: text statistics.

	May 24th	May 31st
#tokens	120 K	98 K
#types	12 K	10 K

Table 3.7: WEBNEWS-PT.RTP-07 dataset: text statistics.

3.2 The Baseline System (AUDIMUS.media)

All the speech recognition work reported in this thesis is done within the AUDIMUS.media ASR system [Meinedo et al., 2003][Meinedo, 2008]. This system is part of a closed-captioning system of live TV broadcasts in European Portuguese that is daily producing online captions for the main news show of one Portuguese Broadcaster - RTP.

Figure 3.1 presents an overview of the AUDIMUS.media ASR system in terms of its main components.

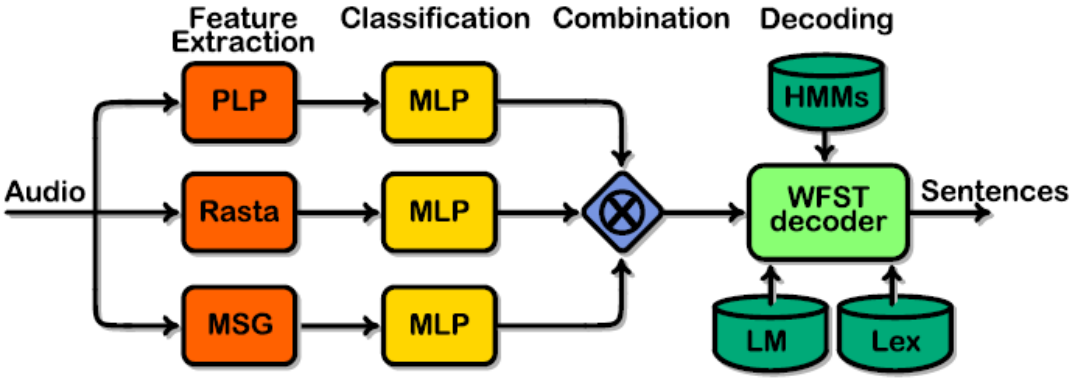

Figure 3.1: AUDIMUS.media ASR system.
(extracted from [Meinedo, 2008])

The next subsections give a brief overview of AUDIMUS.media baseline system in terms of it main components: acoustic modeling, lexical modeling, language modeling, decoding process and its confidence features and associated confidence scoring.

3.2.1 Acoustic Modeling

The AUDIMUS.media acoustic model features a hybrid Hidden Markov Models (HMMs) / Multi-Layer Perceptrons (MLPs) system [Bourlard and Morgan, 1994], using three MLPs, each of them associated with a different feature extraction process, where the MLPs are used to estimate the context independent posterior phone probabilities given the acoustic data at each frame. The phone probabilities generated at the output of the MLPs classifiers are combined using an appropriate algorithm [Meinedo and Neto, 2000]. All MLPs use the same phone set constituted by 38 phones for the Portuguese language plus the silence and breath noises. The training and development of this system was based on the European Portuguese ALERT-SR BN corpus. The acoustic models were trained over 46h of transcribed speech (ALERT-SR.train dataset).

3.2.2 Lexical Modeling

Knowledge about the allowed set of words and their respective pronunciations is represented through the lexical model, or simply the lexicon. Commonly, BN ASR systems developed for the English language are based on vocabularies of 64 K words in size. The vocabulary selection for this baseline system followed the same approach although the European Portuguese language is more inflectional than the English language. From the 604.2 M words of newspaper texts that composed the WEBNEWS-PT corpus at that time, 427 K different words were extracted. From those words around 100 K had an occurrence frequency higher than 50 in the newspapers texts. These 100 K words were selected and classified according to their syntactic classes. From that set of words a subset was selected according to their weighted class frequencies of occurrence. Different weights were used for each syntactic class of words. This subset was augmented with all new words found in the training datasets of ALERT-SR BN corpus giving a final vocabulary of 57,564 words (called in our work as the 57K baseline vocabulary).

This word list was then phonetically transcribed by a rule grapheme to phone system generating an initial set of pronunciations. This automatically generated lexicon was then hand revised by a specialized linguist generating a multi-pronunciation lexicon with 65,585 different pronunciations.

3.2.3 Language Modeling

The AUDIMUS.media baseline language model generated in our research work presented in [Martins et al., 2005] combines a 4-gram backoff LM generated from the WEBNEWS-PT corpus (a total of 604.2 M words consisting of all the newspaper texts collected until the end of 2003), and a 3-gram backoff LM estimated on the 531.7 K word corpus of broadcast news transcripts (ALERT-SR.pilot and ALERT-SR.train datasets). The 4-gram backoff LM was generated using the absolute discounting method and applying cutoff values of 2, 3 and 4 respectively for 2-grams, 3-grams and 4-grams. The 3-gram backoff LM was generated using the Kneser-Ney discounting method without applying any kind of cutoffs.

The two models were combined by means of linear interpolation, generating a mixed model. The optimal interpolation weights were estimated via the Expectation-

Maximization algorithm (EM), using the ALERT-SR.devel dataset as a held-out corpus. The interpolation weights obtained were 0.796 for the newspapers LM and 0.204 for the Broadcast News LM. Figure 3.2 summarizes the details of the lexical and LM components.

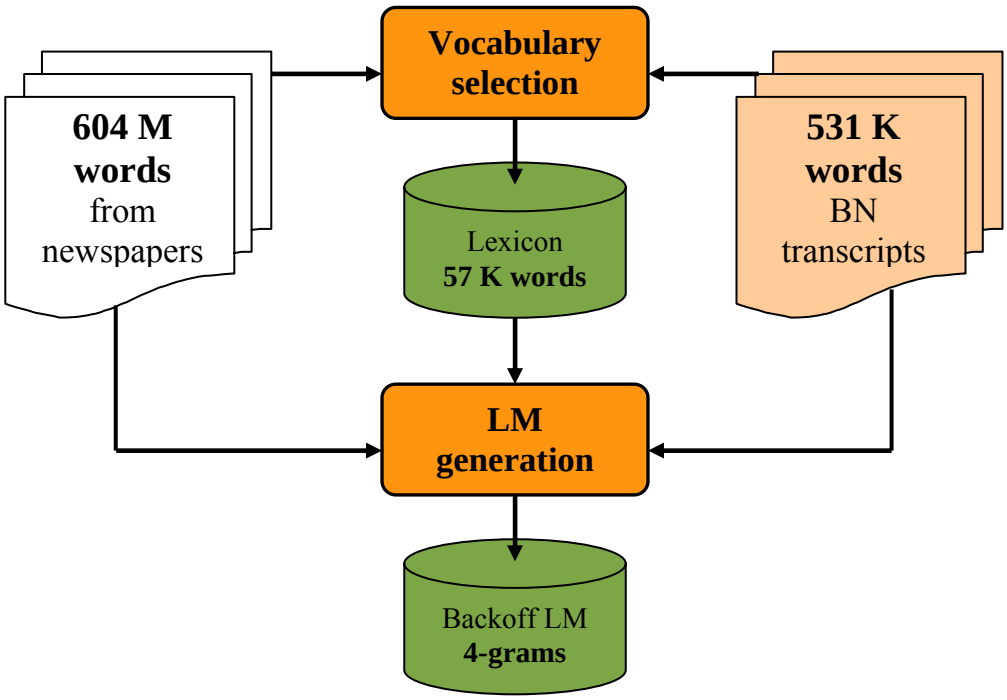

Figure 3.2: Baseline system: lexicon and LM details.

Finally, for the online implementation of our adaptation framework we used an entropy-based pruning technique [Stolcke, 1998], i.e., pruning all the *n*-grams that would increase the relative perplexity by less than a given threshold. Simultaneously, we pruned all the *n*-grams having probabilities lower than the corresponding backed-off estimates. This last pruning, applied to all the language models generated, is especially useful to create language models that can be correctly converted into probabilistic finite-state grammars.

3.2.4 Decoding

The decoder used under this baseline system is based on a weighted finite-state transducer (WFST) approach to large vocabulary speech recognition [Caseiro, 2003]. In this approach, the decoder search space is a large WFST that maps observation distributions to words. This WFST consists of the composition of various transducers representing

components such as: the acoustic model topology H; context dependency C; the lexicon L and the language model G. The search space is thus H ◦ C ◦ L ◦G, which is built "on-the-fly" [Caseiro and Trancoso, 2001][Caseiro and Trancoso, 2002], in opposition to traditional approaches that compile it outside of the decoder and use it statically during the decoding process.

3.2.5 Confidence Scoring

Confidence measures assign a degree of confidence to the recognized words. Hence, using these measures, ASR system can identify the words which are likely to be erroneous and the application using the ASR system can then use corrective actions. A variety of possibilities have been proposed in the past for the confidence score problem, including model-based, word-based, and utterance-based confidence measures (see e.g. [Carpenter et al, 2001]).

The AUDIMUS.media ASR system produces in its decoding process a set of confidence features for each recognized phone of the best hypothesis. These phone confidence features are then combined into word level confidence features, and finally a maximum entropy classifier is used to classify words as correct or incorrect. The maximum entropy classifier [Berger et al., 1996] combines all the word level confidence features according to:

$$P\left(correct\middle|w_i\right)=\frac{1}{Z\left(w_i\right)}\exp\left[\sum_{j=1}^{F}\lambda_j f_j\left(w_i\right)\right] \tag{3.1}$$

where w_i is the word, F is the number of features, f_j is a feature, $Z\left(w_i\right)$ is a normalization factor and λ_j the model parameters. For the AUDIMUS.media ASR system this classifier was trained on the ALERT-SR.train dataset.

In the proposed adaptation approaches presented in our work we used this confidence scoring process to select the most accurately recognized speech segments and reuse them as additional data for unsupervised adaptation purposes.

3.3 Evaluation Metrics

To evaluate the performance of the adaptation approaches proposed in this thesis we used various evaluation metrics. As stated in section 2.3.2, the most common metric for evaluating a language model is the perplexity. However, a meaningful comparison between perplexities of several models can only be made if they have the same vocabulary. For that reason, and since almost all our experiments evaluate and compare language models of different vocabularies, we have not used the perplexity as an evaluation metric.

Thus, for the experimental results we present here we used the WER to consistently evaluate and compare the relative language models performance. For these evaluations the NIST toolkit sclite [NIST, 2000] was used. This software calculates the WER metric given a set of reference sentences and a corresponding set of recognized sentences generated during the decoding process. To evaluate the performance of the vocabulary adaptation algorithms we used out-of-vocabulary (OOV) word rate as another metric.

3.4 Processing Tools

This section briefly describes the processing tools used to implement the various stages of the adaptation framework proposed in this thesis: language modeling, morpho-syntactic tagging and Information Retrieval extraction.

3.4.1 Language Modeling Toolkit

To generate the language models used in this work and evaluate their performance in terms of perplexity values, we used the SRILM Toolkit [Stolcke, 2002], a Language Model Toolkit designed to allow both production of and experimentation with statistical language models for speech recognition and other applications. SRILM is freely available for noncommercial purposes (http://www.speech.sri.com).

Beyond LM production and evaluation, the SRILM toolkit allows us to manipulate LMs in a variety of ways that we needed for our work:

- renormalize a model (recomputing backoff weights);

- approximate an interpolated n-gram with a standard word-based backoff LM;
- prune n-gram parameters, using an entropy criterion [Stolcke, 1998];
- prepare LMs for conversion to finite-state graphs by removing *n*-grams that would be superseded by backoffs.

Moreover, besides the standard word-based n-gram backoff LM models, the SRILM Toolkit supports creation and evaluation of several other LM types, most of them based on *n*-grams as basic building blocks: class-based models, cache models, skip language models, dynamically interpolated LMs, etc.

3.4.2 Morpho-syntactic Tagger

The information obtained by a morpho-syntactic tagging system can be relevant in several areas of natural language processing. For example, knowing the part-of-speech (POS) of a given word allows us to predict which words (or word classes) can occur in its neighborhood. That kind of information maybe useful in the language models used for speech recognition.

For the vocabulary selection algorithm derived in this work we used statistical information related to the distribution of POS tags of some training and adaptation datasets. Those datasets were morpho-syntactically tagged using a morpho-syntactic tagging system developed for the European Portuguese language [Ribeiro et al., 2004]. This morpho-syntactic tagger consists of two sequential modules as illustrated in figure 3.3:

- The morphological analysis module "Palavroso" [Medeiros, 1995], a morphological analyzer developed to address specific problems of the European Portuguese language. As output it gives all possible part-of-speech tags for a given word. If a word is not known, it tries to guess possible part-of-speech tags, always giving an answer;
- The disambiguation module "MARv" [Ribeiro, 2003], a morpho-syntactic ambiguity resolver whose architecture comprehends two components: a linguistic-oriented disambiguation rules component and a probabilistic disambiguation

component. The ambiguity is first reduced by the disambiguation rules component and then the probabilistic component produces a fully disambiguated output.

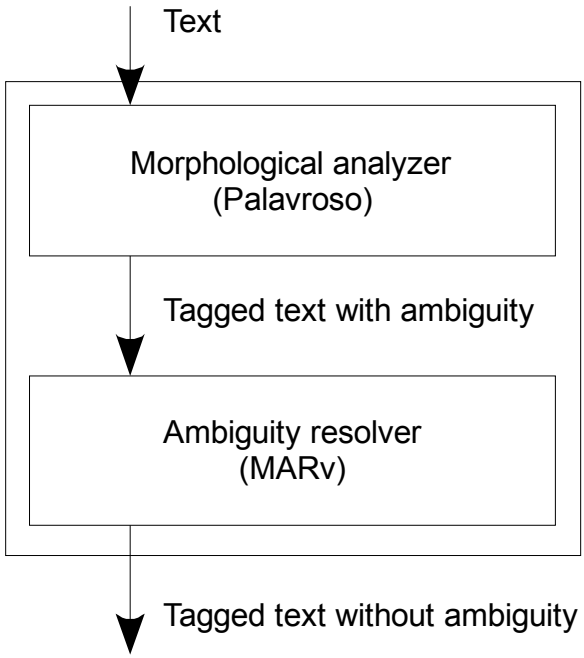

Figure 3.3: Architecture of the morpho-syntactic tagging system.
(extracted from [Ribeiro et al., 2004])

The information coded by the tagset is presented in appendix A.

3.4.3 Information Retrieval Engine

For our framework we looked for an Information Retrieval engine addressing the following requirements: the system architecture should support large-scale text databases, multiple databases, concurrent indexing and querying, fast indexing, different retrieval models, and relevance feedback models support. According to these requirements we chose the Indri search engine [Strohman et al., 2005]. The retrieval model implemented in the Indri search engine combines the best features of inference networks and language modeling in an architecture designed for large-scale applications. The Indri query language can handle both simple keyword queries and extremely complex queries, allowing complex phrase

matching, synonyms, weighted expressions, Boolean filtering, numeric (and dated) fields, and the extensive use of document structure (fields), among others.

Indri is part of the Lemur Toolkit [LEMUR, 2007], an open-source toolkit for language modeling and Information Retrieval.

3.5 Summary

In this chapter we briefly covered the resources relevant to this thesis. We first presented a description of the two corpora used for our experimental studies and to train the AUDIMUS.media ASR system, the baseline system used for those studies. We described each of the components of the system including the lexicon and language model components.

The AUDIMUS.media ASR baseline system is part of a closed-captioning system of live TV broadcasts in European Portuguese that is daily producing online captions for the main news show of one Portuguese Broadcaster - RTP. The current BN ASR system is working in "real time" in a P4 dual core machine @ 2.8 GHz computer with 2 GB memory.

We concluded this chapter by briefly describing other processing tools we used: the modular morpho-syntactic tagger, the language model toolkit (SRILM) and the IR search engine (INDRI).

4

Vocabulary Selection

The daily and real-time transcription of broadcast news is a challenging task both in terms of acoustic, lexical and language modeling. To achieve optimal performance in news transcription, several problems have to be overcome: variety of acoustic conditions, many different speaking styles (from spontaneous conversation to prepared speech close in style to written texts), and topic changing over time leading to unlimited vocabulary. Particularly, when transcribing BN data in highly inflected languages, the vocabulary growth leads to high OOV word rates [Geutner et al., 1998].

In these daily transcription systems the appearance of some new important events brings an increase of OOV. This increase of unknown words leads to degradation in recognition performance. This way, lexical coverage of a vocabulary should be as high as possible to minimize the side effects of OOV on system recognition performance. As stated in [Bigi et al., 2004], vocabulary optimization is mainly dependent on the task, the amount of training data used, and the source and recency of that data. Actually, assuming an open task like the BN data transcription, the topic changing over time leads to unlimited vocabulary. Hence, while a large vocabulary may be desirable from the point of view of lexical coverage, there is also the additional problem of increased acoustic confusability [Rosenfeld, 1995]. Since new words are constantly being introduced into common usage, it is impossible to ever have a complete vocabulary of all spoken words. Thus, the treatment of new lexical items is an essential element.

Several innovative techniques can be exploited to reduce those problems. Various research works show that improvement in system accuracy can be obtained by dynamically

adapting the vocabulary and language model based on additional training data resources. The use of news shows specific information, such as topic-based lexicons, pivot working script, and other sources such as the online written news daily available in the Internet can be added to the information sources employed by the ASR component [Schwarm et al., 2004]. Using multi-phase recognition approaches, word hypotheses can be improved by using adaptive vocabularies and language models [Chen et al., 2004] [Boulianne et al., 2006] [Oger et al., 2008].

This chapter describes in detail the work done in the scope of our thesis, where we are exploring the use of additional sources of information for vocabulary selection of an European Portuguese broadcast news transcription system. Since the vocabulary optimization problem is mainly dependent on the specific linguistic characteristics of the target language, in sections 4.1 and 4.2 we present an analysis of the vocabulary growth, coverage and OOV words for the European Portuguese language using the datasets described in chapter 3. Based on that analysis and its conclusions, we devised new vocabulary selection strategies which take into account the specific characteristics of the European Portuguese language. Thus, in section 4.3 and 4.4 we give a detailed description of the different approaches we derived to improve the automatic selection of the word list for the ASR vocabulary done on a daily basis, presenting some evaluation and comparison results.

4.1 Analysis of Vocabulary Growth and Coverage

A major part of building a language model is to select the vocabulary of the ASR component which will have maximal coverage for the expected task/domain. Thus, the appearance of OOV words during the recognition process is closely related to the way the system vocabulary is chosen. Hence, the growth of the vocabulary as a function of the training corpora plays an important role in the magnitude of the OOV problem. In [Hetherington, 1995] and [Rosenfeld, 1995] the authors present extensive studies related to the vocabulary growth and coverage for different domains. Based on its studies, Hetherington classified the training corpora in three different groups:

- human-to-human written communication, which represents the corpora with the highest vocabulary growth. The Web text news corpus (WEBNEWS-PT) used in this work belongs to this group;
- human-to-human spoken communication, representing the corpora with medium vocabulary growth. The broadcast news corpus ALERT-SR is an example which follows in this category;
- and finally, the human-to-machine communication, encompassing corpora related to spoken dialog systems, being the group with the lowest rate of vocabulary growth.

In figure 4.1 we present the vocabulary growth for the two corpora used in this thesis. For the broadcast news domain (ALERT-SR corpus), the vocabulary growth is plotted for the two training datasets (pilot and train) consisting of about 500K word tokens. For analysis and comparison purposes we plotted the vocabulary growth for a random subset of WEBNEWS-PT corpus, consisting of about 1.5M word tokens. As one can observe, for the WEBNEWS-PT corpus the vocabulary growth is faster than for the ALERT-SR corpus. For a corpus size of about 0.5M word tokens, the ALERT-SR corpus has a vocabulary size of 26K words, while the vocabulary size for the WEBNEWS-PT is 38K, i.e. about 46% more.

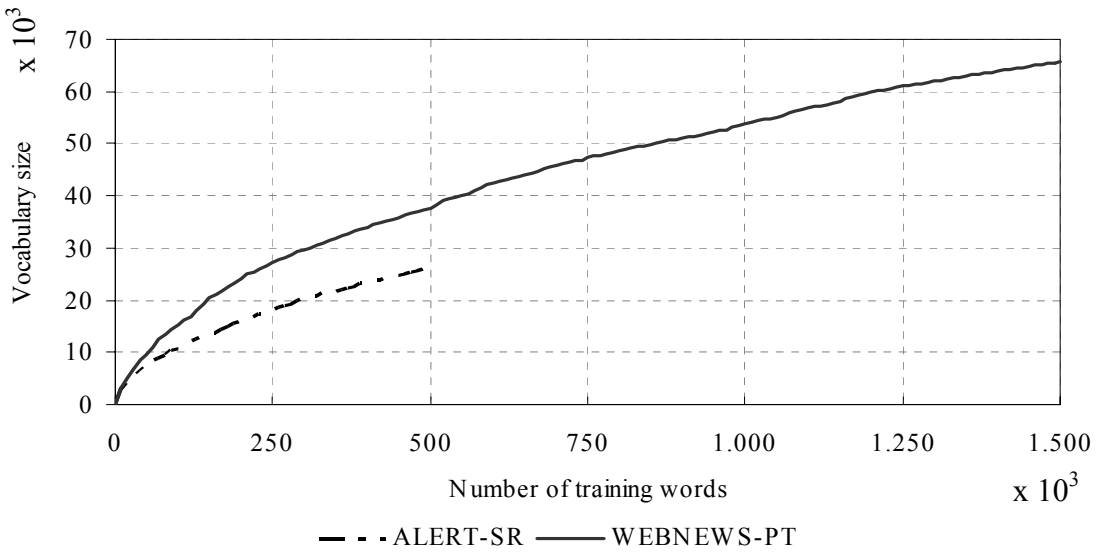

Figure 4.1: Vocabulary growth for the two corpora used in our work: Web text news corpus and broadcast news corpus (pilot and train datasets).

