Machine Translation Enhanced Automatic Speech Recognition

Interactive Systems Laboratories (ISL)
Carnegie Mellon University, Pittsburgh, PA, USA
Universität Fridericiana zu Karlsruhe (TH), Karlsruhe, Germany

Diplomarbeit

by

Matthias Paulik

Advisors:
Dipl.-Inform. Christian Fügen
Dipl.-Inform. Sebastian Stüker
Dr.-Ing. Tanja Schultz
Prof. Dr.rer.nat. Alexander Waibel

May 2005

Hiermit versichere ich, die vorliegende Diplomarbeit selbständig und ohne unzulässige Hilfsmittel verfasst zu haben. Alle verwendeten Quellen sind im Literaturverzeichnis angegeben.

Karlsruhe, den 30. Mai 2005

Matthias Paulik

Abstract

In human-mediated translation scenarios, a human interpreter translates between a source and a target language using either a spoken or a written representation of the source language. In this work the recognition performance on the speech of the human translator spoken in the target language (English) is improved by taking advantage of the source language (Spanish) representations. For this, machine translation techniques are used to translate between the source and target language resources and then bias the target language speech recognizer towards the gained knowledge, hence the name Machine Translation Enhanced Automatic Speech Recognition (MTE-ASR).

Different basic MTE-ASR techniques are investigated, namely restricting the search vocabulary, selecting hypotheses from n-best lists and applying cache and interpolation schemes to language modeling. Given a written representation of the source language and with the help of a non-iterative combination of the most successful basic techniques, it was possible to outperform the English baseline ASR system by a relative word error rate reduction of 30.6%. In the case of a spoken source language representation, where a source language ASR has to be used at first to create a further processable written representation, the reduction is still 23.2%.

With the help of an iterative system design, which recursively applies the improved ASR output to enhance the involved MT system(s) for a further ASR improvement, it was possible to further increase these word error rate reductions to 37.7% and 29.9% respectively.

Zusammenfassung

Durch die EU-Osterweiterung hat sich die Anzahl der zuvor elf Amtssprachen des Europäischen Parlamentes auf 20 erhöht. Amtssprache bedeutet, dass jedes Ausschussdokument und jeder Antrag in diese Sprache übersetzt werden muss. Auch wenn während einer Ausschusssitzung nicht alle der 20 Amtssprachen angeboten werden (können), ist schon alleine der Aufwand der sich durch die Notwendigkeit der Simultanübersetzung der angebotenen Sprachen ergibt, immens. Für Sitzungen der Vereinten Nationen mit ihren sechs Amtssprachen ergibt sich eine ähnliche Situation. In Anbetracht dieser Sachlage ist es wünschenswert, Hilfsmittel zur Verfügung zu haben, die die Arbeit der zahlreichen Übersetzer erleichtert. Ein solches, sehr effektives Hilfsmittel, stellt ein Diktiersystem dar, welches dem Übersetzer erlaubt seine Übersetzung einfach per Spracherkennung eingeben zu können. So haben Experimente gezeigt [1], dass ein menschlicher Übersetzer bis zu viermal schneller arbeiten kann, sofern er seine Übersetzung nicht tippen muss sondern einfach diktieren kann. Noch höhere Steigerungsraten lassen sich durch die direkte Verwendung von Simultanübersetzungen zur Erzeugung von Transkripten einer Rede erzielen. Um den sich, bei auf diese Art und Weise erstellten Dokumenten, ergebenden Nachbearbeitungsaufwand möglichst gering zu halten, sollte die Fehlerrate des Diktiersystems möglichst gering sein. Gerade in der besonderen Situation einer diktierten Übersetzung ergeben sich durch die zur Verfügung stehenden Informationen in der Ursprungssprache (Originaldokument, Sprachsignal des Redners) sehr effektive Möglichkeiten zur Fehlerreduktion. So ist es möglich, diese in der Ursprungssprache gegebenen Informationen mit Hilfe einer maschinellen Übersetzung in die Zielsprache zu übersetzen und den Spracherkenner auf dieses Wissen hin auszurichten. Dieser Ansatz wurde unabhängig voneinander von Dymetman et al. [2] und von Brown et al. [1] bereits im Jahre 1994 vorangetrieben. Seither gab es einige weitere Veröffentlichungen zu diesem Thema.

In dieser Arbeit wird untersucht, wie die Erkennungsleistung eines Spracherkenners in der Zielsprache eines menschlichen Übersetzers (Englisch) mit Hilfe der in der Ursprungssprache (Spanisch) gegebenen und maschinell in die Zielsprache übersetzten Informationen verbessert werden kann[1]. Basierend auf bereits verfolgten Ansätzen zu diesem Thema werden zunächst einige grundlegende Techniken zur Verbesserung des Spracherkenners entwickelt. Zu diesen Techniken zählen die Einschränkung des Erkennervokabulars, die Auswahl von Hypothesen aus den n-besten Listen des Erkenners sowie das Anwenden von Cache- und interpolierten Sprachmodellen. Bei in Schriftform gegebenen ursprungssprachlichen Informationen war es möglich, durch eine nicht iterative Kombination der genannten grundlegenden Verbesserungstechniken, eine relative Wortfehlerratenreduktion von 30.6% zu erzielen. Sind die ursprungssprachlichen Informationen nur in gesprochener Form vorhan-

[1] Natürlich können Ziel- und Ursprungssprache auch einfach vertausch werden, wobei sich dann eine Verbesserung des Spracherkenners auf Seite der Ursprungssprache ergibt.

den, so müssen diese zunächst durch einen Spracherkenner in der Ursprungssprache in Schriftform gebracht werden, um von der maschinellen Übersetzungskomponente weiter verarbeitet werden zu können. Aufgrund dieser zusätzlichen möglichen Fehlerquelle sinkt die Wortfehlerratenreduktion hier auf 23.2% ab.

Neben der nicht iterativen Kombination der grundlegenden Verbesserungstechniken wird eine darauf aufbauende iterative Kombination untersucht. Grundidee dieses iterativen Systementwurfes ist es, die schon verbesserte Spracherkennerausgabe rekursiv zur Verbesserung der beteiligten Übersetzungskomponente(n) zu verwenden und durch die verbesserte maschinelle Übersetzung wiederum eine weitere Steigerung der Spracherkennerleistung zu erzielen. Mit Hilfe dieses iterativen Ansatzes ist es möglich die Wortfehlerratenreduktion auf 37.7% beziehungsweise 29.9% zu steigern.

Acknowledgements

First of all, I would like to thank Prof. Alex Waibel for giving me the opportunity to perfrom my research at the U.S. America part of the Interactive System Labs, at Carnegie Mellon University in Pittsburgh. I am very grateful for the excellent support and advice I experienced from my advisors and/or future colleagues, Sebastian Stüker, Christian Fügen, Tanja Schultz, Thomas Schaaf and Matthias Eck. I would also like to thank everyone else at the Interactive System Labs (both in Pittsburgh and in Karlsruhe) who contributed to the successful completion of this work, by giving me valuable advice on questions concerning the machine translation component (special thanks go here to Ashish Venugopal, Silja Hildebrand, Muntsin Kolss and Dietmar Bernreuther) or by being helpful in administrative matters or by participating in the data collection. In regard to the data collection, I would like to especially thank Anton Prevosti and Dorcas Alexander for their support in the Spanish part of the collection. Finally, I want to thank my parents for their ever-present support.

Contents

Chapter 1

Introduction

1.1 Automatic Speech Recognition

Speech recognition systems for large vocabulary continuous speech recognition are nowadays widely available. Those systems are based on statistical methods, in which the so-called *fundamental equation of speech recognition* is taking center stage:

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

This equation indicates that to find the most probable word sequence $\hat{W}$ given the observed sequence Y of feature vectors extracted from the acoustic signal, the product of $P(W)$ and $P(Y|W)$ has to be maximized (the denominator $P(Y)$ is independent of W and can be ignored). The *language model* (LM) $P(W)$ determines the *a priori* probability of observing the word sequence W. The *acoustic model* $P(Y|W)$ represents the probability of observing the feature vector sequence Y given W. Different central questions of Automatic Speech Recognition (ASR) can be directly derived from equation 1.1:

- Signal preprocessing: Which kind of signal preprocessing should be used to extract the sequence of feature vectors from the acoustic signal?