Moreover, another important aspect to take in account for vocabulary design is to identify and rank the most relevant vocabulary words in order to improve the coverage rate of that vocabulary on unseen data. Figure 4.2 shows the coverage statistics related to the ALERT-SR.11march and WEBNEWS-PT.11march datasets for the baseline vocabulary of 57K words. For this vocabulary the OOV rate measured on the broadcast news dataset averages 1.25%, while the average OOV rate for the Web text news dataset is about 3.16%. These results are consistent with Hetherington and Rosenfeld findings that vocabulary growth and coverage is domain dependent.

The coverage results show a small increase in the OOV rate after the March 11th. We would expect this kind of behavior, with a clear and strong topic change, mainly due to the unexpected and awful events occurring on March 11th of 2004 in Madrid. To figure out what kind of words mainly contributes to these OOV rates, and which adaptation procedures we should pursuit in order to better address this problem specific of highly inflected languages such as the European Portuguese, we derived various analyses at the OOV level which are presented in the next section.

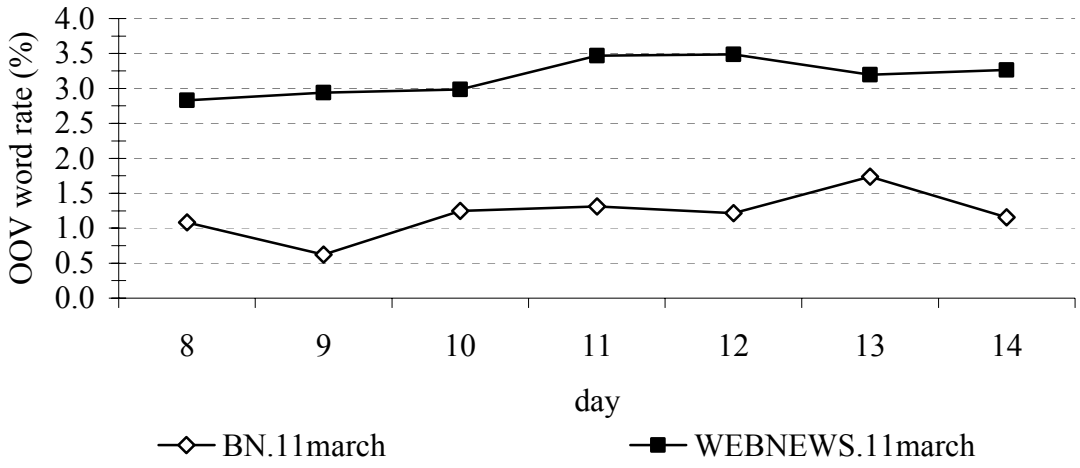

Figure 4.2: OOV rate in the ALERT-SR.11march and WEBNEWS-PT.11march datasets for the 57K words baseline vocabulary.

4.2 Analysis of OOV Words

In this section, we look at some characteristics of OOV words in the broadcast news dataset (ALERT-SR.11march).

First, we examine their classification into part-of-speech (POS) classes. In table 4.1 we break down OOV words into three different categories using the morpho-syntactic tagging system developed for the European Portuguese language and briefly described in section 3.4.2: names (including proper and common names), adjectives and verbs. Other type of words, such as function words, are absent from the list shown in table 4.1 because almost all those words are already in the 57K baseline vocabulary. We simply merged all together ("Others" category in table 4.1)

Class	8th	9th	10th	11th	12th	13th	14th	Week
Names	27.3	15.4	19.0	20.6	28.2	43.1	34.8	28.8
Adjectives	10.1	7.7	18.1	16.8	16.4	8.9	9.6	13.2
Verbs	61.6	69.2	62.9	57.9	53.6	47.2	53.9	56.2
Others	1.0	7.7	0.0	4.7	1.8	0.8	1.7	1.8

Table 4.1: Distribution (in %) of OOV words by POS-classes in the ALERT-SR.11march dataset.

According to findings reported in the literature, OOV words are mostly names. In [Hetherington, 1995], [Bazzi, 2002] and [Allauzen and Gauvain, 2005] a strong correlation between names and OOV words is reported. A similar conclusion is reported in [Palmer and Ostendorf, 2005], with names accounting for 43.66% of the OOV word types. Hence, as a first idea, we would be expecting to observe a similar behavior for ALERT-SR.11march dataset, i.e. a strong relation between names and OOV words, especially for this specific week with new and infrequent words appearing (train station names, terrorist names, journalist names, etc.). However, as one can observe from table 4.1, verbs make up for the largest portion of OOV words. In fact, although verbs represent only 17.5% (see figure 4.3) of the words in the ALERT-SR.11march dataset, they account for 56.2% of the OOV words. Moreover, in this dataset, verbs are also very frequently the source of

recognition errors, representing the largest portion of wrongly recognized words - about 25% (see figure 4.4).

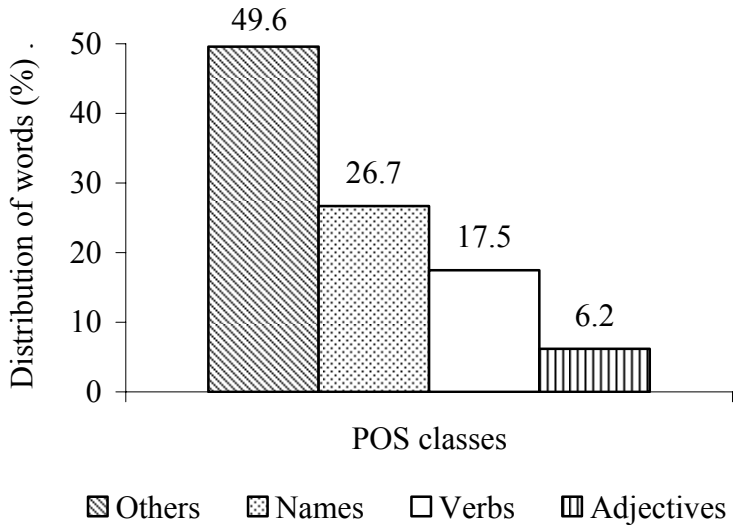

Figure 4.3: Distribution (in %) of words by POS-classes in the ALERT-SR.11march dataset.

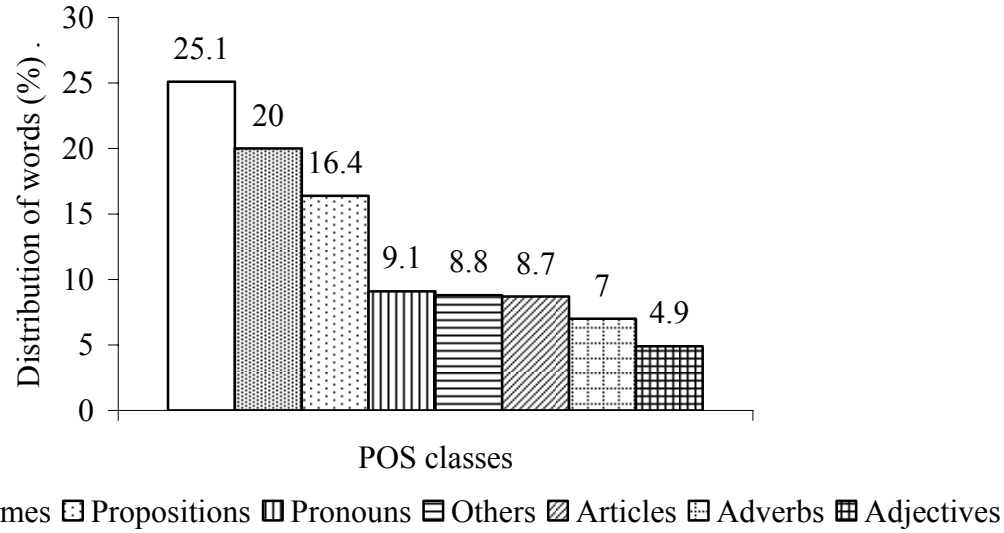

Figure 4.4: Distribution (in %) by POS-classes of the words wrongly recognized in the ALERT-SR.11march dataset. Recognition results obtained with the baseline system.

In a second analysis, and since our adaptation proposal is to take advantage of contemporary written news to dynamically adapt the system vocabulary, we examined the effect of augmenting the vocabulary with new words found in the same day of each tested

BN show. As we have described in section 3.2, the WEBNEWS-PT.11march dataset had an average size of about 280 K tokens collected per day. Thus, taking into account these written text news collected for each day and the 57K words baseline vocabulary, an average of 5K new words was found on a daily basis, accounting for an upgraded vocabulary of 62K words.

	8th	9th	10th	11th	12th	13th	14th
OOV reduction	20.2	15.4	21.6	29.6	36.4	20.3	33.9

Table 4.2: OOV word reduction (in %) in the ALERT-SR.11march dataset by adding new words found in written news on a daily basis.

From table 4.2, one can observe an OOV word reduction ranging between 15.4% (for March 9th) and 36.4% (for March 12th), with an average value around 28.6%. The graph in figure 4.5 gives us an overview about the kind of words (POS-classes) we covered by adding those 5K extra new words. From that, we conclude that a significant OOV word reduction was obtained in the class of names. Remarkably, on the news show of March 12th more than 72% reduction was achieved in the class of names.

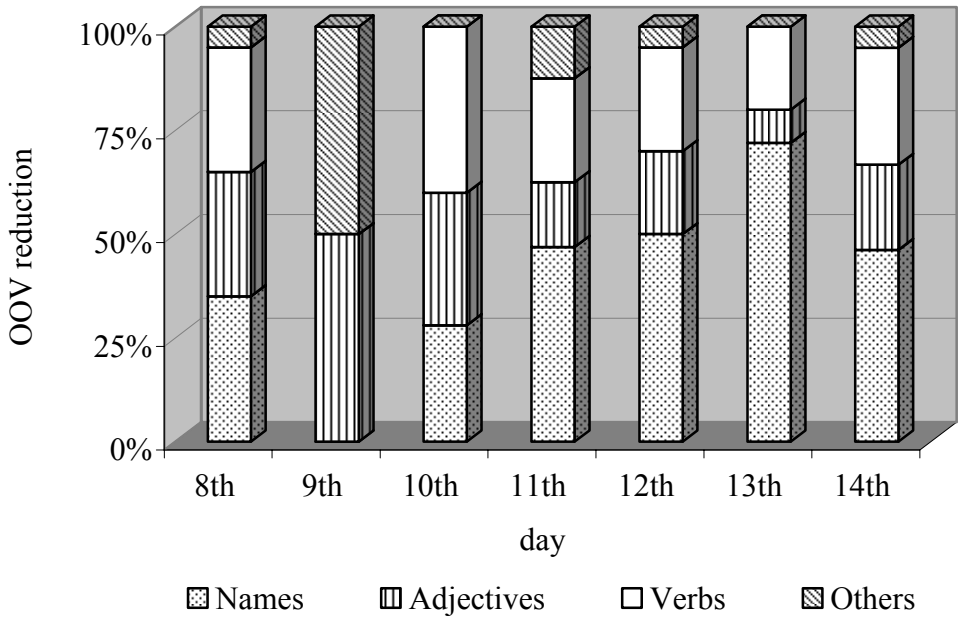

Figure 4.5: OOV word reduction (in %) by POS-classes in the ALERT-SR.11march dataset when adding new words found in written text news on a daily basis.

Based on the above observations, we conclude that the strategy of using contemporary written text news to adapt the baseline vocabulary seems to be useful specially to cover the new names appearing over time. However, even though verbs represent the largest portion of OOV words, the reduction for this class by using contemporary written texts is not so significant. Moreover, the differences in terms of vocabulary growth and coverage for different domains and time periods, makes it necessary to devise new vocabulary selection strategies which take into account those specific characteristics.

In the next sections, we describe the *vocabulary optimization* techniques proposed in this thesis, and their evaluation.

4.3 Vocabulary Adaptation based on Linguistic Knowledge (Lemmas)

As stated before, just generically increasing the system vocabulary size can improve the accuracy for many common words but degrades the recognition rate for less common words [Rosenfeld, 1995]. In our preliminary work [Martins et al., 2005] we tried using a large vocabulary of 213K words selected with an *ad-hoc* approach (all words from the training corpora occurring more than 15 times), obtaining an OOV word rate reduction of 67%. However, this approach does not solve the problem of newly appearing words and infrequent words related to some important events, which are critical and therefore need to be recognized accurately. This is especially true for the broadcast news domain due to the large variety of topics discussed over time. Moreover, looking at tables 3.3 and 3.4 in section 3.1.1, one observes a maximum of 2.2K word types occurring by day. Thus, defining a more rational approach to expand the vocabulary other than by simple frequency of occurrence is need.

4.3.1 Vocabulary Adaptation Algorithm

In [Martins et al., 2006] we proposed a procedure for dealing with the OOV problem by dynamically increasing the baseline system vocabulary, reducing the impact of linguistic

differences over time. Based on the OOV analysis, we focused our work in correcting errors resulting from OOV words mainly on verbs class. Our approach to compensate and reduce the OOV word rate related with verbs was supported by the fact that almost all the OOV verbs were inflections of verbs whose lemmas were already among the lemmas set L_d of the verbs found in contemporary written news of each day d. In fact, using the morphological tool reported in section 3.4.2 we performed a lemmatization over all the words of ALERT-SR.11march and WEBNEWS-PT.11march datasets, concluding that in average 83.1% of the verbal lemmas belonging to the OOV words were present in L_d set (table 4.3).

8th	9th	10th	11th	12th	13th	14th	Week
88.5	77.8	85.8	85.2	84.7	82.7	77.0	**83.1**

Table 4.3: Percentage of verbal lemmas, derived from the OOV verbs present in the ALERT-SR.11march dataset, included on the verbal lemmas set L_d derived from the written news.

On tables 4.4 and 4.5 we present some examples of those verbal lemmas on both ALERT-SR.11march dataset and WEBNEWS-PT.11march dataset (examples selected from the news show of March 13th).

Words	Morpho-syntactic Information	Verbal Lemma
acautelar	[V=f=1s] [V=f=3s] [V=sf1s] [V=sf3s] [V=n]	Acautelar
descarta	[V=ip3s] [V=m2s]	Descartar
ilegalizada	[A=pfs] [V=p==sf]	Ilegalizar
intriga	[Ncfs] [V=ip3s] [V=m2s]	Intrigar
sereno	[Ncms] [A=pms] [V=ip1s]	Serenar

Table 4.4: Examples of verbs present in the WEBNEWS-PT.11march dataset (March 13th).

OOV Words	Morpho-syntactic Information	Verbal Lemma
acauteladas	[A=pfp] [V=p==pf]	Acautelar
descartou	[V=is3s]	Descartar
ilegalizado	[A=pms] [V=p==sm]	Ilegalizar
intrigou	[V=is3s]	Intrigar
serenou	[V=is3s]	Serenar

Table 4.5: Examples of OOV verbs present in the ALERT-SR.11march dataset (March 13th).

The results of table 4.3 motivated our idea of using this linguistic behavior to automatically expand the baseline vocabulary. Thus, the baseline vocabulary of each day d is automatically extended by adding the following words:

- all the new words appearing in the written texts of day d, and
- all the verbal infections observed in the language model training corpora and whose lemmas belong to L_d.

Thus, supposing the baseline vocabulary V_0, the proposed adaptation approach is performed on a daily basis according to the following procedure (see figure 4.6):

1. Every day d, online written news $O(d)$ are downloaded from the Internet;
2. A vocabulary list V_1 consisting of the words found in $O(d)$ is created;
3. The words of V_1 are grammatically classified and lemmatized, and the list of verbal lemmas L_d generated;
4. A new vocabulary list V_2 is generated by selecting all the verbal inflections observed in the language model training corpora and whose lemmas belong to L_d;
5. Vocabulary lists V_0, V_1 and V_2 are merged together to form the adapted vocabulary list V_3.

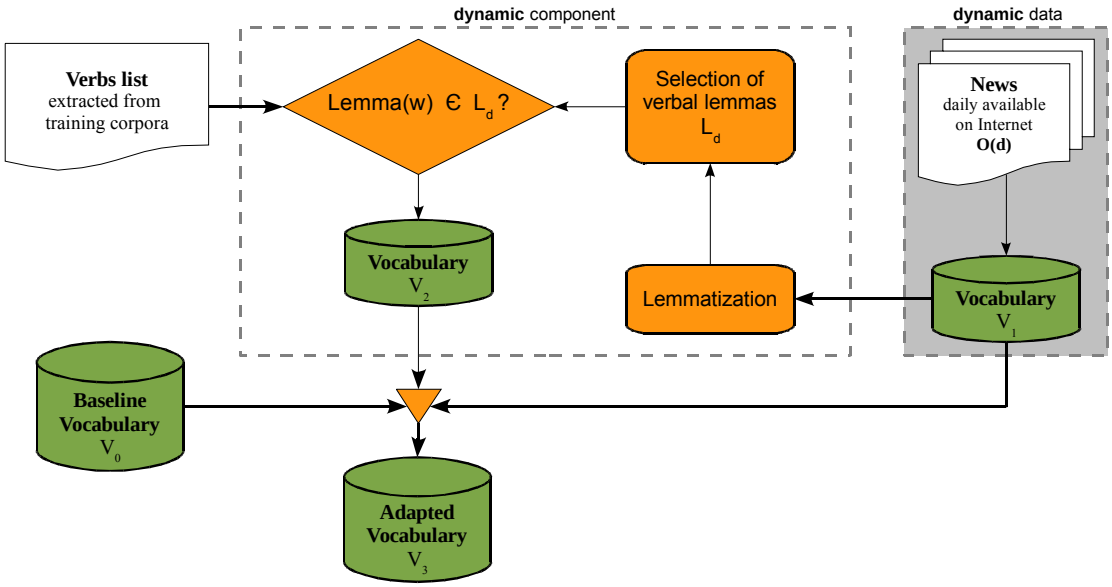

Figure 4.6: Vocabulary adaptation procedure based on linguistic knowledge (lemmas).

Next we present some evaluation results by applying this vocabulary adaptation on the ALERT-SR.11march dataset.

4.3.2 Evaluation Results

To compare the performance of this vocabulary adaptation approach we used the OOV word rate as an evaluation metric. Hence, we applied it on the seven news shows of the ALET-SR.11march dataset, comparing the OOV word rate for different vocabulary sets:

- V_0: baseline vocabulary (V_0) consisting of 57K words;
- $V_0 + V_1$: baseline vocabulary extended with V_1, the list of words appearing on the written news $O(d)$ extracted from the Internet for each day d. In this case the baseline vocabulary was expanded by an average of 5K new words each day, giving a final vocabulary of 62K words per day;
- $V_3 = V_0 + V_1 + V_2$: baseline vocabulary extended with V_1 and V_2 according the procedure here proposed. Applying this adaptation approach, the baseline system vocabulary was expanded by an average of 43K new words each day, resulting in a final vocabulary of 100K words per day;

- $V_0 + V_1 + V_2$': baseline vocabulary extended with V_1 and V_2'. The selection of V_2' is done only by means of word frequency and such that the final vocabulary size of $V_0 + V_1 + V_2$' is the same as $V_0 + V_1 + V_2$, i.e., 100K words per day.

In figure 4.7 we plot the daily OOV word rate on the ALERT-SR.11march dataset using those four vocabularies. As one can observe this adaptation procedure generated a significant improvement in terms of OOV word rate, which was reduced in average by 65.7%, i.e. from 1.25% to 0.43%. Moreover, this approach outperformed all the other methods, being more effective than the common word frequency approach. This improvement was almost uniform across all the seven BN test shows, with the lemmas-based approach being outperformed only for the BN show of March 13th (see figure 4.7). In fact, from table 4.1 one can observe that BN show of March 13th has the lowest percentage of OOV words as verbs.

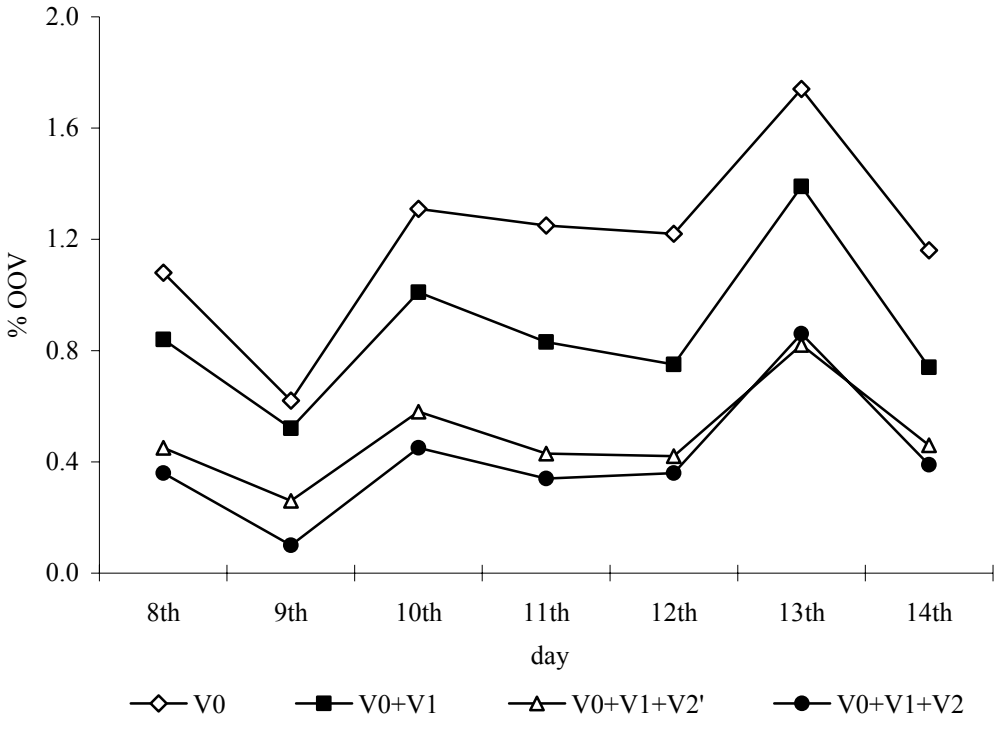

Figure 4.7: OOV word rate comparison for different vocabularies: V_0 (57K), $V_0 + V_1$ (62K), $V_0 + V_1 + V_2$' (100K) and $V_0 + V_1 + V_2$ (100K).

As stated before, the procedure we presented here tries to deal with the OOV problem by dynamically increasing the baseline vocabulary over time. As we would expect, by additionally applying the morphological analysis over the verbal inflections found in the written news, we achieved an OOV word rate reduction of 65.7% (derived vocabulary $V_0 + V_1 + V_2$), against the 28.6% reduction we could obtain by using only the new words found in the written news (derived vocabulary $V_0 + V_1$).

Vocabulary	Names	Adjectives	**Verbs**	Others	Total
Baseline (V_0)	187	87	376	12	662
Adapted ($V_0 + V_1$)	98	49	324	2	473
Adapted ($V_0 + V_1 + V_2$)	98	47	80	2	227
OOV reduction:	47.6%	46.0%	**78.7%**	83.3%	**65.7%**

Table 4.6: Distribution of OOV words using the baseline and adapted vocabularies for all the seven BN shows of ALERT-SR.11march dataset.

Table 4.6 shows the distribution of OOV words by grammatical classes (POS) for both baseline vocabulary and extended vocabularies, which indicates a significant reduction on the OOV words classified as verbs in case of V_3 vocabulary. In fact, using the adapted vocabulary $V_0 + V_1$ we could obtain an average reduction of 13.8%, against 78.7% reduction when applying the proposed vocabulary adaptation algorithm.

4.3.3 Summary

Using the broadcast news dataset formed by the seven news shows (ALERT-SR.11march) we performed some analyses of the type OOV words obtained when applying the baseline vocabulary. From these analyses, we were able to conclude that verbs turned to be the most significant grammatical class in terms of OOV. Hence, a vocabulary adaptation algorithm was proposed that showed to be effective to cope with this specific problem, allowing an average relative reduction of more than 65% compared to the baseline vocabulary.

However, the proposed approach assumes an *a priori* selected static list of words - the baseline vocabulary, just adding new words on a daily basis. This way, the system

vocabulary is always extended resulting in a vocabulary with an average size of 100K words. In the following section a new algorithm for selection and adaptation is proposed. It allows defining the size of the target vocabulary, selecting it from scratch.

4.4 Vocabulary Selection using Morpho-Syntactic Tagging (POS)

When various training corpora of different origins, sizes and recencies are available, we face the problem of how to infer the target vocabulary. In our case, we would like to define an automatic and optimized procedure to daily select the system vocabulary from three different corpora: an out-of-domain dataset (WEBNEWS-PT.train), an in-domain dataset (ALERT-SR.train+pilot) and the adaptation dataset daily collected from the Internet (WEBNEWS-PT.11march). For this purpose, in [Martins et al., 2007] we introduced a modified vocabulary selection technique that takes into account the differences in style across the various corpora, especially in case of written versus spoken style.

In a first step, and using the same morpho-syntatic analysis tool as before, we annotated both the in-domain corpus (ALERT-SR.train+pilot) and a segment of the out-of-domain corpus (WEBNEWS-PT.train) with a similar size, i.e., about 531K word tokens. This segment consisted of news articles randomly selected from the WEBNEWS-PT.train dataset.

In figure 4.8, we summarize the POS statistics obtained for both datasets by breaking down words into four main classes: names (including proper and common names), verbs, adjectives and adverbs. Other type of words, such as functional words, are absent from the list shown in figure 4.8 because they represent closed grammatical classes in the European Portuguese language. These statistics are related to word types and not word tokens, i.e., only unique occurrences of a word/class are counted. As one can see, there is a significant difference in POS distribution when comparing in-domain and out-of-domain datasets, especially in terms of names and verbs. For in-domain data we observe a significant increment (from 30.5% to 36.9%) in the relative percentage of verbs when compared with the out-of-domain data, with the percentage of names decreasing from 45% to 40.6%.

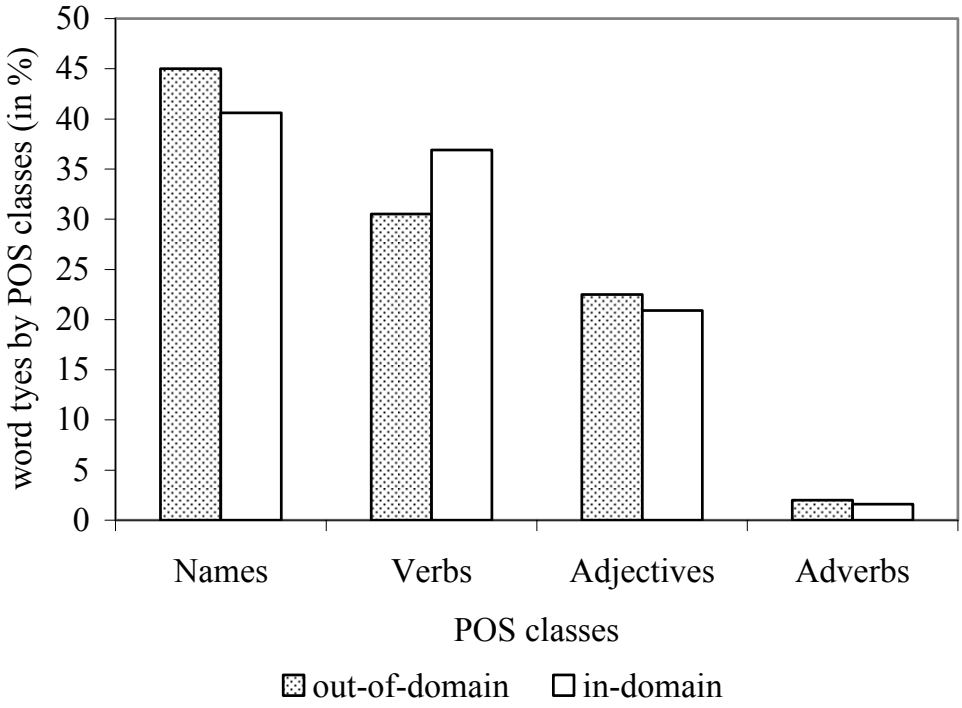

Figure 4.8: Distribution of words types by POS classes (in %).

Based on the above observations, we proposed a new approach for vocabulary selection that uses the part-of-speech word classification to compensate for word usage differences across the various training and adaptation corpora. This approach is based on the hypothesis that the similarities between different domains can be characterized in terms of style (represented by the POS sequences). In [Iyer et Ostendorf, 1997] these similarities have already been integrated to more effectively use out-of-domain data in sparse domains by introducing a modified representation of the standard word n-gram model using part-of-speech labels that compensates for word usage differences across domains. So, in this new approach, instead of simply adding new words to the fixed baseline system vocabulary, as the previously proposed approach, we use now the statistical information related to the distribution of POS word classes on the in-domain corpus to dynamically select words from the various training corpora available.

4.4.1 Vocabulary Selection Algorithm

Assuming we want to select a vocabulary V with $|V|$ words from n training corpora T_j, with $j = 1,\ldots,n$. The proposed approach can be summarized as follows:

1. Definition of the POS classes to use

   In our implementation we used the following set of POS classes:

   $$POSset = \{\text{names, verbs, adjectives, adverbs}\} \quad (4.1)$$

   All the remaining words (mainly functional words) are automatically added to the vocabulary. In fact, in the training corpus used in this work we obtained only 468 words which POS class did not belong to $POSset$.

2. Estimation of POS distribution using an in-domain corpus

   Using an in-domain dataset the distribution of words by POS classes, $M(p)$ with $p \in POSset$, is computed through the maximum likelihood estimation (MLE). Thus, the in-domain is POS-tagged, and its statistics used to estimate the class probability as $M(p) = \dfrac{N(p)}{\sum_{i \in POSset} N(i)}$, with $N(p)$ being the count of occurrences of class p on that dataset.

3. Computation of normalized counts

   Let $c_{i,j}$ be the counts from each one of the available training corpus T_j, for the word w_i. Due to the differences in the amount of available data for each training corpus, we start by normalizing the counts according to their respective corpus length, getting $\eta_{i,j}$ as the normalized counts. The Witten-Bell discounting strategy is used to ensure non-zero frequency words in the normalization process.

4. Estimation of a word weighting factor

From the normalized counts $\eta_{i,j}$ we want to estimate some kind of weighting factor η_i for each word w_i in order to select a vocabulary from the union of the vocabularies of T_1 through T_n that minimizes the OOV word rate for the in-domain task. In [Venkataraman and Wang, 2003] this weighting is obtained by means of linear interpolation of the different counts, with the mixture coefficients calculated in order to maximize the probability of the in-domain corpus. In our work we use a similar method but simply assigning identical values to all the mixture coefficients. Hence,

$$\eta_i = \sum_{j=1}^{n} \lambda_j \eta_{i,j} \quad \text{with} \quad \lambda_j = \frac{1}{n} \tag{4.2}$$

5. Generation of an ordered word list W

   All the words w_i are sorted in descending order according to the weighting factor η_i.

6. Selection of $|V|$ words from the word list W

   According to $M(p)$, the number of words selected from each class p will be $|V| \times M(p)$. Hence, for each class p, the first $|V| \times M(p)$ words of W belonging to class p are selected and included in the target vocabulary V.

   However, since a word can belong to more than one class, the first run of this process can produce a vocabulary list with less than $|V|$ words. In that case, the selection process is iterated until the target number of words is achieved.

4.4.2 Evaluation Results

As before, we used the OOV word rate as an evaluation metric to evaluate the performance of this new approach, applying it on the seven news shows of the ALET-SR.11march dataset. As a first test, we started by comparing the OOV word rate for four different vocabulary sets:

- **Baseline**: baseline vocabulary consisting of 57K words;

- **Baseline+day**: baseline vocabulary extended with the new words appearing on the written news $O(d)$ extracted from the Internet for each day d. In this case the baseline vocabulary was expanded by an average of 5K new words each day, giving a final vocabulary of 62K words per day;
- **Adapted_WF**: selecting a new vocabulary based on all the training/adaptation corpora and using word frequency as the only selection criteria. In this case, the vocabulary is selected on a daily basis, using the two training datasets (in-domain dataset ALERT-SR.train+pilot and out-of-domain datatset WEBNEWS-PT.train) and the adaptation dataset formed by the written news $O(d)$ collected day-by-day (WEBNEWS-PT.11march);
- **Adapted_POS**: selecting a new vocabulary based on all the training corpora and using the new approach proposed in this section. As in Adapted_WF approach, the two training datasets, plus the adaptation one, were used in the vocabulary selection process (in steps 3 and 4 of our algorithm). To estimate the distribution of POS classes (step 2) we used the in-domain dataset ALERT-SR.train+pilot. Table 4.7 presents the $M(p)$ distribution used.

p	names	Verbs	adjectives	adverbs
$M(p)$	40.6	36.9	20.9	1.6

Table 4.7: POS distribution used in our experiments.

In both Adapted_WF and Adapted_POS approaches we used $|V| = 62K$ in order to make results comparisons with Baseline+day approach. Notice that the results presented in table 4.8 correspond to macro-averages over the 7 news shows of the test dataset. As one can observe from table 4.8, the new proposed Adapted_POS approach yields a relative average reduction of 37.8% in OOV word rate, when compared to the results obtained with the baseline vocabulary. Moreover, this approach outperformed all the other methods, being more effective than the common word frequency approach (from 0.82% to 0.78%). This improvement was almost uniform across all the seven BN test shows, with the

Adapted_POS approach being outperformed only for the BN show of March 10th (see figure 4.9).

Approach	%OOV	%reduction
Baseline (57K)	1.25	-
Baseline+day (62K)	0.89	28.5
Adapted_WF (62K)	0.82	34.1
Adapted_POS (62K)	0.78	37.8

Table 4.8: Average OOV word rate for the ALERT-SR.11march dataset applying different methods of vocabulary selection ($|V| = 62K$).

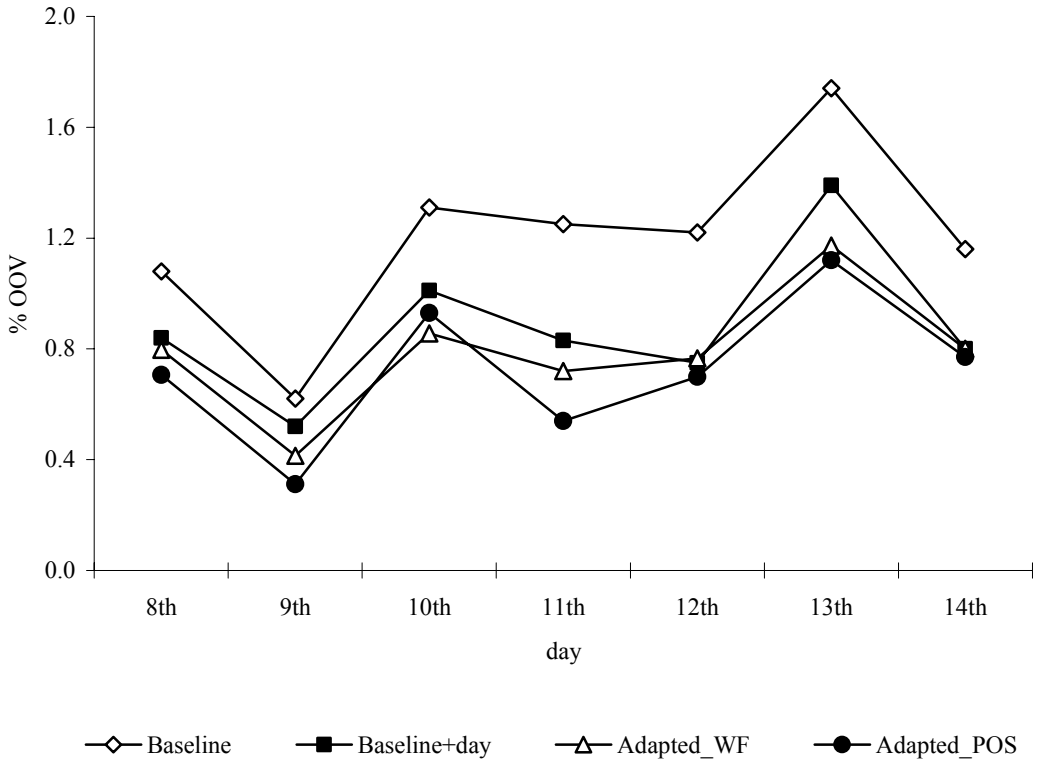

Figure 4.9: OOV word rate for the seven BN shows of the ALERT-SR.11march dataset when applying different methods of vocabulary selection ($|V| = 62K$).

On table 4.9 we present the distribution (in percentage) of words by POS classes for the different vocabularies produced by each one of the selection procedures. One can clearly observe the differences in POS class distribution. By examining the distribution of POS classes and OOV word rates over the test shows for the various vocabulary selection approaches, a correlation between these two values is observed. In fact, the new approach selects words in a more balanced way, especially in the case of names and verbs classes.

Vocabulary	Names	Verbs	Adjectives	Adverbs
Baseline+day	**61.6**	**22.5**	14.1	1.8
Adapted_WF	**56.3**	**26.9**	15.6	1.2
Adapted_POS	**40.6**	**36.9**	20.9	1.6

Table 4.9: Distribution (in %) of words by POS classes for different vocabularies ($|V| = 62K$).

To better understand the performance of this new vocabulary selection procedure for different values of $|V|$ (vocabulary size), we calculated the OOV word rate results for vocabularies of 5K, 25K, 50K, 100K, 150K and 200K words (see table 4.10).

We compared the common word frequency approach with the proposed POS-based approach. Results in table 4.10 show the relative good performance of the Adapted_POS approach for the selection of large-sized vocabularies. Furthermore, as we would expect, for the selection of small vocabularies better results are achieved by using the Adapted_WF method. As one can see, for the vocabulary sizes of 5K and 25K words the Adapted_POS approach does not perform so well. After analyzing the type of OOV words generated by both approaches, one could conclude that for small values of $|V|$, the probability value $M(adverbs)$ is very small, and consequently the Adapted_POS approach does not include in vocabulary some highly frequent adverbs (for example, “demasiado” – too much, “particularmente” – particulary, “todavia” – nevertheless, …).

$\|V\|$	Approach	%OOV	%reduction
5K	Adapted_WF	10.23	
	Adapted _POS	10.89	**-6.4**
25K	Adapted _WF	2.45	
	Adapted _POS	2.49	**-1.8**
50K	Adapted _WF	1.11	
	Adapted _POS	1.05	**5.8**
62K	Adapted _WF	0.82	
	Adapted _POS	0.78	**5.5**
100K	Adapted _WF	0.39	
	Adapted _POS	0.36	**6.8**
150K	Adapted _WF	0.22	
	Adapted _POS	0.20	**8.8**
200K	Adapted _WF	0.14	
	Adapted _POS	0.13	**7.9**

Table 4.10: Word Frequency vs. POS approach results for different values of $|V|$.

4.4.3 Summary

In this section we described a dynamic vocabulary adaptation framework that tries to optimize the trade-off between the expected OOV word rate and the number of added words. It uses POS class information about an in-domain training corpus to select an optimal vocabulary for domain-specific language modeling tasks. When applied to a daily and real-time broadcast transcription task, this procedure showed to be effective in reducing the OOV word rate (a relative average reduction of more than 37%) when compared with the one obtained for the baseline vocabulary, with an increment of 5K words in the vocabulary size (from 57K to 62K). Even with a vocabulary of 50K words, we could get a relative average decrease of 16% (from 1.25% to 1.05%) in the OOV word rate.

In the next section, we compare the performance of the two vocabulary adaptation approaches proposed in this work.

4.5 Comparison of Lemmas-based and POS-based Algorithms

To compare the proposed algorithms for vocabulary selection/adaptation (lemmas and POS approaches) we used two evaluation metrics: OOV word rate and WER over the seven BN shows of ALERT-SR.11march dataset. As stated before, applying the lemmas-based procedure, the baseline vocabulary of 57K was expanded by an average of 43K new words for each day, giving a final vocabulary size of 100K words per day. Thus, to fairly compare the two approaches, we used a vocabulary size of 100K for both.

4.5.1 OOV Rate Results

As one can observe from figure 4.10, the proposed POS-based approach yields a relative reduction of 71.2% in OOV word rate, when compared to the results obtained with the baseline vocabulary, which shows the good performance of this new selection/adaptation technique.

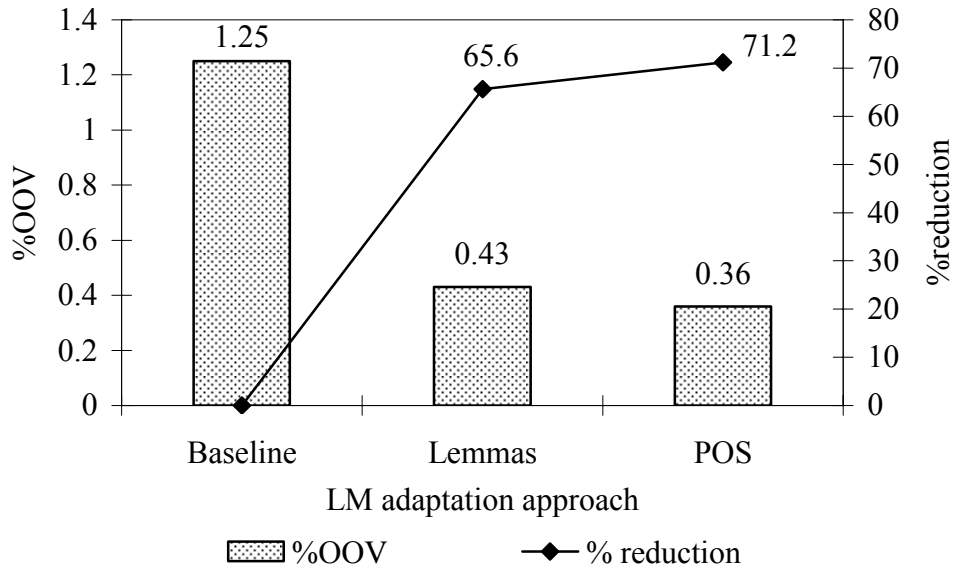

Figure 4.10: OOV word rate for the seven BN shows of the ALERT-SR.11march dataset when applying Lemmas-based and POS-based algorithm for a vocabulary with 100K words.