- Language and acoustic modeling: How should the language model and the acoustic model be represented/computed?

- Decoding: How can the sequence of words $\hat{W}$, which maximizes equation 1.1, be found? (Given the combinatorial explosion associated with large vocabularies, an efficient pruning of the search space is of particular importance to the decoding process.)

Although already published in 1996, [3] still gives a good overview of the principles applied in current Large Vocabulary Recognition (LVR) systems to deal with the mentioned problems.

1.2 Statistical Machine Translation

The basic principle of the statistical methods used in automatic speech recognition were successfully applied to machine translation (MT). The most probable word sequence $\hat{T}$ of words in the target language given the word sequence S in the source language can be computed with the help of the *fundamental equation of statistical machine translation*:

$$\hat{T} = \arg\max_{T} P(T|S) = \arg\max_{T} P(T)P(S|T) \tag{1.2}$$

$P(T)$ is again called the language model (of the target language). The translation model (TM) $P(S|T)$ gives the translation probability of S given T. Again, an efficient search algorithm is needed to find the best target sentence that maximizes equation 1.2. A more detailed introduction to statistical machine translation can be found in [4].

1.3 Machine Translation Enhanced Automatic Speech Recognition

In this work, the term *Machine Translation Enhanced Automatic Speech Recognition* (MTE-ASR) is defined as generic term for all techniques that are aimed to improve the recognition accuracy of an ASR system with the help of available resources in one or more languages different from the ASR system language, whereas these resources are at first being translated by a machine translation component into the language of the ASR system.

Human-mediated translation scenarios in which a speaker of one language communicates with one or several speakers of another language with the help of a bilingual human interpreter are particularly suited for MTE-ASR based applications. One example for such a human-mediated translation scenario is an American aid worker speaking with a non-American victim through a human interpreter. Another example is a Spanish speaker delivering a speech to a non-Spanish audience as commonly seen in European Parliament or United Nations debates. In the latter example, one (or several) interpreters would translate the Spanish spoken presentation into the target language(s) of the listeners. This happens either directly from the spoken speech or with the help of a transcript of the delivered speech. In both examples, it is desirable to have a written transcript of what was said by the speaker in the source language and of what was said by the interpreter(s) in their respective target languages, e.g. for archiving and retrieval, or publication. The most straight-forward technique is to record the speech of the speaker and the interpreter(s) and then use automatic speech recognition to transcribe the recordings. Since additional knowledge in the form of a spoken and/or a written representation of the source/target language is available, it can be used to improve the performance of the ASR. One possibility is the use of machine translation to translate

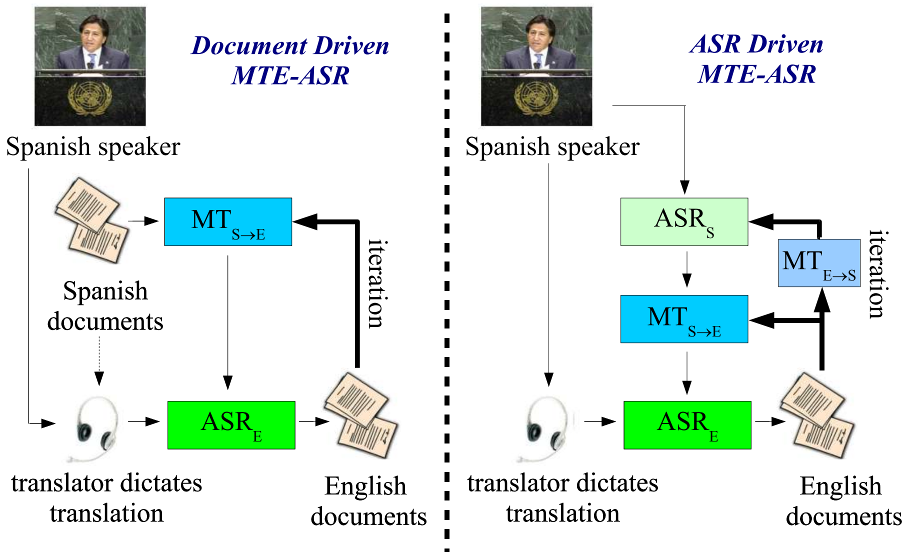


Figure 1.1. Document driven and ASR driven MTE-ASR.

these resources into the language of the respective ASR system. This work concentrates on the specific case where the ASR system for the target language of one interpreter is to be improved. Such a scenario is illustrated in Figure 1.1.

As shown in Figure 1.1, two basic application scenarios can be distinguished: scenarios in which a written representation of the source language is available and scenarios in which such a written representation has to be created first from the spoken representation with the help of a source language ASR system. In the following, the former case will be referred to as *Document Driven MTE-ASR* and the latter as *ASR Driven MTE-ASR*.

1.4 Iterative MTE-ASR

MTE-ASR seeks to improve the performance of automatic speech recognition with the help of available resources in languages different from the ASR system language by using machine translation to translate those resources into the ASR system language. In the same manner, it is possible to improve the performance of a MT system by using automatic speech recognition. A way to accomplish such an improvement would be, for example, to use the translation transcription provided by the target language ASR together with the source documents and/or transcriptions of the source language ASR as additional training data. This motivates the feedback loop of the iterative MTE-ASR system design de-

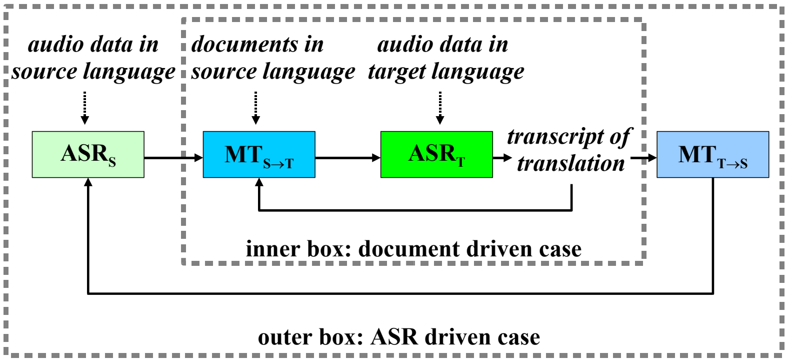

audio data in source language
documents in source language
audio data in target language
ASR_S
MT_S→T
ASR_T
transcript of translation
MT_T→S
inner box: document driven case
outer box: ASR driven case

Figure 1.2. Iterative MTE-ASR.

picted in Figures 1.1 and 1.2. It is noteworthy that for the ASR driven case the improvement of the source language ASR and the target language ASR is automatically combined by this iterative design (feedback loop).

1.5 Objective

Several successful MTE-ASR approaches have been developed in recent years to provide professional translators with a high quality automatic dictation tool. A short overview on those approaches is given in chapter 2. In chapter 3, several basic MTE-ASR techniques that are based on those ideas are developed and compared. Furthermore, the most promising techniques are combined and integrated into the above described iterative MTE-ASR system design to examine the feasibility of this iterative approach. This is done in chapter 4 for the document driven case and in chapter 5 for the ASR driven case. As a consequence of the iterative system design, several techniques are examined that improve the performance of the involved MT systems with the help of the output provided by the involved ASR systems.

Chapter 2

Related Work

Some publications on MTE-ASR for developing an automatic dictation system for professional translators are available. However, given the fact that this is a very specific application, the number of publications is relatively small. This chapter gives a short, to the authors knowledge complete, overview of all these publications.

2.1 The TransTalk Project

Dymetman et al. introduce in [2] a prototype version of their dictation tool TransTalk. The translation direction is English to French and they assume that the transcript of the English sentence is known for each spoken French sentence. The prototype version operates as an isolated-word recognizer over a 20K French vocabulary. They achieve an average word error rate reduction of 24% relative over their baseline system by first using the isolated-word recognizer to prune the 20K word search space to the n (20) most acoustically probable words for each acoustic token and then performing a Viterbi search through the remaining sentence candidates using the translation model together with the available English source sentence.