Moreover, this approach outperformed the other proposed method (lemmas-based), which yields a relative reduction of 65.6%.

4.5.2 WER Results

Adapting the language model of an ASR system requires identifying and adding new words in the system vocabulary, generating proper phonetic transcriptions for those new vocabulary items, and learning the corresponding linguistic constraints to be represented by the language model itself. To evaluate both vocabulary adaptation algorithms in terms of WER, the following language model adaptation procedure was used. Using the 100K words vocabularies, previously selected by each one of the adaptation algorithms, new language models were estimated. They combine a backoff 4-gram language model trained on WEBNEWS-PT.train dataset, a backoff 3-gram language model estimated on ALERT-SR.train+pilot dataset and a backoff 3-gram language model estimated on the adaptation dataset formed by the written news $O(d)$ collected day-by-day (WEBNEWS-PT.11march). These three models were then combined by means of linear interpolation, generating a mixed model. Mixture weights were estimated on ALERT-SR.devel BN dataset. For each vocabulary, the new words were phonetically transcribed using a rule-based phonetizer [Caseiro et al., 2002]. Those phonetic transcriptions were then manually revised by some linguistic specialists, especially in case of foreign names. Thus, the updated language model and vocabulary replace the baseline ones in the automatic transcription system.

Table 4.11 shows the WER results over the seven BN shows of ALERT-SR.11march dataset, using three different vocabularies: baseline (57K words), Lemmas-based and POS-based (the two last ones with 100K words). These adaptation frameworks produced a significant improvement in terms of word error rate, which was reduced on average by 3.2% for Lemmas-based and 4.3% for POS-based approach. Moreover, we would be also expecting that the POS-based algorithm outperformed the Lemmas-based one in terms of recognition results. In fact, we got an average relative reduction in the WER of 1.1% (see table 4.11).

vocabulary	8th	9th	10th	11th	12th	13th	14th	Week
Baseline	29.4	17.7	28.5	28.7	24.2	33.9	27.3	28.1
Lemmas-based	28.3	17.5	28.2	28.2	22.9	32.3	26.4	27.2
POS-based	28.0	17.5	28.0	27.3	22.9	32.1	26.4	26.9

Table 4.11: WER results over the seven BN shows of ALERT-SR.11march dataset using three different vocabularies: Baseline (57K words), Lemmas-based (100K words) and POS-based (100K words).

Analyzing the ASR output produced by each one of the adapted language models, we could conclude that this marginal difference in terms of WER was due to the correct recognition of additional new words in case of POS-based vocabulary when compared to the Lemmas-based one. As shown in table 4.12, for POS-based vocabulary, the average absolute WER reduction was 1.2%, with a ratio between absolute improvement in WER and the OOV rate of 1.4. This conforms to the empirical evidence stated by Hetherington that on average 1.5-2 errors are obtained per OOV [Hetherington, 1995]. However, for Lemmas-based vocabulary, this ratio was only 1.1. In fact, analyzing the relative percentage of new words added to each one of the vocabularies and correctly recognized, we observed a percentage of 55.1% for the POS-based vocabulary and a slightly small percentage of 54.5% for the Lemmas-based one. This difference is mainly due to the fact that in the Lemmas-based approach there are more verbal forms wrongly recognized due to the way the vocabulary is selected (all verbal forms with the same lemma), increasing the acoustic confusability between some verbal forms sharing the same lemma.

Vocabulary	%OOV	WER	ΔWER/ΔOOV
Baseline	1.25	28.1	-
Lemmas-based	0.43	27.2	1.1
POS-based	0.36	26.9	1.4

Table 4.12: Ratio of the absolute error reduction in WER and OOV rate for the ALERT-SR.11march dataset using the Lemmas-based (100K words) and POS-based (100K words) vocabularies.

4.6 Summary

In this section we proposed and described two methods for daily adapting the vocabulary of a speech recognizer to the broadcast news content. While the first approach extends the vocabulary with additional words extracted from written news daily collected from the Internet, the second approach dynamically generates a new vocabulary from scratch on a daily basis, selecting words from various training/adaptation corpora to maximize its lexical coverage. This second approach showed to be more robust both in terms of OOV word rate and WER, allowing to directly defining the desired size of the target vocabulary. Moreover, this adaptation procedure is simple, extensible to any number of available training corpora and experimental results showed that when compared with the common word frequency based approach it gives better results, especially for selection of large-sized vocabularies.

Hence, for the proposed language model adaptation framework we present in next chapter, we used the POS-based algorithm for vocabulary selection.

5

Language Model Adaptation

An up-to-date language model is recognized to be a critical aspect of maintaining the level of performance for a speech recognizer over time for most applications. In particular, for applications such as transcription of broadcast news and conversations where the occurrence of new words is very frequent.

As described in section 2.1.6, language model adaptation can take several forms. One approach might be to perform an offline adaptation, in which the language models are adapted in advance to their use. This adaptation could either be done in a supervised way, or it could also be performed in an unsupervised approach where the language model is adapted in some form based on the sentences that have been recognized already.

Language modeling typically requires large quantities of in-domain training data, i.e., data that matches the task in both topic and style. For broadcast news and conversational speech applications, this is often unrealistic since topics change frequently, and collecting training data is time-consuming and expensive. Thus, the ability to adapt an existing language model over time, also referred to as dynamic LM adaptation, is desirable. As summarized in section 2.1.6, there have been various works using data from the Internet as an additional source of training data for unsupervised language modeling.

This chapter addresses the problem of dynamically adapting over time the language model of our European Portuguese BN transcription system, using adaptation texts extracted from the Internet and the previously described POS-based vocabulary selection algorithm. The next sections present the unsupervised language model adaptation framework proposed in this thesis, and its evaluation. Finally, we briefly describe the

integration and implementation of the proposed approach into a fully functional prototype system for the selective dissemination of multimedia information.

5.1 Multi-phase Adaptation Framework

In [Martins et al., 2007a] we proposed a daily and unsupervised adaptation approach which dynamically adapts the active vocabulary and language model to the topic of the current news segment using a multi-phase speech recognition process. Based on contemporary texts daily available on the Web, a story-based vocabulary is selected using the morpho-syntactic technique described in section 4.4. Using an Information Retrieval engine and the ASR hypotheses as query material, relevant documents are extracted from a dynamic and large-size dataset to generate a story-based language model.

In the next sub-sections we will describe each one of the speech recognition phases.

5.1.1 First-phase (online)

As stated in section 3.2, the baseline AUDIMUS.media ASR system is part of a closed-captioning system of live TV broadcasts, being the state-of-the-art in terms of broadcast news transcription systems for European Portuguese. However, the language modeling component of this baseline system uses a static vocabulary and language model (see figure 5.1), not being able to cope with vocabulary and linguistic content changes over time. To overcome this limitation we proposed and implemented an adaptation approach, which creates from scratch both vocabulary and language model components on a daily basis. Hence, for each day d this approach is performed according to the following steps:

1. Vocabulary Selection

   A new vocabulary V_0 is selected for each day d applying the POS-based algorithm described in section 4.4 and using three corpora as training data: the newspaper texts from WEBNEWS-PT.train dataset (out-of-domain data), the broadcast news transcriptions from ALERT-SR.train+pilot dataset (in-domain data) and the contemporary texts daily extracted from the Web (as adaptation data).

However, only an average of 80K words is being collected per day as adaptation data. Thus, to construct a more homogeneous adaptation dataset and collect enough n-grams containing new words, we merge Web data from several consecutive days. In our work we considered a heuristic time span of seven days. Similar approaches were taken in [Federico and Bertoldi, 2004] and [Allauzen and Gauvain, 2005a]. Hence, for each day d, we use the texts from the current day and the six preceding days (we will denote this adaptation subset as $O_7(d)$ - 7 days of online written news). For the POS-based algorithm, we use the ALERT-SR.train+pilot as the in-domain corpus to estimate the POS distribution function.

2. Language Model Training

   Using the selected vocabulary V_0, three language models are estimated: a generic backoff 4-gram language model (NP-LM) trained on WEBNEWS-PT.train; an in-domain backoff 3-gram language model (BN-LM) trained on ALERT-SR.train+pilot; and an adaptation backoff 3-gram language model (OL-LM) trained on $O_7(d)$. The generic language model (NP-LM) was estimated using the modified Kneser-Ney smoothing, with the absolute discounting being used to estimate the other two language models, BN-LM and OL-LM.

3. Unsupervised Language Model Adaptation

   We use $P_{NP}(w|h)$ to denote the conditional probability of word w based on history h estimated for the NP-LM, $P_{BN}(w|h)$ for the BN-LM, $P_{OL}(w|h)$ for the OL-LM, and $P_{MIX_0}(w|h)$ for the conditional probabilities according to the adapted language model (MIX_0-LM) we want to generate. To perform the unsupervised language model adaptation, we optimize the linear interpolation weights α and β between NP-LM, BN-LM and OL-LM based on the maximum likelihood criterion. The adapted mixture model probabilities $P_{MIX_0}(w|h)$ are as follows:

$$P_{MIX_0}(w|h) = \alpha P_{NP}(w|h) + \beta P_{BN}(w|h) + (1-\alpha-\beta) P_{OL}(w|h) \quad (5.1)$$

The mixture coefficients α and β are estimated using the Expectation-Maximization (EM) algorithm to maximize the likelihood of a held-out dataset. For that propose, we defined as our held-out dataset the set of ASR transcriptions generated by the broadcast news transcription system itself for the 21 preceding days (noted here as $T_{21}(d)$), i.e., 3 weeks of automatically generated captions. However, the confidence measure described in section 3.2.5 is used to select only the most accurately recognized transcription segments. Thus, all the words w_i with a confidence value $P(correct|w_i)$ higher than 91.5% are included in the $T_{21}(d)$ dataset. This is an important issue, since recognition errors can skew the n-gram estimates and thus deteriorate the adapted language model. In fact, in [Tam and Schultz, 2006] and [Wang and Stolcke, 2007] a degradation on the recognition performance was reported when the baseline language model was adapted based on automatic transcriptions, with the authors postulating that this may be caused by the recognition errors that were not smoothed properly. At first, in our experiments, we started by using only 7 days, which showed to produce a small held-out dataset due to the rejection rate imposed by the confidence threshold. Hence, to collect a more homogeneous $T_{21}(d)$ dataset, we used a time span of 21 days. Finally, the mixed language model (MIX_0-LM) is pruned to a reasonable size using entropy-based pruning [Stolcke, 1998].

4. Phonetic Transcriptions Generation

   Finally, the phonetic transcriptions for new words appearing in V_0 vocabulary are automatically derived with a rule-based phonetizer [Caseiro et al., 2002], augmented with a set of exceptions manually revised by some linguistic specialists.

This adaptation framework (figure 5.2) generates from scratch both the vocabulary (V_0) and language model (MIX_0-LM), which are then used by the first run of the ASR to produce the live captions for the TV broadcast on a daily basis.

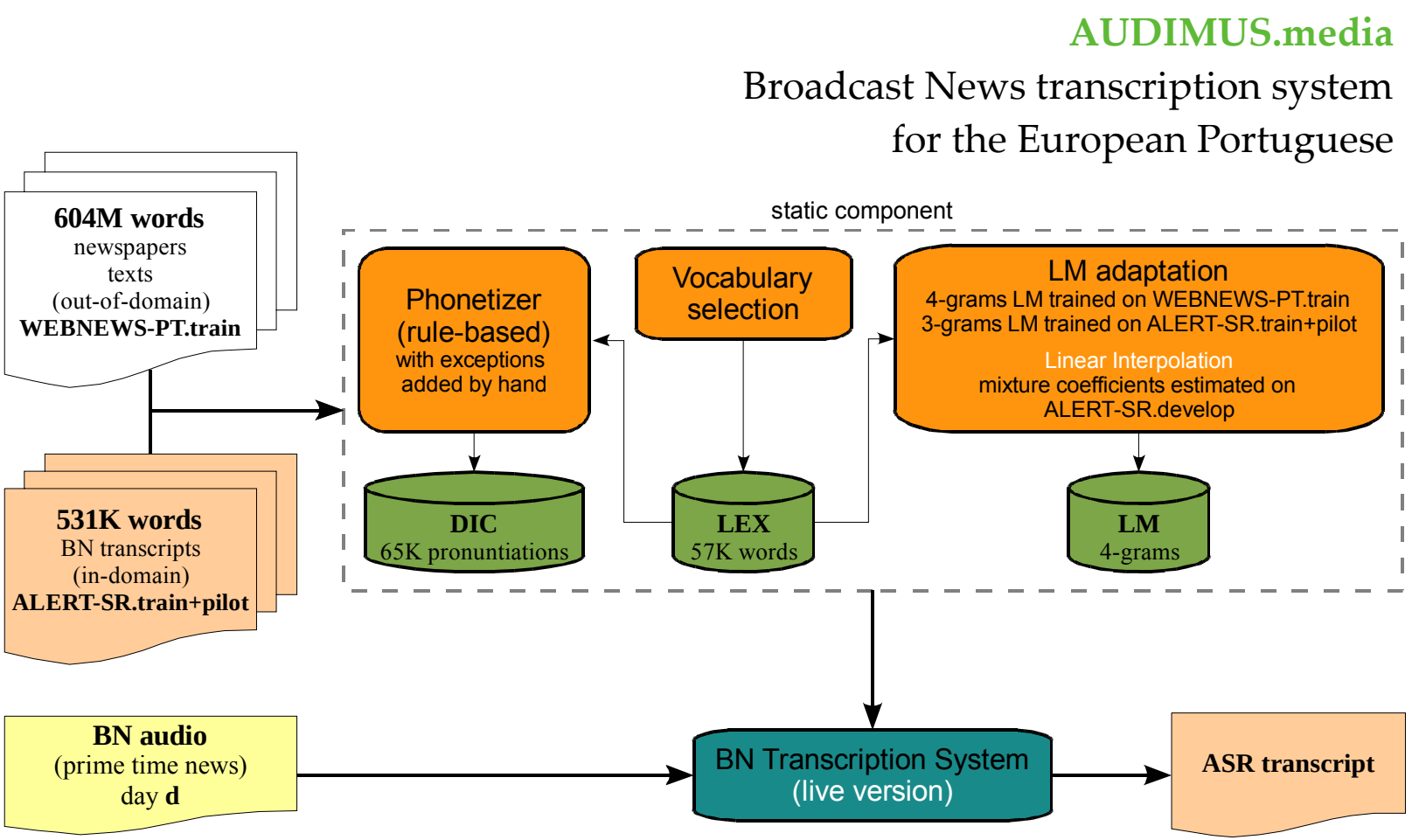

Figure 5.1: Static LM component of the baseline BN transcription system running on a daily basis to produce live captions for European Portuguese TV broadcasts.

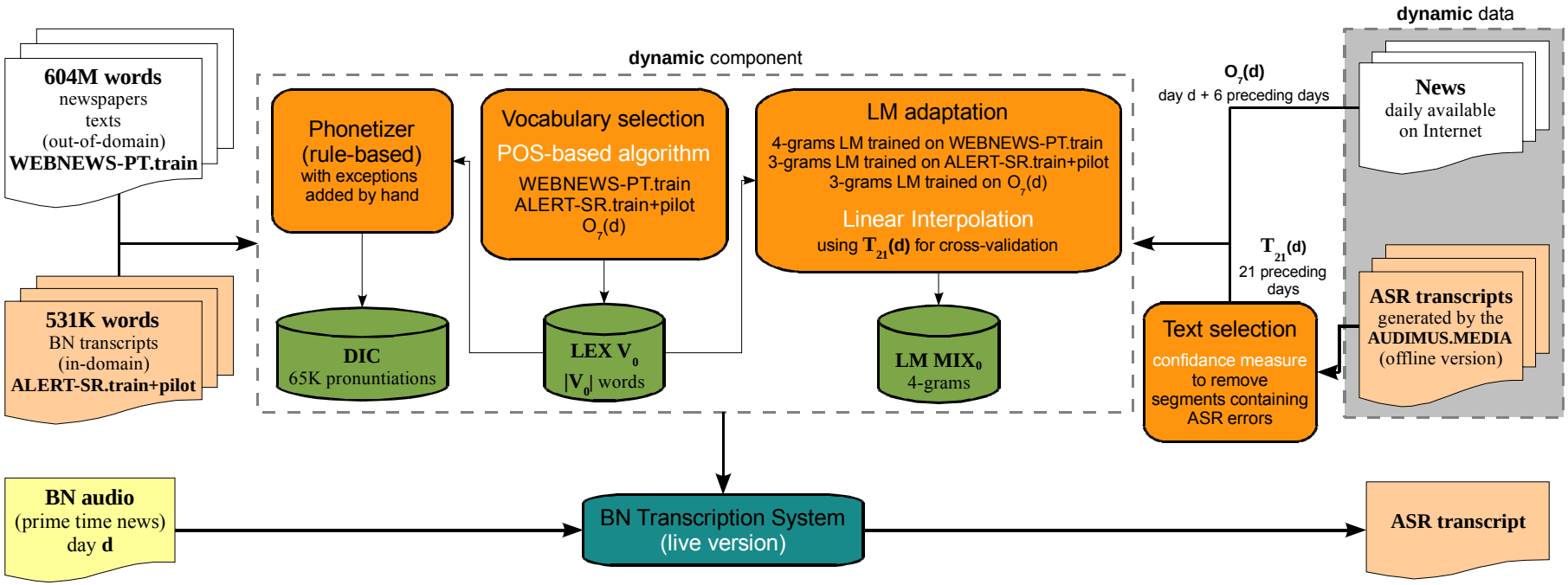

Figure 5.2: Multi-phase adaptation framework: first-pass (online).

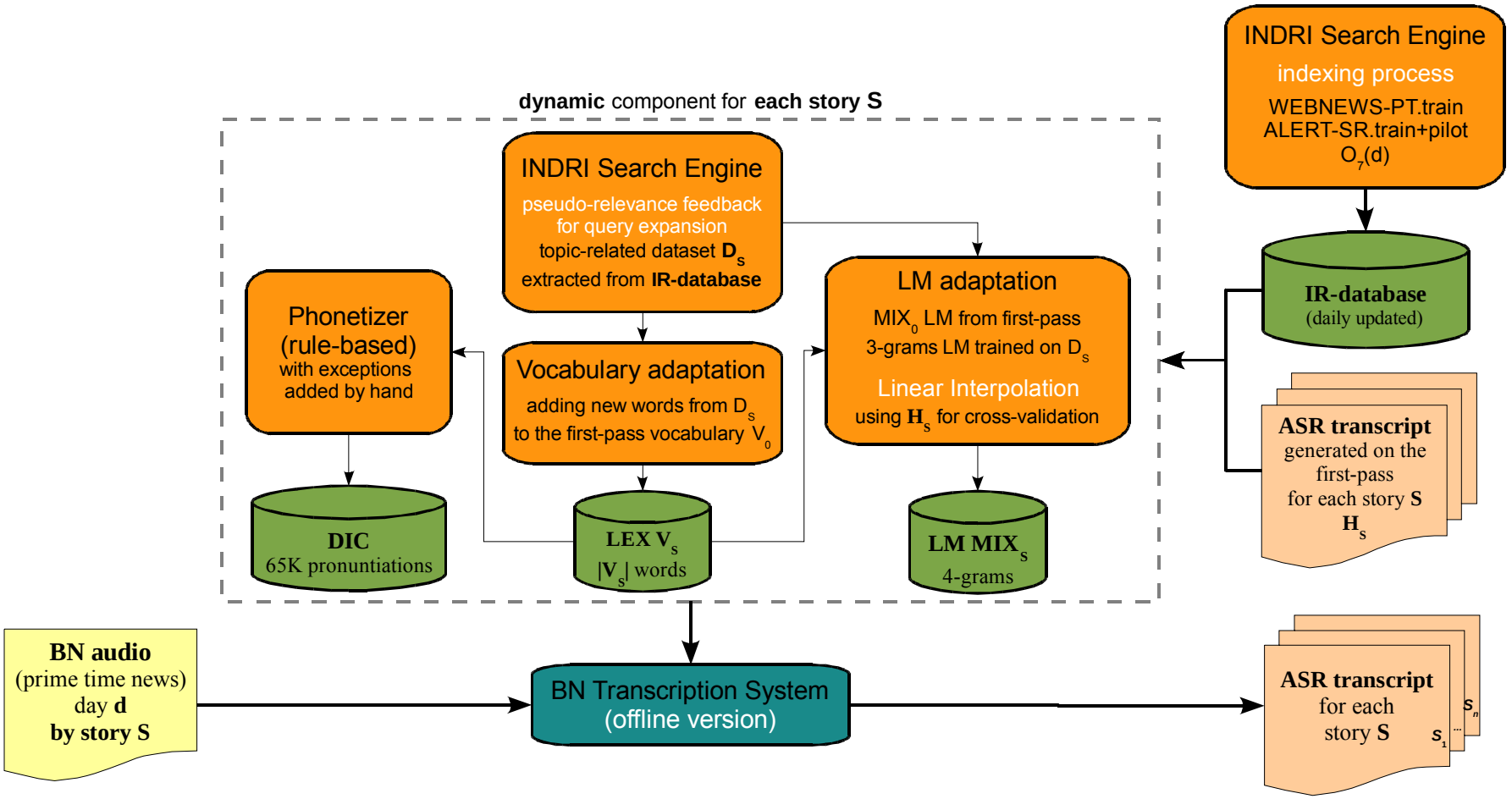

Figure 5.3: Multi-phase adaptation framework: second-pass (offline).

5.1.2 Second-phase (offline)

In this multi-pass adaptation framework, a second-pass is being used to produce improved transcriptions for each day using the initial set of ASR hypotheses generated during the live version.

The initial set of ASR hypotheses (the result of the first decoding pass), which include texts on multiple topics, is automatically segmented into individual stories with each story ideally concerning a single topic. These segmentation boundaries are located by the audio partitioner [Meinedo and Neto, 2005] and topic segmentation procedure [Amaral et al., 2006] currently implemented on the baseline system. The text of each segment can then be used as a query for an Information Retrieval engine to extract relevant documents from a dynamic and large-size database (table 5.1). This way, a story-based dataset is extracted for each segment and used to dynamically build an adapted vocabulary and language model for each story present in the news show being recognized.

In our framework we used the Information Retrieval engine described in section 3.4.3, INDRI - a language-based search engine. For the indexing process we defined as term the concept of word. During the indexing/retrieval process we removed all the function words and the 500 most frequent words, creating a *stoplist* of 800 words. As the starting point, the indexing of all training datasets (WEBNEWS-PT.train and ALERT-SR.train+pilot) has been done, generating a total of about 1.5M articles indexed (called story-segments in case of broadcast news shows). In table 5.1 we present some statistics related to the produced IR-database. After this initial indexation process, the IR database is being updated on a daily basis with the contemporary texts collected for the WEBNEWS-PT corpus, i.e. the texts used to generate the $O_7(d)$ dataset. Thus, for each day d, the texts collected from the Web are provided in an article basis, being dynamically indexed and stored by the Information Retrieval engine (an average of 50K tokens per day).

	#segments	#types	#tokens	size (GB)
IR-database	1,555,807	1,140,429	617,694,657	2.7

Table 5.1: Text statistics for the IR-database.

For the experiments presented in this thesis we used the standard similarity measure for the retrieval phase – the *cosine* measure. During the retrieval process, all articles with an IR score exceeding an empirically determined threshold are extracted for each news story. After some experiments, and to collect enough data for adaptation, we extract the first 1,000 articles with the highest IR score for each one of the queries. However, since the number of words in the hypothesized transcript of each story is usually small and contains transcription errors, one uses a pseudo-relevance feedback mechanism for automatic query expansion [Lavrenko et al., 2001]. This method uses the ASR hypotheses as an initial query, do some processing, and then return a list of expansion terms. The original query is then augmented with the expansion terms and rerun.

Thus, using H_0 to denote the initial set of ASR hypotheses produced by the live version of the BN transcription system, for each day d this second-pass of the ASR system is performed according to the following steps (see figure 5.3):

1. Story Segmentation

   H_0 and the corresponding broadcast news audio are automatically segmented into stories using the topic detection procedure, generating an individual transcript file H_S for each story S.

2. Information Retrieval

   Using the INDRI search engine, and the pseudo-relevance feedback mechanism described above, a topic-related dataset (D_S) is extracted for each story S from the IR dynamic database.

3. Vocabulary Adaptation

   For each story S, all new words found in the corresponding D_S dataset are added to V_0, generating this way a story-specific vocabulary V_S. Note that, for each word added, the vocabulary size is kept constant by removing from V_0 the word with the lowest frequency.

4. Unsupervised Language Model Adaptation

   For each story S, and using its specific vocabulary V_S, an adaptation backoff 3-gram language model trained on D_S is estimated using the modified Kneser-Ney smoothing. As in the first-pass, this new language model is linearly combined with the MIX_0-LM language model to generate a story-specific language model (MIX_S-LM). The H_S set is used to estimate the mixture weights.

5. Phonetic Transcriptions Generation

   As before, the phonetic transcriptions for new words appearing in V_S vocabulary are automatically derived.

Using V_S and MIX_S-LM in a second decoding pass, the final set of ASR hypotheses is generated for each story S. By applying this multi-phase adaptation approach we would be expecting to improve the system performance over the first-pass. In the next section we will describe the experiments we have done for its evaluation.

5.2 Evaluation Results

As before, to evaluate and compare the performance of the proposed adaptation framework we used two evaluation metrics: OOV word rate and WER over the two BN shows of the ALERT-SR.RTP-07 dataset. To fairly compare its performance, with the one obtained for the baseline system, we used a vocabulary size of 57K words.

In addition, comparison of results for both live and offline approaches using different vocabulary sizes is also described. Starting from the baseline vocabulary size of 57K words we defined two new vocabulary sizes to test: a smaller one with about fifty percent of 57K (30K words) and another one with almost the double of the size (100K words). Thus, three different vocabulary sizes are tested: 30K, 57K and 100K words vocabularies.

5.2.1 OOV Rate Results

In figure 5.4 we present the average OOV rate for the two BN shows of the ALERT-SR.RTP-07 dataset when applying the multi-phase adaptation framework for a vocabulary size of 57K words. As one can observe, the proposed second-pass speech recognition approach (2-PASS-POS-IR) using the morpho-syntactic algorithm for vocabulary adaptation (POS-based) and the Information Retrieval engine (IR) for language model adaptation, yields a relative reduction of 65% in OOV word rate, i.e. from 1.40% to 0.49%, when compared to the results obtained for the baseline system. Moreover, this approach outperformed the one based on one single-pass (1-PASS-POS).

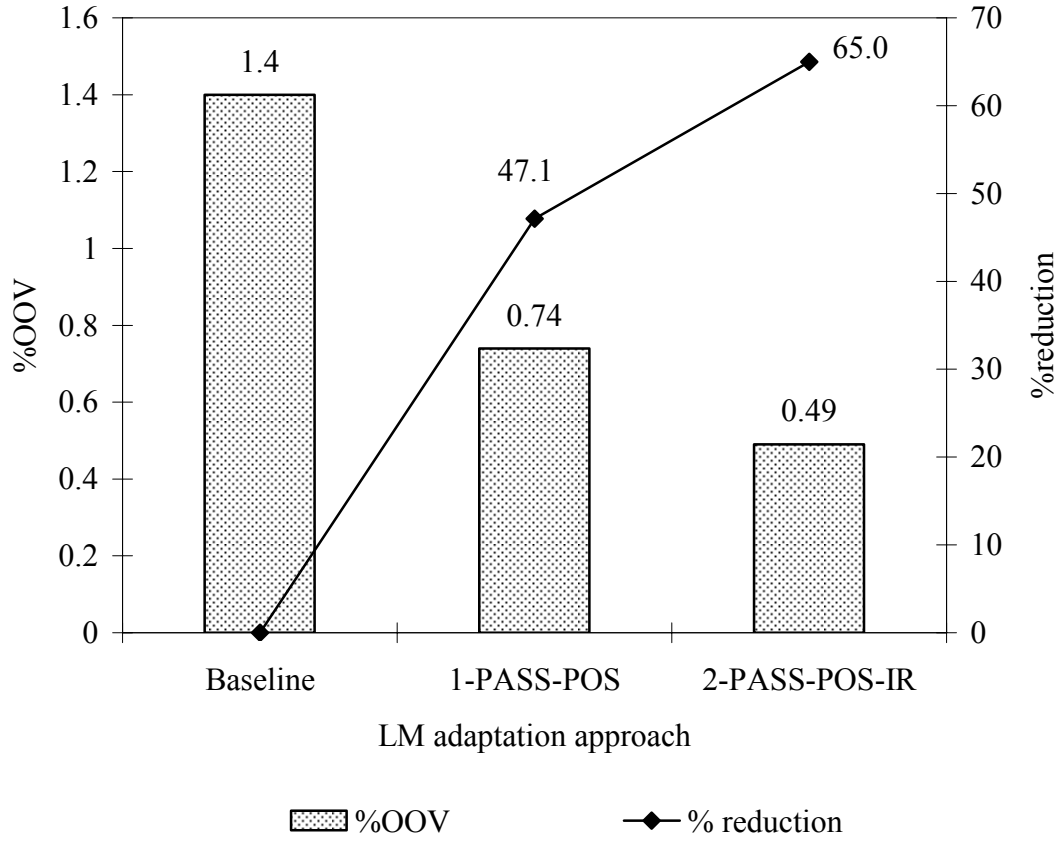

Figure 5.4: OOV word rate for the two BN shows of the ALERT-SR.RTP-07 dataset when applying the multi-phase adaptation framework (vocabulary size of 57K words).

Analyzing the set of OOV words produced by the 2-PASS-POS-IR according to their classification into part-of-speech (POS) classes, we could observe a similar distribution to

the one obtained in section 4.2. In fact, as one can observe from figure 5.5, the verbs class remains the predominant one, with 58.2% of the OOV words belonging to that POS class.

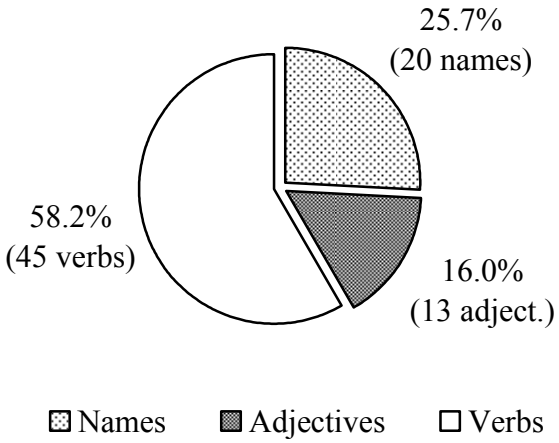

Figure 5.5: Distribution (in %) of OOV words by POS-classes in the ALERT-SR.RTP-07 dataset, after applying the second-pass adaptation approach.

To better understand the performance of this new adaptation procedure we compared the OOV rate results for three different vocabulary sizes (30K, 57K and 100K words).

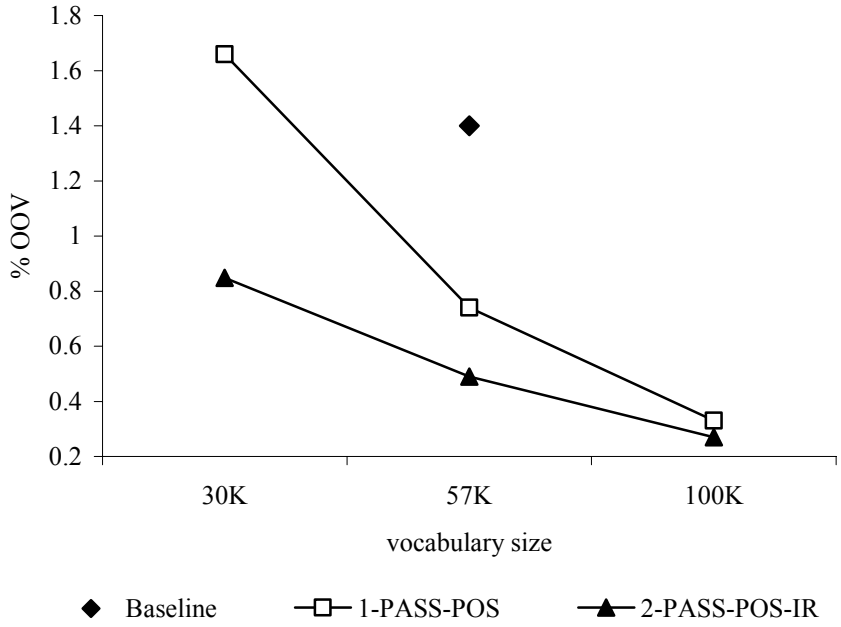

Figure 5.6: OOV rate for the two BN shows of the ALERT-SR.RTP-07 dataset when applying the multi-phase framework with 3 different vocabulary sizes (30K, 57K and 100K).

The graph in figure 5.6 shows the relative good performance of 1-PASS-POS and 2-PASS-POS-IR approaches for the selection of large-sized vocabularies. Furthermore, as we would expect, for the selection of small vocabularies better results are achieved by using the 2-PASS-POS-IR method. In fact, as one can see, with a vocabulary of 30K words we were able to get a better lexical coverage than the one obtained for the baseline system with a vocabulary of 57K words.

5.2.2 WER Results

In table 5.2 we present the average WER for the two BN shows of the ALERT-SR.RTP-07 dataset when applying the multi-phase adaptation framework for a vocabulary size of 57K words. In terms of WER, the new approach (1-PASS-POS-IR) resulted in a 5.7% relative gain, from 21.1% to 19.9%, for a vocabulary size of 57K words. As we would be expecting, the proposed second-pass approach yields a relative reduction of 6.6% in WER when compared to the WER obtained for the baseline system, outperforming the 1-PASS-POS-IR approach (a slight decrease in WER, from 19.9% to 19.7%).

approach	WER	%reduction
Baseline	21.1	-
1-PASS-POS	19.9	5.7
2-PASS-POS-IR	19.7	6.6

Table 5.2: WER for the two BN shows of the ALERT-SR.RTP-07 dataset when applying the multi-phase adaptation framework (vocabulary size of 57K words).

In figure 5.7 we present an WER analysis in terms of word mismatch (substitutions, deletions and insertions). From this analysis, we could conclude that the WER reduction was mainly due to a decrease in the number of word substitutions, i.e. an absolute decrease of 1.1%.

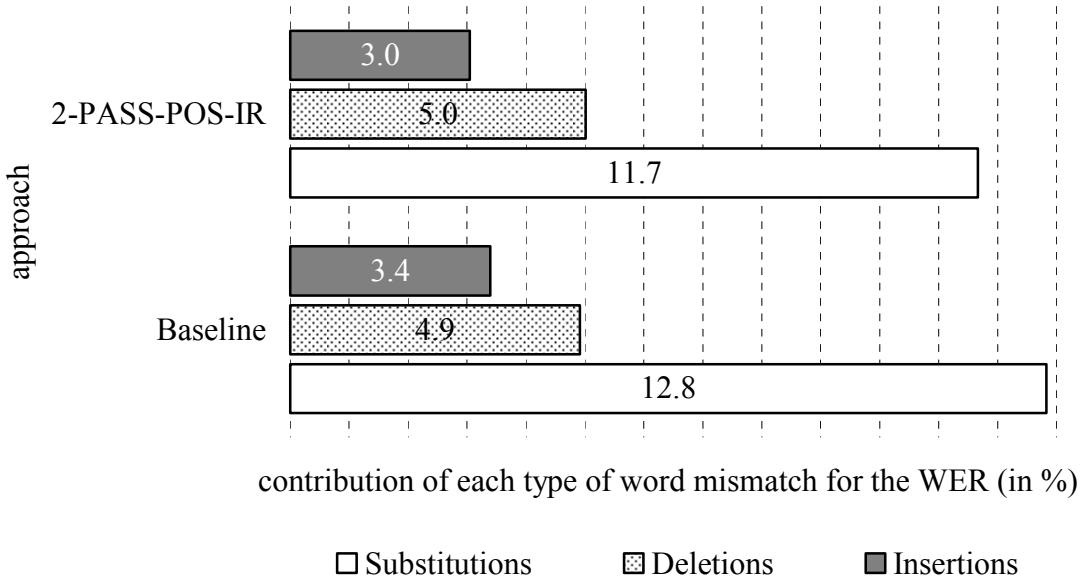

Figure 5.7: Analysis of WER in terms word mismatch (substitutions, deletions and insertions) for the two BN shows of the ALERT-SR.RTP-07 dataset when applying the multi-phase adaptation framework (with a vocabulary size of 57K words).

To better evaluate the accuracy of our approach we performed a more detailed analysis of the WER obtained by the 2-PASS-POS-IR approach with a vocabulary of 57K words. For that analysis, we divided the adapted vocabulary V_S of each story S into 2 sets: the set of word types that were already present in the baseline vocabulary V_0, and the set of all new word types. From this last set (denoted by N_S), we removed all the word types except the ones occurring in the reference transcripts of the tested BN dataset (ALERT-SR.RTP-07). The number of word types in N_S was 86, with 156 occurrences in the reference transcripts. From these 156 occurrences, 108 were correctly recognized by the 2-PASS-POS-IR approach, which means 69.2% of new words found by our IR-based framework were correctly recognized.

In table 5.3 we present the distribution of those 156 occurrences by grammatical category. In the "Names" category we generically include both proper and common names, even the foreign ones. The "Others" category includes other foreign words, acronyms and abbreviations.

POS	% of occurrences	% correctly recognized
Names	**60.3**	**74.5**
Adjectives	10.9	70.6
Verbs	21.2	57.6
Others	7.7	58.3

Table 5.3: Distribution (in %) of new words by grammatical category, and percentage of them correctly recognized by the 2-PASS-POS-IR approach (vocabulary size of 57K words).

As one can observe, more than 60% of those new words found by our algorithm belong to the names class. Moreover, the class of names is the one with the best recognition rate (74.5% of new names were correctly recognized), slightly outperforming the average value (69.2%). This shows that a significant number of relevant terms like proper and common names (including names of persons, locations and organizations) were correctly recognized, making the framework especially useful for novel applications like the information dissemination ones, where those types of words contain a great deal of information.

Finally, we compared the accuracy of the proposed framework for three different vocabulary sizes. Figure 5.8 draws the average WER for the two BN shows of the ALERT-SR.RTP-07 dataset when applying the multi-phase adaptation approach with those 3 different vocabularies (30K, 57K and 100K words). As one can observe that, even using a vocabulary with only 30K words, we were able to get a better WER (20.4%) with our adaptation framework than the one obtained for the baseline system with a 57K words vocabulary (21.1%). Therefore, implementing the proposed multi-pass adaptation approach and increasing the vocabulary size to 100K words we could obtain a relative gain of 8.5% in terms of WER, with a final WER of 19.3% against the 21.1% of the baseline system.

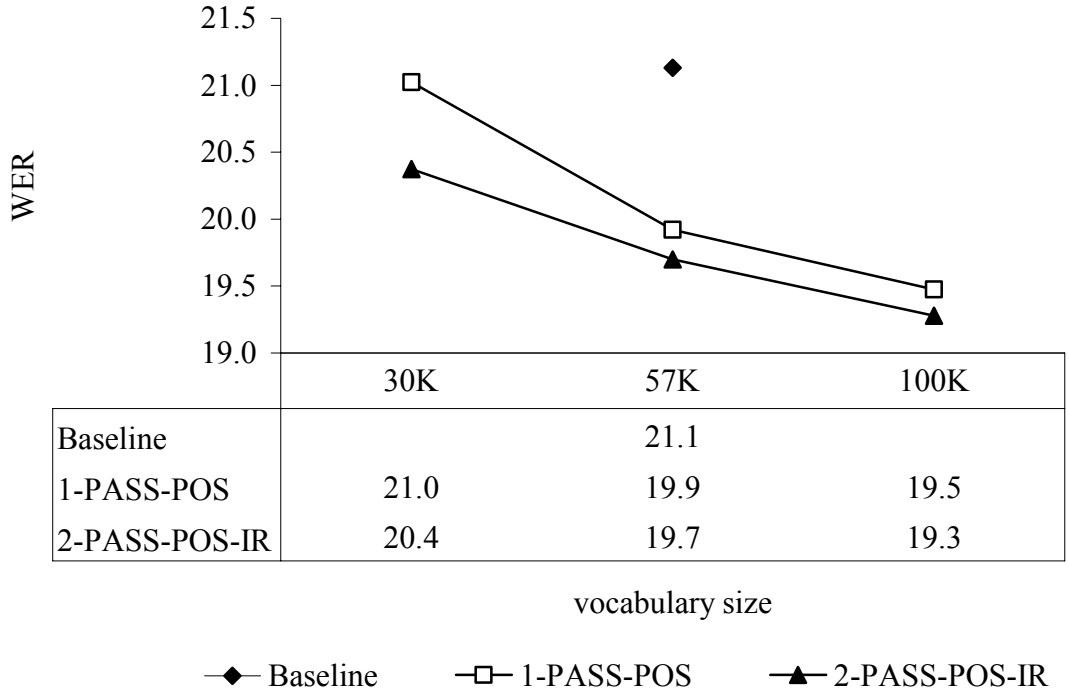

Figure 5.8: WER for the two BN shows of the ALERT-SR.RTP-07 dataset when applying the multi-phase adaptation framework with 3 different vocabulary sizes (30K, 57K and 100K).

5.3 Framework Integration

The work presented in this thesis has been integrated into the fully functional prototype system initially described in section 1.3. Its ASR module were updated and improved by implementing the proposed multi-pass adaptation approach and increasing the vocabulary size to 100K words. Therefore, instead of using a static vocabulary and language model, the updated ASR component takes advantage from that dynamic procedure to better deal with new words appearing in BN data on a daily basis.

For this new framework, the following daily steps have been implemented in the current system:

- Using the RSS News feeds services of 6 different Portuguese news channels, latest news are being collected from the Web, normalized, stored and indexed by the IR engine. This process checks for more news blocks every hour;

- At 5 o'clock in the evening a new vocabulary and LM are generated according the first-pass of the proposed adaptation approach, being used by the ASR module at 8 o'clock to generate closed-captions for the TV news show;
- At the end of the TV news show, the second-pass is processed, generating this way improved BN transcriptions.

The system is fully functional and is processing every day the 8 o'clock evening news show of the Portuguese public broadcast company RTP (http://ssnt.l2f.inesc-id.pt). The live system is generating closed-captions in real-time.