Brousseau et al. describe in [5] version two and three of TransTalk. Version two extends the n-best technique applied in the prototype version to continuous speech recognition. The speech recognizer, which is based on a bi-gram language model, produces a n-best list of French sentence hypotheses and the translation

	word correct	sentence correct
ASR with bi-gram LM	80.7%	4.0%
Rescoring with tri-gram LM	84.5%	8.7%
Rescoring with tri-gram LM and TM	86.0%	12.7%

Table 2.1. TransTalk version 2: Rescoring of ASR n-best hypotheses (n=200).

model, now interpolated with a tri-gram language model, is again used to select one hypothesis. This system was tested on 300 Hansard sentences (6,639 words) without OOV words and only up to 40 words per sentence. The results for version 2 can be found in Table 2.1. It is reported that this approach takes about 93 times real-time.
In version three, the translation model is used before recognition on a French sentence to generate a dynamic vocabulary from the English sentence. The recognizer vocabulary is then constrained to this dynamic vocabulary. The used baseline ASR system runs at 15.8 times real-time and yields a 75.7% word correct rate on the above described test set. Using a dynamic vocabulary with 2,000 words a run time of 5.4 times real-time and 77.1% word correct rate could be accomplished.

2.2 Automatic Speech Recognition in Machine Aided Translation

Brown et al. describe in [1] the possibility of combining speech recognition and machine translation by formulating:

$$\hat{T} = \arg\max_{T} P(T|A,S) = \arg\max_{T} P(A|T)P(T)P(S|T) \tag{2.1}$$

T is the word sequence in the target language, S the word sequence in the source language and A the sequence of acoustic feature vectors. This is identical to the fundamental equation of speech recognition (see equation 1.1) except that the target language model $P(T)$ is now multiplied with the translation model $P(S|T)$. Brown et al. deduce from this that machine translation can be incorporated into speech recognition by "some judicious fiddling with the language probabilities." On a test set of 1,000 Hansard sentences, they accomplish a per-word perplexity decrease from 63.6 to 17.2 by augmenting their standard tri-gram LM with translation model probabilities.

2.3 Cheating with Imperfect Transcripts

In [6] Placeway and Lafferty describe how closed-caption information can be used to improve the quality of an automatic transcription system for television broadcasts. The closed-caption information used is provided in the language of the transcription system. The pursued approach is nevertheless analogous to the MTE-ASR approach presented in [1], as the questions arises how a caption (or rather the "hint" a caption provides) H is being generated from a text W:

$$\hat{W} = \arg\max_{W} P(W|A,H) = \arg\max_{W} P(A|W)P(W)P(H|W) \tag{2.2}$$

A is again the sequence of acoustic feature vectors. The used translation model $P(H|W)$ computes the minimal string edit distance with words as units.

	Standard LM	Interpolated LM
Standard ASR	59.8%	47.8%
ASR + TM	28.5%	18.2%

Table 2.2. Cheating with Imperfect Transcripts: WERs for a NBC Nightly News transcription.

This means that, during decoding, for each partial hypothesis the edit distance to the caption is computed and added in an appropriate way to the score of the hypothesis. It is reported that when keeping all other things equal, the approach affects the search with a 10% slowdown, but that generally a modest increase in overall speed can be observed due to pruning effects. In addition to this approach, an interpolation of the language model with the text of the closed-captions was taken into consideration. Table 2.2 shows the word error rates (WER) for the transcription of a NBC Nightly News show from April 1995.

2.4 MT and Topic-Based Techniques to Enhance Speech Recognition Systems for Professional Translators

The vocabulary approach presented in the TransTalk project [5] was reinvestigated by Ludovik and Zarchaski [7]. For this, the vocabulary used by two independent translators for the translation of 10 Spanish newspaper articles into English was compared to the vocabulary produced by a MT component. Roughly 1/3 of the words used by the professional translators were not included in the vocabulary produced by the MT. Another method examined in [7] used the MT system for topic detection and then chose an appropriate, precomputed, topic-specific language model. With this approach, the error rate of the English ASR system could be reduced from 9.98% to 5.07%.

2.5 Summary

The presented MTE-ASR approaches differ in the way MT knowledge is used to influence the ASR search process. The dynamic vocabulary technique restricts the search space before the actual decoding. Language model interpolation, selecting an appropriate topic specific LM through topic detection and the explicit computation of translation probabilities during decoding (which again can be seen as ”fiddling” with the LM probabilities), influences the search in itself as the probabilities of the considered (partial) hypotheses are being changed. The potential computational overhead, caused by an explicit computation of TM probabilities, can possibly be staved off by pruning effects. Last, but not least, rescoring the n-best ASR hypotheses with the help of MT knowledge does not influence the ASR decoding process itself.

Chapter 3

Comparison of Basic MTE-ASR Techniques

In this chapter, different basic MTE-ASR techniques that are based on the approaches presented in chapter 2 will be introduced and compared. The term basic refers here to the fact that the iterative MTE-ASR system design is not yet taken into consideration. Therefore, only the baseline MT knowledge is used for ASR improvement. Techniques to improve the MT component of the iterative system are presented in chapter 4.

3.1 Experimental Setup

3.1.1 Scenario

The scenario considered for the examined ASR improvement techniques in this chapter can be characterized as document driven and non-iterative:

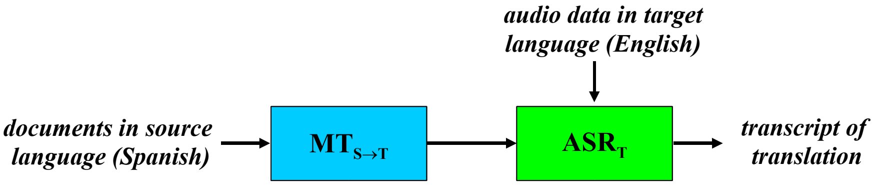

audio data in target language (English)
documents in source language (Spanish)
$MT_{S\rightarrow T}$
ASR_T
transcript of translation

Figure 3.1. Document driven, non-iterative MTE-ASR.

3.1.2 Used Performance Metrics

The primary performance metrics used in this work are word error rate (WER) for measuring ASR performance and BLEU score for measuring MT performance. If not mentioned otherwise, the involved system components will be tuned in regard to these primary performance metrics, i.e. the involved ASR systems will be tuned in regard to WER and the involved MT systems will be tuned in regard to BLEU score. In addition to WER and BLEU score, NIST score and n-best word error rate (nWER) will be given. Throughout this work the nWER is always given for $n = 150$. The BLEU and NIST scores are given in regard to just one reference translation, namely the transcript of the translation spoken by the human translator. BLEU and NIST score were computed with the help of the MT evaluation tool mteval-v09c.pl which can be downloaded from the NIST (National Institute of Standards and Technology) homepage[1].

The word error rate is based on the minimal edit distance between hypothesis and reference sentence, this means it is based on the minimal number of substitutions s, insertions i and deletions d necessary to transform the hypothesis into the reference. With n the number of reference words, the WER is given as:

$$WER = \frac{s+i+d}{n} * 100\% \tag{3.1}$$

The n-best word error rate is the minimal WER found within the n-best hypotheses for a reference, i.e. the nWER is equivalent to the WER of the n-best hypothesis with the best WER.

The BLEU score [8] computes the geometrical mean of the modified n-gram precisions with $n \in \{1; 2; 3; 4\}$ and applies a length penalty to translation hypotheses that are shorter than the, in regard to its length, best matching reference translation. The n-gram precisions are modified in a way to serve the "intuitive" demand for considering a reference n-gram as exhausted after a matching candidate n-gram is identified. In its original definition, the BLEU score ranges from 0 to 1, whereas a translation that is identical to a reference translation attains a score of 1. However, throughout this work the BLEU score will be given in the range form 0 to 100, i.e. multiplied by the factor 100, as it is sometimes seen in MT related publications.

The NIST score [9] is a variation of the BLEU metric but instead of n-gram precisions information-weighted n-gram counts are used, i.e. more informative n-grams are weighted more heavily. The NIST score ranges from the worst possible score 0 to a best maximal value that is dependent of the used reference translations. As the transcripts of the human translator are used as reference translation in this work, there is only one English reference translation for each

[1] http://www.nist.gov/speech/tests/mt/resources/scoring.htm

	WER	nWER	BLEU	NIST
English Baseline ASR	12.6	6.5	82.9	10.8
Spanish to English MT	46.8	34.2	40.4	7.1

Table 3.1. English baseline ASR and Spanish to English MT performance.

Spanish sentence. It is therefore possible to compute the maximal possible NIST scores on the used data sets by using the reference translations as translation hypotheses. The in this manner computed maximal NIST scores can be found in the following sections where the used data sets are described more closely.

3.1.3 Data

The used test data set (data set I) consists of 506 parallel Spanish and English sentences taken from the bilingual Basic Travel Expression Corpus (BTEC). The 506 English sentences were presented four times, each time read by different speakers. After removing some corrupted audio recordings, a total of 2,008 spoken utterances composed of 12,010 (798 different) words was taken as the final data set. This equals 67 minutes of speech from 12 different speakers. The complete data set was used for tuning the parameters of the described MTE-ASR systems. Generalization accuracy over unseen data will be examined along with the iterative MTE-ASR system design on a different test data set (data set II) in chapter 4 and 5. The best possible NIST score on this data set (data set I) is 12.1.

3.1.4 Baseline ASR

For the ASR experiments in this work, the Janus Recognition Toolkit (JRTk) featuring the IBIS single pass decoder [10] was used. The sub-phonetically tied three-state HMM based recognition system has 6 K codebooks, 24 K distributions and a 42-dimensional feature space on MFCCs after LDA. It uses semi-tied covariance matrices, utterance-based CMS and incremental VTLN with feature-space MLLR. The recognizer was trained on 180h Broadcast News data and 96h Meeting data [11]. The back-off tri-gram language model was trained on the English BTEC (not including the test data set), which consists of 162.2 K sentences with 963.5 K running words from 13.7 K distinct words. The language model perplexity on the data set described above is 21.6. The dictionary has 19.8 K entries (18.3 K without pronunciation variants), with the 13.7 K BTEC words as a subset. No gain in recognition accuracy could be observed for reducing the dictionary to the 13.7 K BTEC words; therefore, the original 19.8 K dictionary was kept. The OOV rate on the data set is 0.53%. After system parameter tuning, a word error rate (WER) of 12.6% was achieved. N-best WER, BLEU score and NIST score can be found along with the performance values for the used Spanish to English MT system in Table 3.1.

3.1.5 MT System

The ISL statistical machine translation system [12] was used for the Spanish to English automatic translations. This MT system is based on phrase-to-phrase translations (calculated on word-to-word translation probabilities), extracted from a bilingual corpus, in our case the Spanish/English BTEC (not including the test data set). It produces a n-best list of translation hypotheses for a given source sentence with the help of its translation model (TM), target language model and translation memory. The translation memory works as follows: for each source sentence that has to be translated, the closest matching (in regard to edit distance) source sentence is searched in the training corpus and extracted along with its translation. In case of an exact match, the extracted translation is used. Otherwise, different repair strategies are applied to find the correct translation. The translation model computes the phrase translation probability based on word translation probabilities found in its statistical IBM1 forward and backward lexica regardless of the word order:

$$p(s|h) = \prod_j \sum_i p(s_j|h_i) \tag{3.2}$$

The word order of MT hypotheses is, therefore, appointed by the language model and translation memory. As the same language model is used as in the ASR baseline system, one can say that only the translation memory can provide additional word order information for ASR improvement. The tuned system gave a BLEU score of 40.35. Refer to Table 3.1 for the according NIST score, WER and nWER.

3.1.6 Handling of MT OOV words

The MT system hands unknown Spanish words on without changing them. This means the English translations can contain Spanish words. In the case of words with identical orthography in English and Spanish (this is mostly the case for proper names), it is, therefore, possible to reduce the OOV rate of the ASR system by automatically computing the English pronunciations for unknown MT words. The OOV rate of the ASR system could be reduced from 0.53% to 0.48% with this approach. However, no change in recognition accuracy could be observed. Given the relatively low OOV rate, it is very unlikely to see any significant gains with this approach on the described data set. For this reason, no extension of the ASR dictionary with unknown Spanish words was done for the experiments described in this work.

3.1.7 Used MT n-best List Sizes

The MTE-ASR approaches described in the following make use of the MT n-best translation hypotheses in various ways. Therefore, the question of the optimal n-best list size occurred frequently. It became apparent that for the successful improvement techniques relatively small n-best list sizes, most of the times in the

n	Size of n-best lists vocabulary	Coverage of test set vocabulary	Average number of different translations
1	810	72%	1
10	1159	80%	9.86
20	1393	83%	19.29
40	1669	85%	36.06
80	1967	86%	59.80

Table 3.2. Analysis of MT n-best lists over the complete data set.

range of [1; 40], but always well beneath $n = 100$, were sufficient. To motivate this observation, a basic analysis of the MT n-best lists was done. This analysis showed that with increasing n the n-best list vocabulary size increases notedly faster compared to the coverage of the test set vocabulary. For details refer to Table 3.2.

3.2 Vocabulary Restriction

In the related work presented in chapter 2 inconsistent results have been reported for restricting the search space of the ASR system by restricting its vocabulary. The success of this approach is highly dependent on the quality of the automatically created translations compared to the spoken translations that are to be recognized. Even if the MT would produce ”perfect” translations in respect to reference translations given by another human translator, restricting the ASR vocabulary in the hope to minimize the room for possible recognition errors may be more damaging than helping if the spoken translations differ too much from these reference translations.

To examine the usability of vocabulary restriction on the given system configuration, the baseline ASR system was restricted to the words found within all MT n-best lists, i.e. the vocabulary was not dynamically computed for each sentence as in [5]. For an MT n-best list of size $n = 1$, a WER of 26.0% was achieved, which continuously decreased with larger n, reaching a WER of 19.6% for $n = 150$. A lower bound of 15.0% for $n \to \infty$ was computed by adding all OOV words to the $n = 150$ vocabulary. None of these vocabulary restricted ASR systems could outperform the baseline system. Therefore, the vocabulary restriction approach was not further pursued.

3.3 Language Model Interpolation

Two different approaches were examined to adapt the English baseline ASR language model to the used data set by applying LM interpolation. First, the baseline ASR language model was interpolated with a small back-off tri-gram language model computed on all MT n-best lists that were created for the

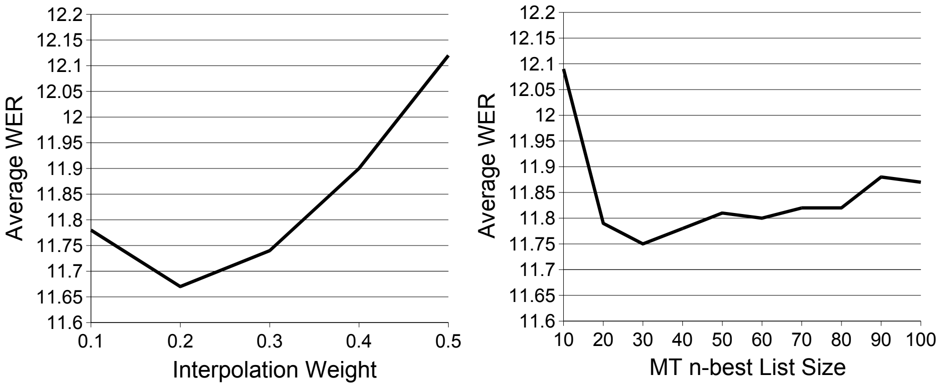
Spanish sentences. Second, for each English sentence a separate interpolated LM was dynamically computed by interpolating the baseline LM with a small back-off tri-gram language model computed on the one MT n-best list that was created for the respective Spanish sentence. In the following the first approach will be referred to as ”standard language model interpolation” and the second as ”dynamic language model interpolation”.