5.4 Summary

In this chapter we presented the work we have done in terms of vocabulary and LM adaptation for European Portuguese ASR. The proposal takes into account the European Portuguese language characteristics such as the high number of verbal inflections. A multi-phase speech recognition framework, using contemporary written texts available on the Web and relevant documents extracted from a general corpus using an IR engine, is proposed. It uses POS class information about an in-domain training corpus to select an optimal vocabulary. When applied to a daily broadcast news transcription task, it showed to be effective, with a relative reduction of OOV word rate (more than 65%) and WER (about 6.6%) when compared to the results obtained for the baseline system with the same vocabulary size (57K words). Moreover, implementing the proposed multi-pass adaptation approach and increasing the vocabulary size to 100K words we could obtain a relative gain of 8.5% in WER.

The baseline transcription system used in this work has been updated according to this multi-phase speech recognition framework. Therefore, instead of using a static vocabulary and language model, the new system takes advantage from this dynamic procedure to better deal with new words appearing in BN shows on a daily basis.

Thus, the first-pass is being used to produce online captions for the closed-captioning system of live TV broadcasts, while the second-pass is being used to improve those captions. On this multi-phase framework, the final set of closed-captions is obtained only

after the end of each live TV show. However, even if this second-pass can not directly benefit the closed-captioning system of live TV broadcasts, it can indirectly improve the overall system performance. First, by reducing the recognition errors we are improving the word accuracy of the held-out dataset ($T_{21}(d)$), which will be used for the unsupervised adaptation performed by the first-pass during the recognition of the next broadcast shows. Moreover, the set of closed-captions obtained by this offline version can be used for other applications where the real-time issue is not important. It is the case of the media monitoring prototype system for the selective dissemination of multimedia information [Meinedo, 2008], which is using the AUDIMUS.media ASR system. For this type of application, our framework showed to be especially useful. In fact, for the ALERT-SR.RTP-07 test dataset, a significant percentage (74.5%) of new words like proper and common names was correctly recognized.

In the next chapter we describe a new method we have proposed to complement the language modeling adaptation framework reported in this chapter that allows including new words in the system vocabulary, even if no well suited training data is available, as is the case of archived documents.

6

Handling Unseen Words

As we described before, different vocabulary and language model adaptation procedures have been proposed to dynamically adapt the language model component of ASR systems, especially for tasks where new words appear on a daily basis as is the case of closed-captioning or information dissemination. Usually, those procedures assume that some kind of well suited sources of data are available to estimate the language model parameters.

However, sometimes we would like to manually add new words to the system vocabulary which are likely to appear on certain broadcast shows, even if no well suited data is available at all, as is the case of archived broadcast news documents [Allauzen, 2003], or just a small amount of data is available but not sufficient to apply the language model adaptation procedures presented in the previous sections. Thus, in this situation estimating the language model parameters for those words is problematic. For example, small amount of data like the anchor working scripts and other prior knowledge information, such as the speakers' names and show summary, can be available and used to extract new words with high probability to appear during the broadcast news show, but not sufficient to extract language model parameters.

As described in section 2.1.1, one of the most commonly used approaches for handling OOV words is the addition of a generic unknown word both in acoustic and language models – the so called filler model [Bazzi, 2002]. However, these filler models can potentially classify segments of the input signal corresponding to in-vocabulary words as OOV words. Moreover, usually an additional step is necessary to transcribe OOV input segments based on phoneme-to-grapheme conversion.

The approach proposed in this thesis is different since we want to explicitly include in the system vocabulary new words, whose orthographic and phonetic transcriptions are known *a priori*, i.e. we just want to know how to incorporate them in the language model, even if no training data is available. The idea is that if no training data is available for the new word, then we will take advantage of morpho-syntactic information related to words which have similar properties in terms of language modeling. In our proposal we use POS word classes to define a new language model unigram distribution associated to the updated vocabulary, assigning probabilities to new words according to their POS classification. Next, we present the proposed method and its evaluation, drawing some conclusions at the end.

6.1 Proposed Method

From an ASR system point of view, adding a new word to its vocabulary implies the following tasks: deriving the possible phonetic transcription(s) associated to that word, and estimating its n-gram distributions within the language model.

Usually, the first task is accomplished by a rule-based phonetizer that automatically derives one or more lexical pronunciations using grapheme-to-phoneme rules. However, estimating the language model parameters for new words is more problematic, especially in cases where no data or insufficient relevant training data is available. As far as no additional training data is available, a new word is no more than an unseen event, which implies estimating n-gram distributions related to unseen words. In a standard approach, various classical smoothing techniques [Chen and Goodman, 1999] exist which can be applied during language model parameters estimation. But, they treat unseen words in the same way, not taking in consideration theirs types or linguistic roles.

By using the BOW classes, the framework proposed in [Allauzen and Gauvain, 2005] takes into account those linguistic differences by clustering words according to their POS tag. However, the estimation of the probability of each new word inside its BOW class relies on some additional adaptation data. The new approach proposed by us allows including new words in the vocabulary even if no well suited training data is available.

Next sub-sections describe the algorithm proposed to define a new LM unigram distribution associated to the updated vocabulary.

6.1.1 Updating Unigram Probabilities

For a standard back-off language model, the n-gram probabilities $P(w_0|h)$ related to unseen words w_0 and given the word history h are derived in the same way, using the unigram estimation $P(w_0)$. However, as no contextual information is available, classical smoothing techniques treat all those w_0 words in a similar basis. Thus, we propose to use classes of words as an alternative to better estimate those unigram probabilities. An additional advantage of classes is that we can gather statistics on the frequency of occurrence of words similar to the unseen ones. The idea is to build a unigram model that uses grammatical information to give a probability to words according to some predefined notion of similarity.

A class-based unigram model is used to implement this idea, where the classes are the parts-of-speech (POS). Therefore, the morpho-syntactic analyzer developed for European Portuguese [Ribeiro et al., 2004] was used to tag all the vocabulary words with their complete morpho-syntactic information. The information coded by this analyzer is described in Appendix A. As an example, for the Portuguese word "fala" (speech) that information consists of five possible tags (see table 6.1), respectively referring to the feminine singular common noun, and four different flexions of the verb "falar" (to speak).

word	morpho-syntactic information
fala	Nc...sf...
	V.ip3s=...
	V.sp1s=...
	V.sp3s=...
	V.m=2s==..

Table 6.1: Complete morpho-syntactic information for the Portuguese word "fala" (speech).

Since keeping all type of morpho-syntactic information for each word would result in too many tags and the training data would be insufficient, we focused only on the syntactic

category of the tag to map words in their corresponding classes. In table 6.2 we list the final tag set used in this work and consisting of 11 grammatical categories. The "Others" category includes foreign words, abbreviations, acronyms and symbols. Hence, the word "fala" in our example, is classified into 2 different classes: class of names (N) and class of verbs (V).

Category	POS	Category	POS
Nouns	N	Prepositions	S
Adjectives	A	Conjunctions	C
Verbs	V	Numerals	M
Pronouns	P	Interjections	I
Articles	T	Others	X
Adverbs	R		

Table 6.2: Part-of-Speech for European Portuguese and their corresponding grammatical categories used in our work.

In the context of a class-based language model, an unseen word can be affected to one or more of these POS classes in order to inherit the contextual properties of the words belonging to these same classes. Thus, in this framework the unigram probabilities $P(w)$ are re-estimated as

$$P(w) = \sum_{c_i \in C(w)} P(w|c_i) P(c_i) \tag{6.1}$$

where $C(w)$ represents the set of POS classes c_i assigned to word w. Therefore, after defining $C(w)$ for all the vocabulary words, the corresponding unigram distribution needs to be re-estimated. The next subsection describes the proposed method for its estimation that allows assigning non-zero probabilities for unseen words.

6.1.2 Parameters Estimation

In (6.1) the emission probability of a word given its class $P(w|c_i)$ and the class probability $P(c_i)$ are both computed through the maximum likelihood estimation (MLE)

approach. For $P(c_i)$ estimation, only the in-domain dataset (ALERT-SR.train+pilot) was used. This decision was based on findings of section 4.4, where we could observe a significant difference in POS distribution when comparing in-domain and out-of-domain datasets, especially in terms of names and verbs. Hence, the in-domain corpus was POS-tagged using the morpho-syntactic ambiguity resolver [Ribeiro, 2003] which gives the POS of a word in its context (table 6.3 presents an example of a tagged sentence). The statistics of occurrence of POS classes in this in-domain corpus were then used to estimate $P(c_i)$, for $i = 1,\ldots,11$.

	Sentence
original text	tenha um bom fim de semana
tagged text	tenha/V um/T bom/A fim/N de/S semana/N

Table 6.3: Example of a sentence tagged by the morpho-syntactic ambiguity resolver.

Due to the small size of the ALERT-SR.train+pilot training dataset, a sub-corpus of WEBNEWS-PT.train and the $O_7(d)$ subset of WEBNEWS-PT corpus were also POS-tagged, and their statistics used for maximum likelihood estimation of the emission probability of a word given its class as

$$P(w|c_i) = \frac{N(w/c_i)}{N(c_i)} \tag{6.2}$$

with $N(w/c_i)$ being the count of occurrences of word w in the context of c_i class. However, the most problematic task is to estimate this probability distribution for new words since we assume that no additional data is available for training. To overcome this problem, we derived a heuristic approach to affect non-zero probabilities for those words by using the morpho-syntactic information of each word. Thus, considering w_0 as an unseen word to be introduced in the system vocabulary, $M(w_0)$ as its complete morpho-

syntactic information, and $S = \{w : M(w) = M(w_0)\}$ as the set of all vocabulary words sharing the same morpho-syntactic information as w_0, we define

$$N(w_0/c_i) = \max_{w \in S} \left(N(w/c_i) \right) \tag{6.3}$$

In our experiments, we studied other functions as *min* and *average* functions, being the *max* function the one that produced the best results. In fact, choosing the *max* function we heuristically assign a probabilistic mass to new words within classes, turning those words as probable as the ones with the highest probabilistic value in the same contextual position. Finally, for the estimation of $P(w|c_i)$ we applied the Kneser-Ney discounting method [Chen and Goodman, 1999] for smoothing purposes.

Next, we will report the experiments we performed to assess the effectiveness of this new method.

6.2 Evaluation Results

To evaluate and compare the performance of our framework, experimental results are reported according to two evaluation metrics: the word error rate (WER) and the percentage of new words introduced in the vocabulary and correctly recognized. The recognition experiments were carried out using the two BN shows of the ALERT-SR.RTP-07 dataset. For these experiments we compared three approaches:

- **Baseline**: in this case, as our baseline system, we used the vocabulary and language model resulting from the 1-PASS-POS adaptation approach described in section 5.1.1, with a vocabulary of 57K words;
- **Standard-addition**: baseline vocabulary extended with the unseen words, and re-estimating the language model as it has been done in the 1-PASS-POS approach. In this case, and to simulate our assumption of no additional data available for training the new words being added to the vocabulary, we removed from all the training corpora all the n-grams containing at least one of those new words.

- **POS-addition**: baseline vocabulary extended with the unseen words, and just re-estimating the unigram probabilities of the baseline language model according to the proposed method.

As explained before, the main goal of the proposed method is to easily and effectively allow the introduction in the system vocabulary of small amount of new words like the ones provided by the anchor working scripts and other prior knowledge information. However, since for the work reported in this thesis we did not have access to this kind of data, we performed an oracle experiment to simulate the use of the proposed method. In this oracle experiment and considering the baseline vocabulary of 57K words, all the OOV words contained in the manual orthographic transcripts of the evaluation dataset (ALERT-SR.RTP-07) were added to the vocabulary. Hence, an average of 48 new words per BN show were added to the baseline vocabulary of 57K words. With this experiment an upper-bound on the gain that could be obtained by applying our approach is estimated.

The WER results are summarized in figure 6.1. As a reference, we present the WER obtained for the baseline system with a vocabulary size of 57K words (Baseline). As said before, the baseline system is exactly the one defined by the 1-PASS-POS approach. However, due to some improvements to the acoustic model [Meinedo, 2008] done after we have run the multi-phase adaptation experiments, the WER obtained for the baseline was different from the previous one (19.0% against the 19.9% reported on table 5.2).

As one can observe, applying the proposed LM updating framework (POS-addition) for the addition of new words to the baseline vocabulary, yields a relative reduction of 6.3% in terms of WER, from 19.0% to 17.8%, with a ratio between absolute improvement in WER and the OOV word rate of about 1.6, which conforms with the assertion: in average 1.5 to 2 errors are obtained per OOV. Moreover, this new approach clearly outperformed the standard one (Standard-addition). Thus, applying the POS-addition approach we could get an absolute improvement of 0.8% over the standard one.

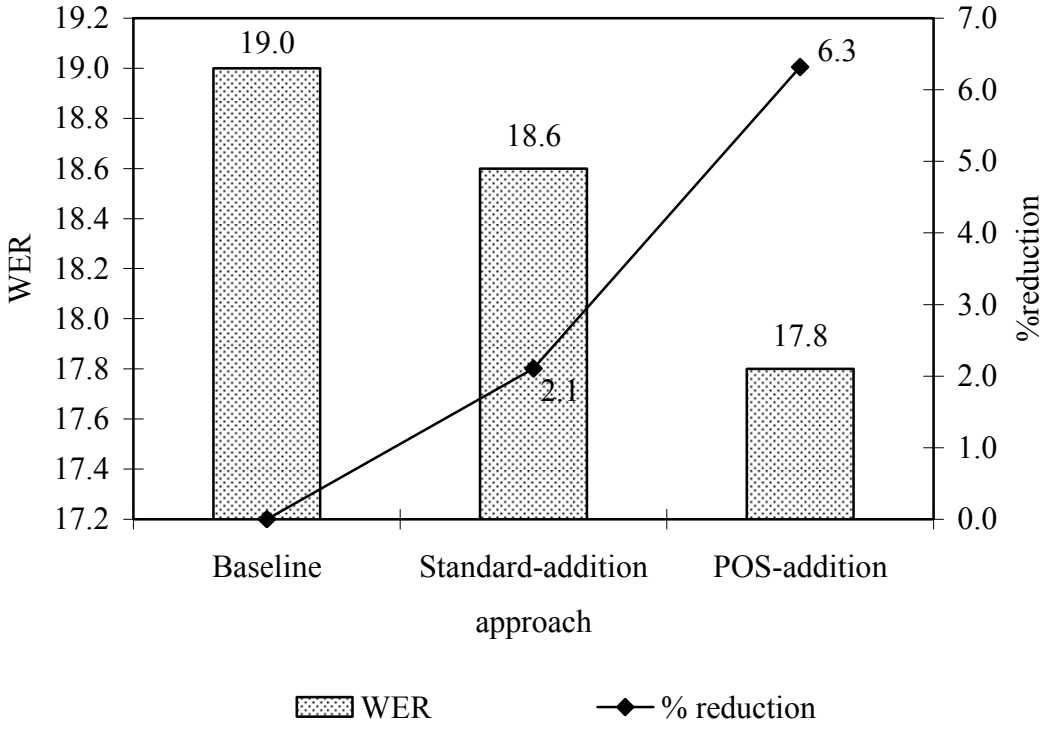

Figure 6.1: WER for the two BN shows of the ALERT-SR.RTP-07 dataset when applying different approaches to estimate LM parameters for unseen words (57K words vocabulary).

The relative percentage of new words introduced in the vocabulary and correctly recognized, is another important metric to measure the performance of the proposed framework. In table 6.4 we present these statistics, by evaluation show, for both language model updating strategies. While only 31.8% of new words were correctly recognized when applying the standard LM approach, a significant improvement has been observed when we used the POS-addition one, with 78.2% of those words being correctly recognized.

Approach	May 24th	May 31st	Average
Standard-addition	25.9	37.5	**31.8**
POS-addition	72.2	83.9	**78.2**

Table 6.4: Percentage of new words correctly recognized with both LM updating strategies: standard-addition and POS-addition.

After analyzing the ASR results (presented in figure 6.1) we could observe that both names and adjectives classes had a recognition rate above the 80%, with only 71.2% of verbs being correctly recognized (table 6.5). In table 6.5 one can also observe the distribution of the new words introduced in the vocabulary by the oracle experiment. This distribution conforms to all the previously reported results, i.e. verbs make up for the largest portion of OOV words (in this case 47.3%).

POS	% of occurrences	% correctly recognized
Names	31.8	80.0
Adjectives	20.9	91.3
Verbs	47.3	71.2

Table 6.5: Distribution (in %) of unseen words by grammatical category, and percentage of them correctly recognized by the POS-addition approach (vocabulary size of 57K words).

Analyzing the words wrongly recognized, we could observe that some new words were wrongly recognized. This occurs mainly in case of foreign words (especially names), abbreviations, and acronyms. However, in case of verbs, globally the phonetic transcription is not the major source of errors. Acoustical conditions, overlapped speech, spontaneous speech style, and the proper statistics of the LM contribute for those errors.

REF: A MIM CORREU-ME BEM só QUE SÓ não fiz foi a primeira a primeira pergunta **HYP**: DE INCUMBEM só SE EU não fiz foi a primeira a primeira pergunta
REF: E NESTA altura FAZ razão nós CHAMARMOS a atenção AQUELA CÉLEBRE frase **HYP**: NESSA altura A FAVA razão nós CHAMAMOS a atenção À COLAÇÃO frase
REF: o tribunal não acreditou no arrependimento e CONDENOU-O a seis anos de prisão **HYP**: o tribunal não acreditou no arrependimento e CONDENOU a seis anos de prisão
REF: se o primeiro-ministro e a ministra da educação querem que os LEVEMOS a sério **HYP**: se o primeiro-ministro e a ministra da educação querem que os GOVERNOS a sério

Table 6.6: Examples of some ASR transcripts containing new words wrongly recognized.

In table 6.6 we present 4 examples of ASR transcripts produced by the POS-addition approach containing new words wrongly recognized. The last two examples represent a typical situation where the language model contributed to the recognition error. In fact, the perplexity values for the reference sentences are 104.6 and 41.5, while the corresponding perplexity values for the ASR hypotheses are 85.1 and 33.8, respectively.

6.3 Summary

The results of our experiments showed us the effectiveness of this new approach for language model re-estimation when new words need to be added to the ASR system in an easy and automatic way, even if no adaptation data is available for that. This is especially useful for inclusion of new words with high probability to appear during the BN show, as is the case of words contained in anchor scripts for example.

The proposed approach assigns non-zero probabilities for those words by using the morpho-syntactic information of each word and POS classes. This way, the LM unigram distribution associated to the updated vocabulary is re-estimated, even if no sufficient data is available. The re-estimation process can be done in a couple of minutes, which allows updating the language model just some minutes before the BN show starts.

However, as shown in table 6.5, our approach does not perform so well for verbs. Hence, 62.5% of those wrongly recognized words were verbs. This last observation suggests us that special focus should be given to this syntactic category of words in our future research trends.

7

Conclusions and Future Directions

In this chapter we briefly discuss the results obtained, providing the main conclusions of the work done in this thesis, and deriving some directions about future research lines.

7.1 Results Discussion

The transcription of broadcast news is a challenging task due to several problems. In terms of language modeling, the topic changes over time and the frequent occurrence of new words appearing every day (out-of-vocabulary words), are two of those problems. The appearance of new words is even more problematic when transcribing BN data in highly inflected languages. To recognize those new words, the vocabulary and the language model of the speech recognition system need to be periodically updated. In this thesis, we addressed the task of incremental language modeling for automatic transcription of European Portuguese broadcast news speech, proposing a framework to dynamically adapt both vocabulary and language model on a daily basis.

In our work, we proposed two vocabulary optimization algorithms. In a first step, we expanded the baseline system vocabulary of 57K words with new words found on texts collected from the Internet, on a daily basis. However, with this procedure we could only reduce the OOV word rate by an average of 28.6%. Moreover, analyzing the distribution of OOV words according to their grammatical property in a given sentence (i.e. their POS), we found that more than 56% of them were verbs. This was an important result since from

other findings published in the literature, OOV words are mostly names. This suggested us to study the interest of using specific linguistic knowledge (POS tagging) to improve the lexical coverage of a selected vocabulary. In fact, all the remaining work of this thesis was motivated and influenced by this finding. Hence, both vocabulary optimization and language model adaptation approaches were based on the integration of different knowledge sources and techniques (language modeling, Information Retrieval and **morphological knowledge**).

Our first approach to compensate and reduce the OOV word rate related with verbs was supported by the fact that almost all the OOV verb tokens were inflections of verbs whose lemmas were already among the lemmas set (L) of the words found in contemporary written news of each day. Thus, the baseline system vocabulary is additionally extended with all the words observed in the language model training texts and whose lemmas belong to L. This lemmas-based approach achieved an OOV word rate reduction of 65.6%. However, it assumes an a priori selected static list of words - the baseline vocabulary, just adding new words on a daily basis. This way, the system vocabulary is always extended, resulting in a vocabulary with an average size of 100K words. Thus we derived a new approach defining a vocabulary from scratch and allowing the selection of its size.