3.3.1 Standard Language Model Interpolation

For these experiments 10% of the English sentences found in the BTEC were randomly selected as held-out data set and a new English baseline LM was computed on the reduced BTEC. The interpolation weight w of the small MT language model was automatically computed with tools provided by the SRI Language Modeling Toolkit [13] so that the perplexity on the 10% held-out data set became minimal. The in this manner found best interpolation parameters were $n = 30$, $w = 0.2035$. The perplexity of the interpolated LM on the test data set was now 17.0, the WER was 11.9%.
In addition to the described experiment the development of the WER for using different values of n and w when interpolating the original baseline LM, as it was described in section 3.1, was examined. The best setting for the interpolation weight, based on the average WER (averaged on WERs for the different used values of n), was $w = 0.2$. The best setting for the MT n-best list size, again based on the, accordingly for different values of w computed average WER, was $n = 30$. This is in compliance with the results from the first experiment. Figure 3.2 (a) shows the average WERs for the different interpolation weights and Figure 3.2 (b) shows the average WERs for the different n-best list sizes. The system with $w = 0.2$, $n = 30$ had a WER of 11.62%. The best parameter setting based on the absolute WER (not the average WER) was $w = 0.2$, $n = 20$ and yielded a WER of 11.60%. This is an absolute gain of 1.0% compared to the baseline WER of 12.6%. The MT word context information given in the MT n-best hypotheses could, therefore, successfully be applied to adapt the baseline LM resulting in an improved ASR recognition performance.

3.3.2 Dynamic Language Model Interpolation

The idea behind dynamic LM interpolation was to make additional use of the available alignment information: for each English sentence the corresponding Spanish sentence is known. With this knowledge it is possible to further adapt the LM to each individual English sentence by dynamically computing an interpolated LM for each sentence with the help of its MT n-best list. No gain in performance compared to the baseline system could be observed for this approach. The best interpolation weight (in regard to WER) was again $w = 0.2$, but the best MT n-best list size was with $n = 90$ now three times as high as for the non-dynamic case. The system with these settings yielded a WER of 13.2%. The higher MT n-best list size can be explained by the less of available adaption data, i.e. to compensate the missing information that is additionally



(a) (b)

Figure 3.2. Average WERs for LM interpolation.

given when using all MT n-best lists the size of the used MT n-best list has to be increased. However, this compensation is accompanied with the use of lower ranking translation hypotheses that are of a smaller value for the LM adaption.

3.4 Hypothesis Selection by Rescoring

The n-best WER found within the ASR 150-best lists of the baseline system is 6.5% showing the huge potential of rescoring the ASR n-best lists. In contrast to this, no such potential is evident for rescoring the MT 150-best lists as only a minimal WER of 34.2% can be achieved on these. However, when combining the n-best lists of ASR and MT, the nWER reduced to 4.2% which proves that complementary information is given in the n-best lists of both components. In fact, a performance gain could be observed for enriching the ASR 150-best lists with the first best MT hypothesis prior to rescoring.

All rescoring experiments mentioned in this work use ASR 150-best lists that are enriched with the first best MT hypothesis, i.e. there are up to 151 hypotheses in the n-best lists used for rescoring.

The applied rescoring algorithm computes new scores (negative log-probabilities) for each sentence by summing over the weighted and normalized translation model score, language model score, and ASR score of this sentence. To compensate for the different ranges of the values for the TM, LM and ASR scores, the individual scores in the n-best lists were scaled to [0; 1].

$$s_{final} = s'_{ASR} + w_{TM} * s_{TM} + w_{LM} * s_{LM} \tag{3.3}$$

The ASR score output s_{ASR} by the JRTk is an additive mix of acoustic score, weighted language model score (with the weight lz), word penalty lp and filler word penalty fp. The language model score within this additive mix contains fixed discounts for special words or word classes.

$$\begin{aligned} s_{ASR} = & s_{acoustic} + lz * (s_{LM} - DiscountsForSpecialWords) \\ & + lp * n_{words} + fp * n_{fillerwords} \end{aligned} \tag{3.4}$$

The rescoring algorithm allows to directly change the word penalty, and the filler word penalty added to the acoustic score. Moreover, four new word context classes with their specific LM discounts are introduced: MT mono-, bi-, tri-grams and complete MT sentences. MT n-grams are n-grams included in the MT n-best list of the respective sentence; MT sentences are defined in the same manner. The ASR score in equation 3.3 is, therefore, computed as:

$$\begin{aligned} s'_{ASR} = & s_{ASR} + lp' * n_{words} + fp' * n_{fillerwords} \\ & - md * n_{MTmonograms} - bd * n_{MTbigrams} \\ & - td * n_{MTtrigrams} - sd * \delta_{isMTsentence} \end{aligned} \tag{3.5}$$

Parameter optimization was done by manual gradient descent. The best parameters turned out to be $w_{TM} = 0.2$, $w_{LM} = 0.4$, $md = 58$, $fp' = -35$, $n = 20$, and all other parameters set to zero (the baseline system had a LM weight of $lz = 32$ and the settings $lp = -5$, $fp = 25$). The parameter n assigns the size of the MT n-best lists used for defining the above mentioned word context classes. The system yielded a WER of 10.5%, which corresponds to a relative reduction of 16.9%. The MT is not able to produce/score non-lexical events seen in spontaneous speech. This accounts for the negative rescoring filler penalty of $fp' = -35$; the ASR score has to compete with the filler penalty free TM and LM scores during rescoring.

This approach offers a successful way of applying MT knowledge for ASR improvement without changing the ASR system. MT knowledge is applied in two different ways; by computing the TM score for each individual hypothesis and by introducing new word class discounts based on MT n-best lists. Of the word class discount parameters only the mono-gram discount is different from zero. This shows that the word context information provided by the MT is of little value to the ASR. On the other hand, the mono-gram discount contributes largely to the success of this approach: the best WER found without any word class discounts was 11.5%. Thus, the MT is not very useful for getting additional word context information, but very useful as a provider for a "bag of words," that predicts which words are going to be said by the human translator.

3.5 Cache Language Model

Since the mono-gram discounts have such a great impact on the success of the rescoring approach, it is desirable to use this form of MT knowledge not

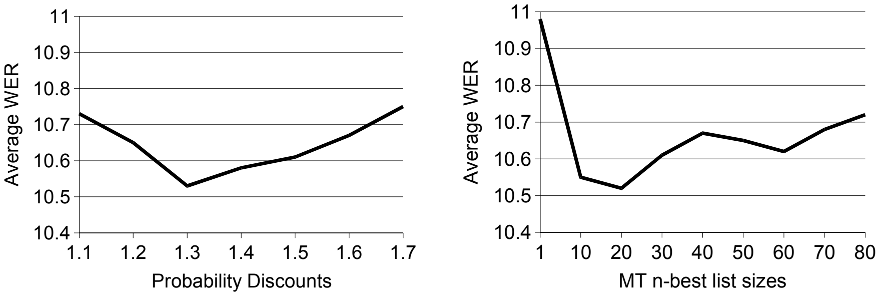


Figure 3.3. Average WERs for the cache LM approach.

only after, but already during ASR decoding. This will influence the pruning applied during decoding in a way that new, correct hypotheses can be found.

A classical cache language model has a dynamical component (a ”cache”) that remembers the recent word history of m words to adjust the language model probabilities based on this history. Similar to this definition, the cache LM used in this experiment has a dynamically updated ”cache” and the LM probabilities are influenced based on the content of this cache. However, the cache is not used to remember the recent word history but to hold the words (mono-grams) found in the respective MT n-best list of the English sentence that is being decoded at the moment. This cache LM was realized by taking advantage of the in the JRTk given possibility to define language model discounts for special words or word classes (compare equation 3.4), i.e. the cache LM was realized by introducing the word class MT mono-gram in the same manner as in section 3.4 but now defining its members dynamically during decoding. Resulting from this proceeding, two cache LM parameters have to be adjusted: the MT n-best list size n and the log probability discount d of the word class MT mono-gram.

In addition to testing different cache LM parameter settings, different settings for lz, lp and fp were taken into consideration. It could be observed that the optimal values for these parameters are interdependent, i.e. the best performance can be expected when tuning all of these parameters together. However, for all reasonable settings of lz, lp and fp (settings with a good performance on the baseline system), settings for the cache LM parameters n and d could be found that yielded similar good word error rates. The best performing system used the settings: $n = 20$, $d = 1.3$, $lz = 32$, $lp = 10$ and $fp = 40$. It had a WER of 10.4%. Figure 3.3 shows the average word error rates for different n-best list sizes and different log probability discounts.

This approach yields a similar performance as the rescoring approach, but in contrast to the rescoring approach, only two parameters have to be tuned (as mentioned above was the additional tuning of lz, lp and fp of less importance). Moreover, the expectation to find new, correct hypotheses could be fulfilled;

the nWER for the Cache LM system output was now 5.5% in comparison to 6.5% of the baseline system.

The applied method was quite simple: a fixed LM probability discount was used for all MT mono-grams. A more sophisticated approach would be, for example, to increase the probabilities of words that occur very often in the respective n-best list by a greater value than the probabilities of words that occur less often. Some additional experiments referring to this idea have been done but were not further pursued because of their small gain in performance. Descriptions for these experiments can be found in appendix A.

3.6 Combination of Different Techniques

The MTE-ASR techniques examined so far applied different forms of MT knowledge with varying success. For example, language model interpolation used MT word context information found within the MT n-best lists in the form of trigrams (and bi- and mono-grams for back-off). In contrast to this, the cache LM approach only used MT mono-grams. Therefore, the question arises if it is possible to further improve the recognition accuracy by a combination of the introduced techniques. Several experiments for a direct, non-iterative combination of the MTE-ASR procedures described so far were performed. For all these experiments, the parameters for word penalty, filler word penalty and language weight were fixed to $lz = 32$, $lp = 10$ and $fp = 40$.

3.6.1 Cache + Interpolated LM

For combining the cache and interpolated LM schemes a minimal WER of 10.1% was obtained when using the cache LM parameters $n_c = 20$, $d = 1.4$ and the interpolation LM parameters $w = 0.1$, $n_i = 60$. This is only a small improvement compared to the cache LM. We can argue that the MT context information used within the interpolated LM is of little value and that the success of the interpolated LM approach is largely due to mono-gram backing-off. As the cache LM approach is already based on MT knowledge provided through MT mono-grams, the combination with the interpolated LM can only yield small improvements.

3.6.2 Hypothesis Selection on Cache LM System Output

For this experiment, the rescoring algorithm described above was used on the n-best lists produced by the best found cache LM system. The best WER found was 9.35% when using the parameter setting $w_{TM} = 0.075$, $w_{LM} = 0.025$, $bd = 2$, $sd = 2$, $fp' = -20$, $lp' = 5$, $n = 20$ and all other parameters set to zero. The WER is only slightly different if no word class discounts are used. This can be explained by the fact that MT knowledge in the form of mono-gram discounts is already optimally used by the cache LM. Moreover, when keeping all rescoring parameters fixed to zero except for the translation model weight w_{TM},

Rescoring using	best WER	parameter settings
-	10.34	-
only fp',lp'	10.30	$fp' = -10$, $lp' = 5$
only w_{LM}	10.28	$w_{LM} = 0.025$
only fp',lp' and MT n-gram discounts	10.03	$bd = 18$, $n = 20$ $fp' = -10$, $lp' = 5$
only w_{TM}	9.55	$w_{TM} = 0.075$
all parameters	9.35	$w_{TM} = 0.075$, $w_{LM} = 0.025$, $bd = 2$, $sd = 2$, $n = 20$, $fp' = -20$, $lp' = 5$

Table 3.3. Rescoring on cache LM system output which had a WER of 10.41%. The results are given in regard to different parameter settings. Unlisted parameters were set to zero. The decline in WER for keeping all rescoring parameters fixed to zero can be explained by the positive effect of enriching of the ASR 150-best list with the first best MT hypotheses.

a minimal WER of 9.55% was accomplished, again for $w_{TM} = 0.075$. Vice versa, when keeping the translation model weight fixed to zero, a minimal WER of 10.03% was accomplished for the parameter setting $bd = 18$, $fp' = -10$, $lp' = 5$, $n = 20$ and all other parameters zero. This shows that, although $w_{TM} = 0.075$ is comparatively low, primarily the discriminative capabilities of the TM lead to a further reduction in WER. In this context it is interesting to note, that the MT bi-gram discount is now more heavily weighted. This can be explained by the fact that TM score and MT n-gram discounts only are different representations of MT knowledge: the TM score is an implicit part of the MT n-grams as it decides out of which words an MT n-gram is formed. Furthermore, the TM score has a great impact on the ranking of the n-best hypotheses within the MT n-best lists. Therefore, by using only the bi-grams of the 20-best MT hypotheses, the TM score is again implicitly applied. The now more heavily weighted MT bi-grams are only in part responsible for the decrease of the WER from 10.41% for the cache LM to now 10.03%. Two more factors influence this decline in WER: the positive effect of enriching the ASR 150-best list with the first best MT hypothesis and the rescoring due to changing the ASR system parameters fp' and lp'. When keeping all rescoring parameters fixed to zero, which is equivalent to only using the ASR score, the WER decreases to 10.34%, which is due to the additionally used first best MT hypotheses. When only using fp' and lp' for rescoring the WER further decreases to 10.30%. This shows that rescoring in regard to the ASR system parameters fp' and lp' yields only small gains in recognition accuracy. This was to be expected, as these ASR system parameters were already tuned for the baseline ASR system and re-adjusted for the cache LM system. Finally it should be noted that only a minimal decline in WER to 10.28% can be accomplished by only adjusting the language model weight w_{LM}. This is due to the fact that the ASR score already includes the LM score and that the language model weight lz was already tuned for the baseline

ASR system. An overview on the mentioned results and parameter settings can be found in Table 3.3. In summary it has to be noted that the additionally computed TM score has the greatest impact on the success of the subsequent rescoring of cache LM system output.

3.6.3 Hypothesis Selection on Cache + Interpolated LM System Output

When performing the hypothesis selection on the cache and interpolated LM system output, a WER of 9.7% could be achieved for $w_{TM} = 0.12$, $w_{LM} = 0.15$, $fp' = -10$, $lp' = 5$, $n = 20$, $sd = 2.5$ and all other parameters zero. This WER is higher than the WER for hypothesis selection on the cache LM output. A more elaborate attempt to explain this fact will be given in the following summarization (section 3.7). At the moment it should be only noted that the difference in WER compared to rescoring on cache LM system output is statistically insignificant.

For testing statistical significance a sentence based T test against 5% is used throughout this work.

3.7 Summarization

The LM interpolation approach uses MT context information in the form of tri-grams (and bi- and mono-grams for back off). The small gain in WER, compared to the rescoring and cache LM approach, can be explained by the small value of the MT context information for ASR improvement.
Two forms of MT knowledge are very successfully applied by the hypothesis selection approach:

- MT mono-grams: the MT acts as a provider of a ”bag of words,” thereby stating these words as likely to be seen in the translation of the human translator. However, no information on the translation probability of the individual words is given.

- TM scores: the TM scores constitute the word order independent sentence translation probability.

In addition to that, it is possible to incorporate MT context information in the form of bi-, tri-gram and sentence discounts. However, for rescoring the baseline ASR output no gains in performance could be observed in doing so. For higher bi-, tri-gram and sentence discounts, a rapid deterioration in recognition accuracy could be observed. This again proofs the small value of the MT context information for ASR improvement. The great advantage of the rescoring approach only to operate on the ASR output without changing the ASR process in itself, is also its most apparent disadvantage: the success of the approach stands and falls with the quality of the ASR n-best lists.