This second approach automatically induces the vocabulary from various training corpora, possibly from different domains, taking the implicit assumption that at least an in-domain corpus is available. It is based on the hypothesis that the linguistic similarities between different domains can be characterized in terms of style (represented by their POS sequences). Hence, instead of simply adding new words to the fixed baseline vocabulary, we use the statistical information related to the distribution of POS word classes on the in-domain corpus to dynamically select words from the various training corpora available. For the ALERT-SR.11march test dataset this POS-based approach yields a relative reduction of 71.2% in OOV word rate, outperforming the lemmas-based one that yields a relative reduction of 65.6% for the same vocabulary size (100K words). Additionally, this last approach is more versatile, allowing the automatic selection of a vocabulary from any number of available training corpora. However, we think there is space to improve the result of our POS-based approach. If more in-domain data was available, we could define a held-out corpus to estimate the mixture coefficients for linear interpolation in equation 4.1

instead of assigning identical weights to all the training corpora. Moreover, we could define the word classes set (*POSset*) with a finer level of granularity as we will discuss in the next section. Unfortunately, as we can observe from section 3.1, we have more than 740 million words of newspapers texts (out-of-domain data), but a maximum of 852K words of broadcast news transcripts (in-domain data), where only about 531K of them are being used as training data by the current ASR system. Comparing these linguistic resources with the ones reported for other languages one can conclude that we have one of the largest corpora in terms of newspaper texts, but a quite small in-domain corpus. State of the art ASR systems for English and French were trained at least with 190 h of manually transcribed training data. Given this we clearly would benefit from more broadcast news training data.

Using adaptation texts extracted from the Internet and the previously described POS-based vocabulary selection algorithm, we proposed and implemented a multi-pass speech recognition approach which creates from scratch both vocabulary and LM components on a daily basis. Therefore, a generic LM is linearly interpolated with a small LM estimated from the adaptation data. This first-pass is being used to produce online captions for a closed-captioning system of live European Portuguese TV broadcasts. In this multi-pass adaptation framework, a second-pass is being used to produce offline transcripts for each day using the initial set of ASR hypotheses generated on the first-pass and automatically segmented into individual stories, with each story ideally concerning a single topic. Using an Information Retrieval engine and the text of each story segment as query material, relevant documents are extracted from a dynamic and large-size database to generate a story-based vocabulary and LM. For the ALERT-SR.RTP-07 test dataset, and considering the same vocabulary size as the baseline one (57K words), a relative gain of 6.6% in the WER and relative reduction of 65.2% in OOV word rate were observed. Moreover, 69.2% of new words found by our IR-based framework were correctly recognized, with slightly better recognition rate (74.5%) for names, which makes this framework especially useful for novel applications like the information dissemination ones, where those type of words contain a great deal of information. In our multi-pass approach we can identify some points which can be improved, namely the estimation of each story-specific language model (MIX_S-LM). In the current approach that LM is generated by means of linear interpolation using the H_S set for cross-validation. However, due to the small size of H_S, we think our

approach could be improved by using other strategies for this LM estimation as we will propose in the next section. Concluding, the proposed multi-pass framework allowed an average relative OOV word rate reduction of 65% with the same vocabulary size as the baseline vocabulary, and 6.6% in terms of WER. These results compare favorably with the ones reported in the related research (section 2.1.6), where a maximum reduction of 58% in OOV word rate and 4.7% in WER were obtained.

In addition to that multi-pass framework, we proposed a language model adaptation scheme that allows us to include unseen words in the vocabulary and providing a heuristic estimation for their probability of appearance, without the need of additional data or LM retraining. As in the vocabulary selection algorithm, it uses morpho-syntactic knowledge about the in-domain corpus and POS tagging to estimate a new unigram distribution associated to the update vocabulary. Our experiments showed that 78.2% of new words were correctly recognized, with both names and adjectives classes reporting a recognition rate above 80%, with only 71.2% of verbs being correctly recognized. This last observation suggests us that a special focus should be given to this syntactic category of words in our future research trends. As in case of vocabulary optimization, the use of more in-domain data and a POS set with a finer level of granularity would likely lead to a higher rate of new words correctly recognized. Comparing our results with the ones reported for the similar approach described in section 2.1.1, one can observe a slightly worst performance for our approach, i.e. 80% against the 78.2% of new words correctly recognized. However, in our approach no additional data is used to estimate the LM probabilities for those new words.

7.2 Main Conclusions

In this work we have investigated the use of additional sources of information to dynamically adapt the language model component of a European Portuguese broadcast news transcription system.

A first conclusion, and the one that influenced the here proposed framework, was the finding that verbs were the dominant source of OOV words for our BN transcription system. Based on this finding a new algorithm for vocabulary selection has been derived,

which showed to be effective in dealing with this specific characteristic of the European Portuguese. This new algorithm has been integrated in a multi-pass ASR framework using an Information Retrieval engine to improve the language model estimation process. Our experiments showed the effectiveness of this adaptation framework by allowing to dynamically incorporate new words in the system vocabulary and to re-estimate its probabilities. As our initial proposal, this framework allows to dynamically and automatically adapt the language model component of our BN transcription system. Moreover, we also showed that a significant number of relevant terms like proper and common names (including names of persons, locations and organizations) were correctly recognized, making the framework especially useful. Lastly, in a pilot study, the proposed algorithm for inclusion of unseen words in the system vocabulary using the morpho-syntactic information of each word and POS classes showed to be effective, allowing an easy update of the system vocabulary even when no additional adaptation data is available.

From the results obtained in our work, we could conclude that the usage of morphological knowledge, as it has been shown, seems to be a promising technique which can be successfully employed both for vocabulary optimization and language model estimation. As suggested in the next section, further research can be taken in order to study its usage for language model probabilities estimation and their smoothing. Additionally, the integration of this morphological knowledge with IR-based techniques would lead to further performance gains. In particular, by exploring the POS class of verbs.

While our focus was on European Portuguese broadcast news, where morphological variants such as inflectional verb endings showed to be an important problem to address, we believe the here proposed framework would likely lead to improved performance for other inflectional languages and/or applications, specially the POS-based vocabulary selection procedure and the suggested algorithm for inclusion of unseen words in the system vocabulary.

In the next section we conclude this thesis by presenting some future research directions.

7.3 Future Work

While we have been able to made progress in tackling the language modeling adaptation problem for European Portuguese by defining a dynamic and unsupervised adaptation framework, there are other extensions that in our opinion can be investigated to enhance the global performance of the language model component. Especially in case of the task addressed in this thesis – broadcast news speech transcription, but extensible to other challenging domains like meetings, course lectures and broadcast conversation (BC).

Next, we summarize some of the research trends we judge to be worth addressing in future works and related to: **vocabulary optimization**, **data selection** and **language model adaptation**.

Using large-sized vocabularies may be desirable from the point of view of lexical coverage. However, there is always the additional problem of increased acoustic confusability [Rosenfeld, 1995]. Moreover, as we reported in section 3.1.1, each BN show comprises an average of 8,300 word tokens and only 2,200 word types. Thus, the majority of the vocabulary words are irrelevant when adapting to a single BN show. Therefore, in our opinion, future research should focuses on methods to better constrain vocabulary growth while preserving adaptation performance. Thus, related to vocabulary selection we highlight two research trends:

- Using a small amount of data like anchor working scripts, BN show subtitles and summaries together with IR techniques to select more accurate adapted vocabularies;
- Improving POS-based vocabulary selection algorithm by defining the word classes set ($POSset$) with a finer level of granularity.

Therefore, we believe that better results can be achieved by exploring more deeply the linguistic knowledge of the in-domain corpus. We can use not only the grammatical property of words (POS), but also its morphological information (gender, number, conjugation, etc.). However, to follow this research trend, two resource constraints must be overcome: more in-domain data and a morphological analyzer for European Portuguese which could give us that level of morphological information with an acceptable accuracy.

One solution to increase the BN training data would be to use the confidence scores for unsupervised selection of text segments from the BN transcripts produced by the ASR system during a pre-defined time span. In fact, the current implemented framework uses this data but only for estimation of the mixture coefficients.

In the speech recognition community, there is a long-standing belief that "there is no data like more data". However, different works have reported significant gains by defining topic-specific subsets and prune less useful documents from training data, both for acoustic and language model adaptation [Hwang et al., 2007] [Ramabhadran et al., 2007] [Wu et al., 2007]. Thus, we can investigate the use of the IR techniques implemented in our multi-pass framework to select and cluster data, reducing its redundancy and improving the generic language model estimation.

Context setting is an important aspect of spoken language communication. In [Zue, 2007] context awareness is enumerated as one of the research challenges to address in order to turn today's speech-based interfaces more organic. For BN transcription, for example, the use of audio segmentation to mark changes in environment, topic and talker, are all attempts to establish context in order to improve speech recognition performance. The entire LM adaptation framework proposed in this thesis is in fact an attempt to take advantage of context.

One strategy for improving the ASR performance of the BN transcription system used in our work was to build speaker adapted acoustic models for certain important speakers like news anchors [Meinedo, 2008]. Therefore, in a similar way, the system can make use of the audio pre-processing information to adapt not only the acoustic model, but also the vocabulary and language model. Since news shows consist of "talking head" style broadcasts (generally one person reading a news script), and other blocks more interactive and spontaneous in style, one can investigate the use of different language models for each one these two genres – broadcast news (BN) and broadcast conversation (BC) [Mrva and Woodland, 2006] [Hwang et al., 2007a] [Wang, 2007].

Approaches taking advantage of semantic and syntactic knowledge are expected to lead to more effective solutions for language model adaptation [Bellegard, 2004]. In the adaptation framework proposed in this thesis, morpho-syntactic knowledge showed to

produce good results both for vocabulary and language model adaptation. To extend the effectiveness of the new method proposed for automatic estimation of LM parameters for unseen words, we can investigate its application as a smoothing method for n-grams of higher order. Using the minimum discrimination information (MDI) based LM adaptation approach, the POS-based unigram marginals can be used to restrict the allowed adaptive n-grams. Moreover, we can try to redefine the word classes by using more details from the morpho-syntactic information available for each word, and giving special attention to verbs. Thus, related to language model adaptation we highlight three research trends:

- Using "context-dependent" language models: BN versus BC language model adaptation;
- Define word classes with a finer level of granularity;
- Unsupervised language model adaptation using POS-based Marginals.

A

Morpho-Syntactic Tagset

The tagset used by the Morpho-syntactic tagger used in our work has about 200 tags with information that varies from grammatical category to morphological features that can be combined to form composed tags (resulting in about 400 different tags) [Ribeiro et al., 2004]. The information coded by this tagset is presented in Table A.1 (in Portuguese).

Each tag is an array, and each position of the array codes one of the features presented in Table A.1, saving the first position for the grammatical category and the second position for the subcategory. When a position (category, subcategory or feature) is not used, its code is replaced by an equal sign. For example, R=n means adverb with no subcategory, in normal degree.

Atributo	Valor	Símbolo	Posição
Categoria gramatical	Nome	N	1
Subcategoria gramatical	comum	c	2
	próprio	p	2
Género	masculino	m	3
	feminino	f	3
	comum	n	3
Número	singular	s	4
	plural	p	4
	invariante	n	4

Categoria gramatical	Verbo	V	1
Modo	indicativo	i	3
	conjuntivo	s	3
	imperativo	m	3
	condicional	c	3
	infinitivo	n	3
	infinitivo pessoal	f	3
	particípio	p	3
	gerúndio	g	3
Tempo	presente	p	4
	pretérito imperfeito	i	4
	futuro	f	4
	pretérito perfeito	s	4
	pretérito mais-que-perfeito	q	4
Pessoa	primeira	1	5
	segunda	2	5
	terceira	3	5
Número	singular	s	6
	plural	p	6
Género	masculino	m	7
	feminino	f	7
	comum	c	7

Categoria gramatical	Adjectivo	A	1
Grau	positivo	p	3
	comparativo	c	3
	superlativo	s	3
Género	masculino	m	3
	feminino	f	3
	comum	n	3
Número	singular	s	4
	plural	p	4
	invariante	n	4

Categoria gramatical	Pronome	P	1
Subcategoria gramatical	pessoal	p	2
	demonstrativo	d	2
	indefinido	i	2
	possessivo	o	2
	interrogativo	t	2
	relativo	r	2
	exclamativo	e	2
	reflexivo	f	2
Pessoa	primeira	1	3
	segunda	2	3
	terceira	3	3
Género	masculino	m	4
	feminino	f	4
	comum	c	4
Número	singular	s	5
	plural	p	5
	invariante	n	5
Caso	nominativo	n	6
	acusativo	a	6
	dativo	d	6
Formação	simples	s	8
	fusão	f	8

Categoria gramatical	Artigo	A	1
Subcategoria gramatical	definido	d	2
	indefinido	i	2
Género	masculino	m	3
	feminino	f	3
Número	singular	s	4
	plural	p	4
Categoria gramatical	Advérbio	R	1
Grau	positivo	p	3
	comparativo	c	3
	superlativo	s	3
Categoria gramatical	Preposição	S	1
Formação	simples	s	3
	fusão	f	3
Género	masculino	m	4
	feminino	f	4
	comum	c	4
Número	singular	s	5
	plural	p	5
	invariante	n	5
Categoria gramatical	Conjunção	C	1
Subcategoria gramatical	coordenativa	c	2
	subordinativa	s	2
Categoria gramatical	Numeral	M	1
Subcategoria gramatical	cardinal	c	2
	ordinal	o	2
Género	masculino	m	3
	feminino	f	3
	comum	c	3
Número	singular	s	4
	plural	p	4
Categoria gramatical	Interjeição	I	1
Categoria gramatical	Marcador da voz médio-passiva	U	1
Categoria gramatical	Residual	R	1
Subcategoria gramatical	estrangeirismo	f	2
	abreviatura	a	2
	acrónimo	y	2
	símbolo	s	2
Categoria gramatical	Pontuação	O	1

Table A.1: Tagset: morpho-syntactic information.

BIBLIOGRAPHY

[Allauzen, 2003] Allauzen, A. (2003). Modélisation linguistique pour l'indexation automatique de documents audiovisuals. In PhD Thesis, LIMSI-CNRS, 2003.

[Allauzen and Gauvain, 2005] Allauzen, A., and Gauvain, J. (2005). Open Vocabulary ASR for Audiovisual Document Indexation. In Proceedings of ICASSP, 2005.

[Allauzen and Gauvain, 2005a] Allauzen, A., and Gauvain, J. (2005). Diachronic vocabulary adaptation for broadcast news transcription. In Proceedings of Interspeech 2005, Lisbon, Portugal, 2005.

[Amaral et al., 2006] Amaral, R., Meinedo, H., Caseiro, D., Trancoso, I. and Neto, J. (2006). Automatic vs. Manual Topic Segmentation and Indexation in Broadcast News. In IV Jornadas en Tecnologia del Habla, pages 123--128, November 2006.

[Bach et al., 2007] Bach, N., Noamany, M., Lane, I. and Schultz, T. (2007). Handling OOV Words In Arabic ASR Via Flexible Morphological Constraints. In Proceedings of Interspeech 2007, Antwerp, Belgium, 2007.

[Bahl et al., 1977] Bahl, L., Baker, J., Jelinek, F. and Mercer, R. (1977). Perplexity — A Measure of The Difficulty of Speech Recognition tasks, Program of the 94th Meeting of the Acoustical Society of America, 1977.

[Bahl et al., 1983] Bahl, L., Jelinek, F. and Mercer, R. (1983). A maximum likelihood approach to continuous speech recognition. In IEEE Transactions on Pattern Analysis and Machine Intelligence, volume PAMI-5, March 1983.

[Bazzi, 2002] Bazzi, I. (2002). Modeling Out-of-Vocabulary Words for Robust Speech Recognition. In PhD Thesis, Massachusetts Institute of Technology, 2002.

[Bellegarda, 2000] Exploiting Latent Semantic Information in Statistical Language Modeling. In Proceedings of IEEE 88 (8), 1279-1296, August 2000.

[Bellegarda, 2004] Bellegarda, J. (2004). Statistical language model adaptation: review and perspectives. Speech Communication 42, 2004.

[Bengio et al., 2003] Bengio, Y., Ducharme, R., Vincent, P., and Javin, C. (2003). A Neural Probabilistic Language Model. In Journal of Machine Learning Research 3 (2003) 1137–1155.

[Berger et al., 1996] Berger, A. L., Pietra, S. D., and Pietra, V. J. D. (1996). A maximum entropy approach to natural language processing. Computational Linguistics, 22(1):39–71.

[Beyerlein et al., 2002] Beyerlein, P., Aubert, X., Haeb-Umbach, R., Harris, M., Klakow, D., Wendemuth, A., Molau, S., Ney, H., Pitz, M. and Sixtus, A. (2002). Large vocabulary continuous speech recognition of Broadcast News - The Philips/RWTH approach. In Speech Communication, Volume 37, Issue 1-2 (May 2002).

[Bigi et al., 2004] Bigi, B., Huang, Y. and Mori, R. (2004). Vocabulary and Language Model Adaptation using Information Retrieval. In Proceedings of ICSLP, 2004.

[Bilmes and Kirchhoff, 2003] Bilmes, J. and Kirchhoff, K. (2003). Factored language models and generalized parallel backoff. In Proceedings of HLT/NACCL, 2003, pp. 4-6, 2003.

[Blei and Jordan, 2003] Blei, A. and Jordan, M. (2003). Latent Dirichlet Allocation. In Journal of Machine Learning Research, 2003.

[Bourlard and Morgan, 1994] Bourlard, H. and Morgan, N. (1994). Connectionist Speech Recognition - A Hybrid Approach. Kluwer Academic Publishers, Massachusetts, EUA, 1994.

[Boulianne et al., 2006] Boulianne, G., et al. (2006). Computer-assisted closedcaptioning of live TV broadcast in French", in Proceedings of Interspeech 2006, Pittsburgh, PA, USA, 2006.

[Buckley et al., 1995] Buckley, C., Allan, J., Salton, G. and Singhal, A. (1995). Automatic query expansion using SMART: TREC 3. In Proceedings of the Third Text REtrieval Conference (TREC-3), pages 69–80. NIST Special Publication 500-225, April 1995.

[Bulyko et al., 2003] Bulyko, I., Ostendorf, M. and Stolcke, A. (2003). Getting more mileage from web text sources for conversational speech language modeling using class-dependent mixtures. In Proceedings of HLT-NAACL, Edmonton, Canada, 2003.

[Bulyko et al., 2007] Bulyko, I., Ostendorf, M., Siu, M., Ng, T., Stolcke, A. and Cetin, O. (2007). Web resources for language modeling in conversational speech recognition. ACM Trans. on Speech and Language Processing, 2007.

[Carpenter et al., 2001] Carpenter, Paul., Jin, Chun., Wilson, Daniel., Zhang, Rong., Bohus, Dan. and Rudnicky, Alex. (2001). Is this conversation on track?. In Proceedings of Eurospeech 2001, Aalborg, Denmark, 2001.

[Caseiro and Trancoso, 2001] Caseiro, D. and Trancoso, I. (2001). Transducer composition for "on-the-fly" lexicon and language model integration. In Proceedings of ASRU Workshop 2001, Madonna di Campiglio, Trento, Italy, 2001.

[Caseiro and Trancoso, 2002] Caseiro, D. and Trancoso, I. (2002). Using dynamic wfst composition for recognizing broadcast news. In Proceedings ICSLP 2002, Denver, USA, 2002.

[Caseiro et al., 2002] Caseiro, D., Trancoso, I., Oliveira, L. and Ribeiro, M. (2002). Grapheme-to-Phone Using Finite-State Transducers. In 2002 IEEE Workshop on Speech Synthesis, September 2002.

[Caseiro, 2003] Caseiro, D. (2003). Finite-State Methods in Automatic Speech Recognition. In PhD Thesis, Lisbon, Portugal: Instituto Superior Técnico, Universidade Técnica de Lisboa, 2003.

[Chelba et al., 2008] Chelba, C., Hazen, T. and Saraçlar, M. (2008). Retrieval and Browsing of Spoken Content. In IEEE Signal Processing Magazine, 39, May 2008.

[Chen et al., 2001] Chen, L., Gauvain, J., Lamel, L., Adda, G. and Adda, M. (2001). Using Information Retrieval Methods for Language Model Adaptation. In Proceeedings of EUROSPEECH'01, pp. 255-258, 2001.

[Chen et al., 2004] Chen, L., Gauvain, J., Lamel, L. and Adda, G. (2004). Dynamic Language Modeling for Broadcast News. In Proceedings of ICSLP, 2004.

[Chen and Goodman, 1998] Chen, S. and Goodman, J. (1998). An empirical study of smoothing techniques for language modeling. Technical Report TR-10-98, Computer Science Group, Harvard University, 1998.

[Chen and Goodman, 1999] Chen, S. and Goodman, J. (1999). An empirical study of smoothing techniques for language modeling. Computer Speech and Language, 13(4):359{394, October 1999.

[Choueiter et al., 2006] Choueiter, G., Povey, D., Chen, S. and Zweig, G. (2006). Morpheme Based Language Modeling for Arabic LVCSR. In Proceedings of ICASSP, 2006.

[Clarkson, 1999] Clarkson, P. (1999). Adaptation of Statistical Language Models for Automatic Speech Recognition. In PhD thesis, Cambridge University Engineering Department, 1999.

[Cole et al., 1995] Cole, R., Mariani, J., Uszkoreit, H., Zaenen, A. and Zue, V. (1995). Survey of the State of the Art in Human Language Technology. Novembro, 1995.

[Creutz et al., 2007] Creutz, M., Hirsimäki, T., Kurimo, M., Puurula, A., Pylkkönen, J., Siivola, V., Varjokallio, M., Arisoy, E., Saraclar, M. and Stolcke, A. (2007). Morph-Based Speech Recognition and Modeling of Out-of-Vocabulary Words Across Languages. ACM Transactions on Speech and Language Processing, Vol. 5, Issue 1, ACM, pp. 1-29, New York, December 2007.