The cache LM approach inherits the way the "bag of words"-knowledge is used from the rescoring approach. In doing so, it is not only capable of providing similarly good (even slightly better, although statistically not significant) results, but it also produces ASR n-best list with a lower n-best WER (and a lower average WER). These n-best lists, therefore, offer once again a promising basis for hypothesis selection by rescoring with its ability to easily apply the above mentioned additional forms of MT knowledge. In fact, hypothesis selection on cache LM n-best lists yields the best results with a WER of 9.4%. This is equivalent to a BLEU score of 86.8 and a NIST score of 11.1 on the used data set.
No absolutely satisfying explanation could be found for why rescoring of cache + interpolated LM output does not provide the same or even slightly better results as rescoring on cache LM output. Considering this discrepancy in performance it has at first to be noted that the observed difference in WER of about 0.4 absolute is statistically not significant on this data set. However, one possible explanation goes as follows; the LM interpolation weight was chosen in regard of the WER produced by the combination of cache and interpolation scheme and not in regard of the WER produced by an additional rescoring. As already stated, it is not desirable to make overly use of MT context information, which is of course inherent to the interpolated LM in the form of tri- and bi-grams. The damaging influence of this MT context information becomes apparent in the additional rescoring. For the successful combination of cache LM and interpolated LM, one can argue that the TM score, which is implicitly given in the MT n-best lists by the positioning of the individual hypotheses, is to be credited. ASR hypotheses equal to the n-best MT hypotheses are favored by the interpolated LM. This once again shows another aspect of why the additional rescoring may not be as successful; during rescoring, the TM score of the ASR hypotheses is considered along with their ASR and LM score. However, when using an interpolated LM, the ASR hypotheses equivalent or similar to the MT n-best hypotheses used for LM interpolation already have an implicit share of the TM score in their LM and ASR score and are potentially overly favored.

Table 3.4 gives an overview on the performance of the described basic MTE-ASR techniques.

Technique	WER	Relative Gain
Baseline ASR	12.6	0.0%
Vocabulary Restrictions	> 15.0	-19.0%
Dynamic LM Interpolation	13.2	-4.8%
LM Interpolation	11.6	8.0%
Hypothesis Selection (on Baseline)	10.5	16.9%
Cache LM	10.4	17.6%
Cache & Interpolated LM	10.1	20.0%
Hypothesis Selection on Cache & Interp. LM	9.7	23.0%
Hypothesis Selection on Cache LM	9.4	26.0%

Table 3.4. Comparison of basic MTE-ASR techniques.

Chapter 4

Document Driven Iterative MTE-ASR

In this chapter, it is at first examined which of the basic MTE-ASR systems introduced in chapter 3 are most suited for an integration into the document driven iterative MTE-ASR system design depicted in Figure 4.1. This system component selection is done in section 4.1 with the help of the data set used so far. As the iterative design is based on an additional improvement of the involved MT component, the examinations will also include different MT improvement techniques that will be introduced at the beginning of section 4.1. Based on the results of this system component selection, a final iterative system design will be derived and then re-investigated on a second data set.

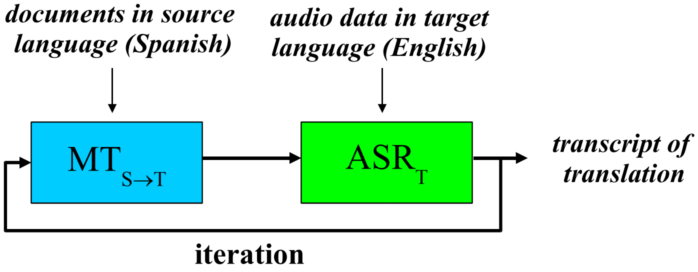

documents in source language (Spanish)
audio data in target language (English)
$MT_{S\rightarrow T}$
ASR_T
transcript of translation
iteration

Figure 4.1. Document driven iterative MTE-ASR.

4.1 System Component Selection

The scenario for which the basic MTE-ASR techniques of chapter 3 were developed is equivalent to iteration 0 of the document driven iterative system design. Therefore, it is convenient to start iteration 1 with the output provided by one of the described basic MTE-ASR systems. Hypothesis selection on cache LM yielded not only the best first best hypotheses, i.e. the best WERs, but also the most promising n-best lists in regard to nWER and average WER. For this reason, hypothesis selection on cache LM was greedily selected as vantage point for iteration 1.

The used data set (refer to chapter 3 for a closer description) was read four times. This means that, after iteration 0, there are four different ASR n-best lists containing English translation hypotheses for each Spanish source sentence. Using all of these four lists for the following iterations would change the iterative system into some sort of a voting system that choses between the n-best hypotheses provided by four ASR passes. For this reason, the data set was split into four disjoint subsets. Based on these four subsets, four different iterative MTE-ASR systems had to be examined. However, if not stated otherwise, only the average performance, calculated on the four individual system results, is presented in the following.

4.1.1 MT System Improvement

An important part of the iterative system design is the improvement of the MT system component with the help of the ASR output computed in the preceding iteration. Three approaches for MT improvement have been investigated, namely interpolating the MT target LM with a small ASR language model computed on the ASR n-best lists, retraining the MT system with the ASR n-best lists as additional training data and combining these two methods. An overview on the performance gains of the individual MT improvement techniques is given in Table 4.1. In the following, the three approaches will be shortly described. At the end of this section, the results will then be evaluated in the given MTE-ASR context, i.e. in the context of a further improvement of the English ASR system.

Language Model Interpolation

In a first experiment, the optimal settings for the ASR n-best list sizes and the interpolation weight of the small ASR language model were computed for each of the four systems by minimizing the perplexity on the complete English data set. For all four systems, the settings were $n = 10$ and w in the range of [0.915; 0.944]. The average performance was BLEU = 53.1.

In a second experiment, the average performance was computed for different combinations of ASR n-best list sizes and interpolation weights. The optimal settings in regard to BLEU score (as well as in regard to WER and nWER) were now $n = 3$ and $w = 0.8$ which yielded an average performance of BLEU

	BLEU	NIST	WER	nWER
Baseline MT	40.4	7.1	46.8	34.2
LM Interp	53.4	8.3	35.0	26.0
Updated Translation Memory				
- Retraining	70.2	9.9	21.4	7.0
- Combination	84.7	10.9	10.2	6.5
Fixed Translation Memory				
- Retraining	42.1	7.3	45.4	30.0
- Combination	54.2	8.4	34.8	25.8

Table 4.1. Comparison of MT improvement techniques.

= 53.4. Overall a similar MT performance (less than 4% relative deviation in BLEU and NIST score and less than 8% relative deviation in WER and nWER) could be observed for n-best lists of size $1 \leq n \leq 10$ and interpolation weights of $0.6 < w < 1.0$.

Retraining

For retraining, new IBM1 lexica (forward and backward lexicon) were computed. This was done by adding the ASR n-best lists together with their respective source sentence several (x) times to the original training data. Two sets of experiments were run: the first with the translation memory fixed to the original training data and the second with an updated translation memory. In both cases, it turned out that the parameter range yielding best performances was $1 \leq n \leq 5$, $1 \leq x \leq 4$. The best performance in regard to BLEU score (as well as WER and NIST score) was found for the parameters $n = 1$ and $x = 4$ (fixed and updated translation memory). The system with the fixed translation memory gave a BLEU score of 42.1. The system with the updated translation memory yielded BLEU score of 70.2.

Retraining Combined with LM Interpolation

The above described systems for LM interpolation and retraining were combined. The range for the parameter settings with the best performance was equal to the parameter ranges described for the individual systems. The best parameter setting was $n_{LM} = 1$, $i = 0.9$ for LM interpolation and $n_{RT} = 1$, $x = 1$ for retraining. Using a fixed translation memory, a BLEU score of 54.2 was computed. Updating the translation memory improved the performance to a BLEU score of 84.7.

Conclusions

The combined approach of language model interpolation and retraining provides the best results, both for keeping the translation memory fixed and for updating

the memory, whereas the influence of the retrained IBM1 lexica is only small compared to the interpolated language model and the updated translation memory. As the objective was to further improve the recognition accuracy of the English ASR system by an improvement of the MT performance it seems appropriate to chose the MT improvement technique yielding the best MT performance, namely the combination of LM interpolation and retraining with an updated translation memory. However, the basic idea behind MTE-ASR was that it should be possible to improve the ASR system with the additional complementary knowledge provided by the MT system. But when using an updated translation memory one can argue that the complementary information given in the MT n-best lists is being strongly minimized by updating the translation memory. The updated memory sees to it that the ASR n-best hypotheses added to the training data are part of the newly created MT n-best lists. Moreover, if only the added ASR hypotheses are present as translation examples, and if $n_{MT} \leq n_{ASR}$, then we can speak of a simple rescoring of the ASR hypotheses by the translation model and the language model when using an updated translation memory. In the context of our iterative system design, which is aimed at a further improvement of the ASR with additional complementary MT knowledge, it is, therefore, possible that updating the translation memory is more damaging than helping. To prove or disprove this theoretical consideration, the combination of LM interpolation and retraining with a fixed/updated translation memory was considered for MT improvement. As we will see in section 4.2.2, it is in fact more effective to keep the translation memory fixed for further improvement of ASR recognition accuracy.
A more sophisticated approach than just not to update the translation memory would be a "cautious" updating of the memory with the help of a reliable confidence measure. Given such a confidence measure, it would be possible to update the translation memory only with ASR translation hypotheses that are most likely correct. That way, it should be possible to further improve the MT component without losing valuable MT knowledge. This approach was not examined in this work due to a lack of time.