[Decadt et al., 2002] Decadt, B., Duchateau, J., Daelemans, W. and Wambacq, P. (2002). Transcription of Out-Of-Vocabulary Words in Large Vocabulary Speech Recognition Based on Phoneme-To-Grapheme Conversion. In Proceedings of ICASSP, vol 1, pp 861-864, Orlando, Florida, USA, 2002.

[Deerwester et al., 1990] Deerwester, S., Dumais, S. T., Landauer, T. K., Furnas, G. W. and Harshman, R. A. (1990). Indexing by Latent Semantic Analysis. of the Society for Information Science, 41(6), 391-407.

[Dempster et al., 1977] Dempster, A., Laird, N. and Rubin, D. (1977). Maximum likelihood from incomplete data via the EM algorithm. In Journal of the Royal Statistical Society, Series B, 39(1):1–38.

[Federico, 1999] Federico, M. (1999). Efficient Language Model Adaptation through MDI Estimation. In Proceedings of Eurospeech'99, 1583-1586, 1999.

[Federico and Bertoldi, 2004] Federico, M. and Bertoldi, N. (2004). Broadcast News LM Adaptation over Time. Computer Speech and Language. 18(4): pp. 417-435. October, 2004.

[Gales et al., 2006] Gales, M., Kim, D., Woodland, P., Chan, H., Mrva, D., Sinha, R. and Tranter, S. (2006). Progress in the CU-HTK broadcast news transcription system. In IEEE Transactions on Speech and Audio Processing, 14 (5). pp. 1513-1525, 2006.

[Gauvain et al., 2001] Gauvain, J., Lamel, L. and Adda, G. (2001). Audio partitioning and transcription for broadcast data indexation. In MTAP Journal, 14(2):187–200, 2001.

[Gauvain et al., 2002] Gauvain, J., Lamel, L. and Adda, G. (2002). The LIMSI Broadcast News Transcription System. Speech Communication, vol. 37, 2002.

[Gauvain et al., 2005] Gauvain, J., Adda, G., Adda-Decker, M. Allauzen, A., Gendner, V., Lamel, L. and Schwenk, H. (2005). Where Are We In Transcribing French Broadcast News?. In Proceedings of Interspeech 2005, Lisbon, Portugal, 2005.

[Geutner et al., 1998] Geutner, P., Finke, M., Sheytt, P., Waibel, A. and Wactlar, H. (1998). Transcribing Multilingual Broadcast News using Hypothesis Driven Lexical Adaptation. In proceedings of ICASSP, 1998.

[Goodman, 2000] Goodman, J. (2000). Putting it all together: Language model combination. In Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, volume 3, pages 1647-1650, Istanbul, June 2000.

[Goodman, 2001] Goodman, J. (2001). A Bit of Progress in Language Modeling – Extended Version. Machine Learning and Applied Statistics Group - Microsoft Research, Redmond, WA MSR-TR-2001-72, 2001.

[Gutkin, 2006] Gutkin, A. (2006). Log-Linear Interpolation of Language Models. In PhD Thesis, University of Cambridge, 2006.

[Hakkani-Tur and Rahim, 2006] Hakkani-Tur, D. and Rahim, M. (2006). Bootstrapping Language Models for Spoken Dialog Systems From The World Wide Web. In Proceedings of ICASSP, 2006.

[Hazen and Bazzi, 2001] Hazen, T. and Bazzi, I. (2001). A Comparison and Combination of Methods for OOV Word Detection and Word Confidence Scoring," Proc. of ICASSP, Salt Lake City, 2001.

[He and Young, 2004] He, Y. and Young, S. (2004). Robustness Issues in Data- Driven Spoken Language Understanding System. In HLT/NAACL04 Workshop on Spoken Language Understanding for Conversational Systems, 2004.

[Hetherington, 1995] Hetherington, I. (1995). A characterization of the problem of new, out-of-vocabulary words in continuous-speech recognition and understanding. In PhD Thesis, Massachusetts Institute of Technology, 1995.

[Hwang et al., 2007] Hwang, M., Peng, G., Wang, W., Faria, A., Heidel, A. and Ostendorf, M. (2007). Building a highly Accurate Mandarin Speech Recognizer. In Proceedings of ASRU 2007, Kyoto, Japan, 2007.

[Hwang et al., 2007a] Hwang, M., Wang, W., Lei, X., Zheng, J., Cetin, O. and Peng, G. (2007). Advances in Mandarin Broadcast Speech Recognition. In Proceedings of InterSpeech 2007, Belgium, 2007.

[Ircing and Psutka, 2002] Ircing, P. and Psutka, J. (2002). Lattice Rescoring in Czech LVCSR System Using Linguistic Knowledge, Specom02, pp. 23-26, St. Petersburg, Russia, 2002.

[Iyer et Ostendorf, 1997] Iyer, R. and Ostendorf, M. (1997). Transforming out-of-domain estimates to improve in-domain language models. In Proceedings of Eurospeech, 1997.

[Iyer et al., 1999] Iyer, R. and Ostendorf, M. (1999). Modeling long distance dependence in language: Topic mixtures versus dynamic cache models. In IEEE Transactions on Acoustics, Speech and Audio Processing, 7:30–39, January.

[Jelinek and Mercer, 1980] Jelinek, F. and Mercer, R. (1980). Interpolated estimation of markov source parameters from sparse data. In E. Gelsema and L. Kanal, editors, Pattern Recognition in Practice, 1980.

[Jelinek, 1990] Jelinek, F. (1990). Self-Organized Language Models for Speech Recognition. In Waibel, A. and Lee, K.-F., editors, Readings in Speech Recognition, pages 450-506. Morgan Kaufman Publishers.

[Jelinek, 1997] Jelinek, F. (1997). Information Extraction From Speech And Text. In MIT Press, 1997.

[Jelinek, 2000] Frederick Jelinek. Statistical Methods for Speech Recognition. The MIT Press, Cambridge, Massachusetts, 2000.

[Jurafsky and Martin, 2000] Jurafsky, D. and Martin, J. (2000). Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition. Prentice Hall Series in Artificial Intelligence, 2000.

[Katz, 1987] Katz, S. (1987). Estimation of Probabilities from Sparse Data for the Language Model Component of a Speech Recognizer. In IEEE Transactions on Acoustics, Speech and Signal Processing, 35(3):400-401.

[Kemp and Waibel, 1998] Kemp, T. and Waibel, A. (1998). Reducing the OOV rate in broadcast news speech recognition. In Proceedings of ICSLP, pp. 1839-1842, 1998.

[Kirchhoff, 2002] Kirchhoff, K. (2002). Novel Speech Recognition Models for Arabic. Final Report of Johns-Hopkins University Summer Research Workshop 2002.

[Kirchhoff et al., 2006] Kirchhoff, K., Vergyri, D., Duh, K., Bilmes, J. and Stolcke, A. (2006). Morphology-based language modeling for Arabic speech recognition. Computer Speech and Language 20(4), pp. 589-608, 2006.

[Klakow, 1998] Klakow, D. (1998). Log-linear interpolation of language models. In Proceedings of Intl. Conf. on Spoken Language Processing, Sydney, Australia, 1998.

[Kneser and Ney, 1995] Kneser, R. and Ney, H. (1995). Improved backing-off for m-gram language modeling. In Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, volume 1, pages 181{184, Detroit, May 1995.

[Kneser et al., 1993] Kneser, R. and Steinbiss, V. (1993). On the dynamic adaptation of stochastic language models. In Proceedings of ICASSP, volume II, pages 586–589, Minneapolis, MN, 1993.

[Kneser et al., 1997] Kneser, R., Peters, J. and Klakow, D. (1997). Language Model Adaptation Using Dynamic Marginals. In Proceedings of EuroSpeech97, Rhodes, Greece, 1997.

[Kurimo et al., 2006] Kurimo, M., Creutz, M., Varjokallio, M. (2006). Unsupervised segmentation of words into morphemes – Morpho Challenge 2005 Application to Automatic Speech Recognition. In Proceedings of ICASSP, 2006.

[Kuhn et al., 1992] Kuhn, R. and De Mori, R. (1992). Correction to a cache-based natural language model for speech reproduction. In IEEE Transactions on Pattern Analysis and Machine Intelligence, 14(6):691–692.

[Kwon and Park, 2003] Kwon, O. and Park, J. (2003). Korean Large Vocabulary Continuous Speech Recognition With Morpheme-based Recognition Units. Speech Communication, Vol. 39, No. 3-4, pp. 287-300, 2003.

[Lavecchia et al., 2006] Lavecchia, C., Smaïli, K. and Haton, J. (2006). How to Handle Gender and Number Agreement in Statistical Language Models?. In Proceedings of InterSpeech 2006, Pittsburgh, PA, USA, 2006.

[Lavrenko et al., 2001] Lavrenko, V., and Croft, W. (2001). Relevance-Based Language Models. In proceedings of SIGIR'01, 2001.

[LDC-Hub4, 2000] LDC-Hub4 (2000). http://www.ldc.upenn.edu/Projects/Corpus_Cookbook/transcription/broadcast_speech/english/index.html.

[LEMUR, 2007] LEMUR (2007). The Lemur Toolkit - for Language Modeling and Information Retrieval. http://www.lemurproject.org.

[Lo and Gauvain, 2005] Lo, Y. and Gauvain, J. (2005). Topic Tracking on Audio Documents.

[Malouf, 2002] Malouf, R. (2002). A comparison of algorithms for maximum entropy parameter estimation. In Proceedings of Sixth Conf. on Natural Language Learning, pages 49-55, 2002.

[Martin et al., 1999] Martin, S., Ney, H. and Zaplo, J. (1999). Smoothing methods in maximum entropy language modeling. In Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, volume I, pages 545{548, Phoenix, March 1999.

[Martin et al., 2000] Martin, S., Kellner, A. and Portele, T. (2000). Interpolation of stochastic grammar and word bigram models in natural language understanding. In Proceedings of Intl. Conf. on Spoken Language Processing, Beijing, China, 2000.

[Martins, 1998] Martins, C. (1998). Modelos de Linguagem no reconhecimento de Fala Contínua. Master's thesis, Lisbon, Portugal: Instituto Superior Técnico, Universidade Técnica de Lisboa, 1998.

[Martins et al., 2005] Martins, C., Teixeira, A., and Neto, J. (2005). Language models in automatic speech recognition. Revista Electrónica e Telecomunicações, Departamento de Electrónica e Telecomunicações, Universidade de Aveiro, Aveiro, vol. 4, nº 4, 2005.

[Martins et al., 2006] Martins, C., Teixeira, A., and Neto, J. (2006). Dynamic Vocabulary Adaptation for a daily and real-time Broadcast News Transcription System. IEEE/ACL Workshop on Spoken Language Technology, December 2006.

[Martins et al., 2007] Martins, C., Teixeira, A., and Neto, J. (2007). Vocabulary Selection for a Broadcast News Transcription System using a Morpho-syntactic Approach. In Proceedings of InterpSpeech 2007, Antwerp, Belgium, 2007.

[Martins et al., 2007a] Martins, C., Teixeira, A., and Neto, J. (2007). Dynamic Language Modeling for a daily Broadcast News Transcription System. In Proceedings of ASRU 2007, Kyoto, Japan, 2007.

[Medeiros, 1995] Medeiros, J. (1995). Processamento Morfológico e Correcção Ortográfica do Português. Master's thesis, Instituto Superior Técnico, Lisbon, Portugal, 1995.

[Meinedo and Neto, 2000] Meinedo, H. and Neto, J. (2000). Combination of acoustic models in continuous speech recognition. In Proceedings of ICSLP 2000, Beijing, China.

[Meinedo et al., 2003] Meinedo, H., Caseiro, D., Neto, J. and Trancoso, I. (2003). AUDIMUS.media: A Broadcast News Speech Recognition System for the European Portuguese Language. In Proceedings of PROPOR 2003, Portugal, 2003.

[Meinedo and Neto, 2005] Meinedo, H., and Neto, J. (2005). A Stream-based Audio Segmentation, Classification and Clustering Preprocessing System for Broadcast News using ANN Models. In Proceedings of Interspeech 2005, Lisbon, Portugal, 2005.

[Meinedo, 2008] Meinedo, H. (2008). Audio Pre-processing and Speech Recognition for Broadcast News. In PhD Thesis, Lisbon, Portugal: Instituto Superior Técnico, Universidade Técnica de Lisboa, 2008.

[Moore and Young, 2000] Moore, G. and Young, S. (2000). Class-based language model adaptation using mixtures of word-class weights. In Proceedings of ICSLP, 2000.

[Mrva and Woodland, 2006] Mrva, D. and Woodland, P. (2006). Unsupervised language model adaptation for Mandarin broadcast conversation transcription. In Proceeedings of InterSpeech 2006, Pittsburgh, PA, USA, 2006.

[Neto et al., 2003] Neto, J., Meinedo, H., Amaral, R., and Trancoso, I. (2003). The development of an automatic system for selective dissemination of multimedia information. In Proceedings of Third International Workshop on Content-Based Multimedia Indexing – CBMI 2003, Rennes, France.

[Ney et al., 1994] Ney, H., Essen, U. and Kneser, R. (1994). On Structuring Probabilistic Dependencies in Stochastic Language Modeling. In "Computer Speech and Language", vol. 8, 1994.

[Ney et al., 1997] Ney, H., Martin, S. and Wessel, F. (1997). Statistical language modeling using leaving-one-out. In S. Young and G. Bloothooft, editors, Corpus-Based Methods in Language and Speech Processing, chapter 6, pages 174{207. Kluwer Academic Publishers, Dordrecht, 1997.

[Nguyen et al., 2005] Nguyen, L., Xiang, B., Afify, M., Abdou, S., Matsoukas, S., Schwartz, R., and Makhoul, J. (2005). The BBN RT04 english broadcast news transcription system. In Proceedings of InterSpeech 2005, Lisbon, Portugal, 2005.

[NIST, 2000] NIST (2000). Speech recognition scoring toolkit (SCTK). http://www.nist.gov/speech/tools/.

[Oger et al., 2008] Oger, S., Linares, G., Bechet, F. and Nocera, P. (2008). On-demand New Word Learning using World Wide Web. In Proceedings of ICASSP, Las Vegas, 2008.

[Orengo and Huyck, 2001] Orengo, V. and Huyck, C. (2001). A stemming algorithm for the Portuguese language. In Proceedings of the Eighth International Symposium on String Processing and Information Retrieval, 2001.

[Ostendorf et al., 2005] Ostendorf, M., Shriberg, E., and Stolcke, A. (2005). Human Language Technology: Opportunities and Challenges. In Proceedings of ICASSP, Philadelphia, 2005.

[Palmer and Ostendorf, 2005] Palmer, D. and Ostendorf, M. (2005). Improving out-of-vocabulary name resolution". Computer Speech and Language, vol. 19, 2005.

[Peters and Klakow, 2000] Peters, J. and Klakow, D. (2000). Capturing Long Range Correlations using Log-Linear Language Models. Verbmobil: Foundations of Speech-to-speech Translations ed. W. Wahlster (2000).

[Ponte, 1998] Ponte, J. (1998). A language modeling approach to information retrieval. PhD thesis, University of Massachusetts at Amherst, 1998.

[Rabiner and Juang, 1993] Rabiner, L. and Juang, B. (1993). Fundamentals of Speech Recognition. Prentice Hall Signal Processing Series, Alan V. Oppenheim, Series Editor, 1993.

[Ramabhadran et al., 2007] Ramabhadran, B., Siohan, O. and Sethy, A. (2007). The IBM 2007 Speech Transcription System for European Parliamentary Speeches. In Proceedings of ASRU 2007, Kyoto, Japan, 2007.

[Ribeiro, 2003] Ribeiro, R. (2003). Anotação morfossintáctica desambiguada do português. Master's thesis, Instituto Superior Técnico, Lisbon, Portugal, 2003.

[Ribeiro et al., 2004] Ribeiro, R., Mamede, N. and Trancoso, I. (2004). Morpho-syntactic Tagging: a Case Study of Linguistic Resources Reuse. Chapter of the book "Language Technology for Portuguese: shallow processing tools and resources", Edições Colibri, Lisbon, Portugal, 2004.

[Robertson, 1977] Robertson, S. (1977). The probabilistic ranking principle in IR. Journal of Documentation, 33:294–304, 1977.

[Rosenfeld, 1994] Rosenfeld, R. (1994). Adaptive Statistical Language Modeling: A Maximum Entropy Approach. In PhD Thesis, School of Computer Science, Carnegie Mellon University, 1994.

[Rosenfeld, 1995] Rosenfeld, R., (1995). Optimizing Lexical and n-gram Coverage via judicious use of Linguistic Data. In Proceedings of Eurospeech, vol. 2, 1995.

[Rosenfeld, 2000] Rosenfeld, R. (2000). Two decades of statistical language modeling: Where do we go from here?. In Proceedings of the IEEE, Volume 88, Issue 8, Aug 2000.

[Rotovnik et al., 2003] Rotovnik, T., Maucec, M., Horvat, B. and Kacic, Z. (2003). Slovenian Large Vocabulary Speech Recognition With Data-Driven Models Of Inflectional Morphology. In Proceedings of ASRU, 2003.

[Rotovnik, 2004] Rotovnik, T. (2004). Large Vocabulary Continuous Speech Recognition of Inflectional Language with Subword Units Stem – Ending. In PhD Thesis, Faculty of Electrical Engineering and Computer Science, Univeristy of Maribor, 2004.

[Rybach et al., 2007] Rybach, D., Hahn, S., Gollan, C., Schlüter, R. and Ney. H. (2007). Advances in Arabic Broadcast News Transcription at RWTH. In Proceedings of IEEE ASRU, Kyoto, Japan, December 2007.

[Salton et al., 1975] Salton, G., Wong, A. and Yang, C. (1975). A Vector Space Model for Automatic Indexing. Communications of the ACM, vol. 18, nr. 11, pages 613–620.

[Schwarm et al., 2004] Schwarm, S., Bulyko, I. and Ostendorf, M. (2004). Adaptive Language Modeling with Varied Sources to Cover New Vocabulary Items. IEEE Transactions on Speech and Audio Processing, vol. 12, n. 3, May 2004.

[Shriberg, 2005] Shriberg, E. (2005). Spontaneous speech: How people really talk, and why engineers should care.

[Singhal, 2001] Singhal, A. (2001). Modern Information Retrieval: A Brief Overview. Bulletin of the IEEE Computer Society Technical Committee on Data Engineering 24 (4): 35-43.

[Song and Croft, 1999] Song, F. and Croft, W. (1999). A general language model for information retrieval. In Proceedings of Eighth International Conference on Information and Knowledge Management (CIKM'99), 1999.

[Stolcke, 1998] Stolcke, A. (1998). Entropy-based Pruning of Backoff Language Models. In Proceedings of DARPA News Transcription and Understanding Workshop, Lansdowne, VA, 1998.

[Stolcke, 2002] Stolcke, A. (2002). SRILM - An Extensible Language Modeling Toolkit. in Proceedings of International Conference on Spoken Language Processing, Denver, USA, 2002.

[Strohman et al., 2005] Strohman, T., Metzler, D., Turtle, H., and Croft, W.B. (2005). Indri: A language-model based search engine for complex queries (extended version). CIIR Technical Report, 2005.

[Tam and Schultz, 2006] Tam, Y. and Schultz, T. (2006). Unsupervised Language Model Adaptation Using Latent Semantic Marginals. In Proceedings of InterSpeech 2006, Pittsburgh, PA, USA, 2006.

[Turtle and Croft, 1991] Turtle, H. and Croft, W. (1991). Evaluation of an inference network based retrieval model. Trans. Inf. Syst., 9(3):187-222, 1991.

[Venkataraman and Wang, 2003] Venkataraman, A. and Wang, W. (2003). Techniques for Effective Vocabulary Selection. In Proceedings of Eurospeech, 2003.

[Vergyri et al., 2004] Vergyri, D., Kirchhoff, K., Duh, K. and Stolcke, A. (2004). Morphology-based language modeling for Arabic speech recognition. In Proceedings of ICSLP '04, Jeju Island, Korea, Oct. 2004.

[Wang and Stolcke, 2007] Wang, W. and Stolcke, A. (2007). Integrating MAP, Marginals, and Unsupervised Language Model Adaptation. In Proceedings of InterSpeech 2007, Antwerp, Belgium, 2007.

[Wu et al., 2007] Wu, Y., Zhang, R. and Rudnicky, A. (2007). Data Selection For Speech Recognition. In Proceedings of ASRU 2007, Kyoto, Japan, 2007.

[Xiang et al., 2006] Xiang, B., Nguyen, K., Nguyen, L., Schwartz, R. and Makhoul, J. (2006). Morphological Decomposition for Arabic Broadcast News Transcription. In Proceedings of ICASSP, 2006.

[Yokoyama et al., 2003] Yokoyama, T., Shinozaki, T., Iwano, K. and Furui, S. (2003). Unsupervised Class-Based Language Model Adaptation for Spontaneous Speech Recognition. In Proceedings of ICASSP, 2003.

[Yu et al., 2000] Yu, H., Tomokiyo, T., Wang, Z. and Waibel, A. (2000). New developments in automatic meeting transcription. In Proceedings of the ICSLP, Beijing, China, October 2000.

[Zhu and Rosenfeld, 2001] Zhu, X. and Rosenfeld, R. (2001). Improving trigram language modeling with the world wide web. In Proceedings of ICASSP, Salt Lake City, Utah, 2001.

[Zue, 2007] Zue, V. (2007). On Organic Interfaces. In Proceedings of InterSpeech 2007, Belgium, 2007.