As so mentioned above, it was necessary to split the data into four disjoint subsets as not to make use of the additional information provided by the fact that the data set was read four times. In realistic application scenarios, it is in fact highly unlikely to have the audio stream of several translators at hand that are translating into the **same** target language. Nevertheless, this scenario was examined for MT system improvement. When using all available ASR n-best hypotheses of the effective four ASR passes along with an updated translation memory, a BLEU score of 90.0 could be accomplished (the NIST score was 11.4, the WER was 5.8% and the nWER was 2.1%). This shows the high ability of the translation model to function as a voting mechanism in the case of multiple translation hypotheses provided by automatic speech recognition on multiple target audio streams.

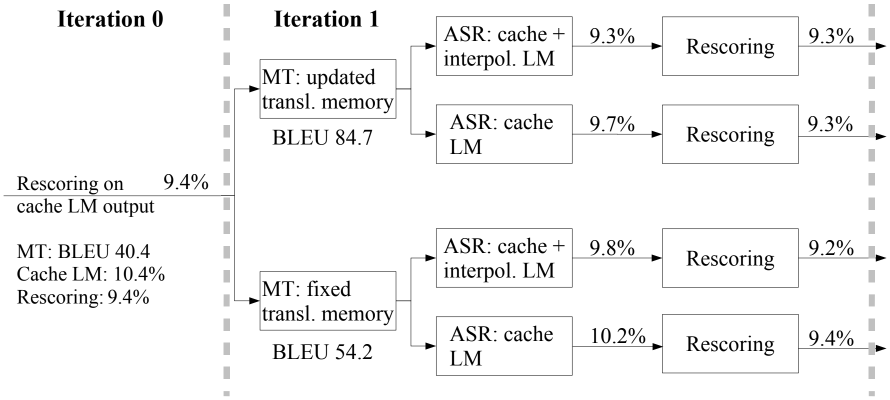


Figure 4.2. Iteration 1: Examined System Component Combinations & Respective WERs

4.1.2 Iteration Results

Based on the insights gained so far, the combined MT improvement technique with a fixed or updated translation memory and the ASR improvement techniques ”rescoring on cache LM system output” and ”rescoring on cache + interpolated LM system output” seem to be most promising for the following iterations. For iteration 1, the resulting four combinations together with their respective WERs are shown in Figure 4.2. No significant word error rate reduction, compared to iteration 0, could be observed. The same was true for iteration 2; therefore, no further iterations have been carried out.
The parameter settings[1] used for iteration 1, again found by manual gradient descent, are shown in Table 4.2. The better performance of the MT system with the updated translation memory is reflected in the smaller MT n-best list sizes n (n_c, n_i), the slightly higher probability discounts d and the higher LM interpolation weight w. The smaller amount of MT knowledge applied in the case of the cache LM system without LM interpolation is being compensated by higher TM weights w_{TM} and higher MT n-gram discounts (bi-gram discount bd and sentence discount sd).

4.1.3 Conclusions

The difference in word error rate for the examined component combinations was to small to a allow a justified decision for one of these combinations. Moreover, no significant reduction of WER was seen for applying the iterative scheme. One possible explanation for both observations could be the fact that the complete data set was used for system parameter tuning. Especially when looking at the relatively high number of parameters used for rescoring on cache LM output,

[1] Although there were in fact four separate systems, one per data subset, the same settings were used for all four systems.

	Updated Transl. Memory	Fixed Transl. Memory
Cache LM	$n = 1$, $d = 1.5$	$n = 20$, $d = 1.3$
Rescoring	$w_{TM} = 0.225$, $w_{LM} = 0.1$, $fp' = -20$, $lp' = 5$, $n = 20$, $bd = 2$, $sd = 6$	$w_{TM} = 0.175$, $w_{LM} = 0.1$, $fp' = -17.5$, $lp' = 10$, $n = 20$
Cache + Interpol. LM	$n_c = 1$, $d = 1.4$, $n_i = 5$, $w = 0.1$	$n_c = 20$, $d = 1.3$, $n_i = 10$, $w = 0.05$
Rescoring	$w_{TM} = 0.125$, $w_{LM} = 0.15$, $fp' = -35$, $n = 20$	$w_{TM} = 0.15$, $w_{LM} = 0.1$, $fp' = -35$, $n = 20$

Table 4.2. Parameter settings for iteration 1. Unlisted parameters were set to zero.

it is questionable if the same very good performance can be accomplished on unseen data not used for parameter tuning. One could, therefore, argue that the possibly unrealistically good rescoring performance excels potentially given positive iteration effects as well as differences in the examined component combinations. In this context, it should be noted that the slightly more visible differences in WER for the ASR output in iteration 1 become clearly smaller after rescoring.

Another possible reason for the failure of the iterative approach could be the very good match of the used data set and the baseline language model. The perplexity of the LM on the data set was very low (21.60). Therefore, room for further improvements by applying word context knowledge provided by the improved MT system is relatively small.

4.2 Final System

4.2.1 Experimental Setup

Final System Design

The results gained so far for the different system component combinations introduced in 4.1 do not allow a justified decision for one of these combinations. For this reason, all of these combinations will be re-investigated.

Data

The second data set consists of 500 English and Spanish sentences in form and content close to the BTEC. The English sentences were read 4 times, each time by 5 different speakers with 10 speakers overall. The data was split into four parts so that each sentence occurred just once per subset. Overall, there were four MTE-ASR systems, one per subset. One tenth of each subset was randomly selected as held-out data for tuning the parameters of the respective MTE-ASR system. The final performance was measured over the complete output of all four systems. Because of some flawed recordings, the reduced data set consisted

	WER	nWER	BLEU	NIST
Baseline ASR	22.3	10.3	68.0	9.6
Baseline MT	50.1	34.5	32.5	6.9

Table 4.3. Performance of baseline components on data set II.

only of 1,747 sentences composed of 13,398 (959 different) words. The audio data equals 68 min. The best possible NIST score on this data set is 12.3.

Baseline Components

The same baseline systems (ASR and MT) were used as for the experiments on the first data set (refer to 3.1 for a closer description). The OOV rate of the ASR system on the second data set was now 0.51%. The perplexity of the language model used by both baseline systems was now 85.2 on the new data set and, thereby, approximately four times higher than on the first data set. Table 4.3 gives an overview on the baseline performance.

4.2.2 Iteration Results

The system component combinations for iteration 1, introduced in section 4.1.2, were based on the use of the cache LM system without language model interpolation in iteration 0. With the given higher perplexity of the baseline LM on data set II, the question arises if it is still reasonable to forgo language model interpolation in iteration 0 as it was done on data set I. It turned out that similar results could be observed in iteration 0 on data set II. The combination of cache LM and interpolated LM yielded a better word error rate than the cache LM system alone; however, rescoring on cache LM system output finally led to the best WER:

	WER	nWER	BLEU	NIST
Cache LM	18.2	7.5	72.6	10.0
Rescoring	15.5	7.5	76.5	10.4
Cache + Interpol. LM	16.9	8.0	74.3	10.2
Rescoring	15.9	8.0	76.3	10.4

Table 4.4. Results for Iteration 0 on data set II.

Therefore, the same component combinations were taken into consideration as before. Figure 4.3 shows the respective word error rates on data set II for iteration 1. No noteworthy changes in word error rate could be observed for iterations > 1. In general, better ASR results can be gained when working with a fixed translation memory. The loss of MT knowledge when updating the translation memory, which was already mentioned in 4.1.1, becomes not only evident in the first best word error rates of the respective systems, but also

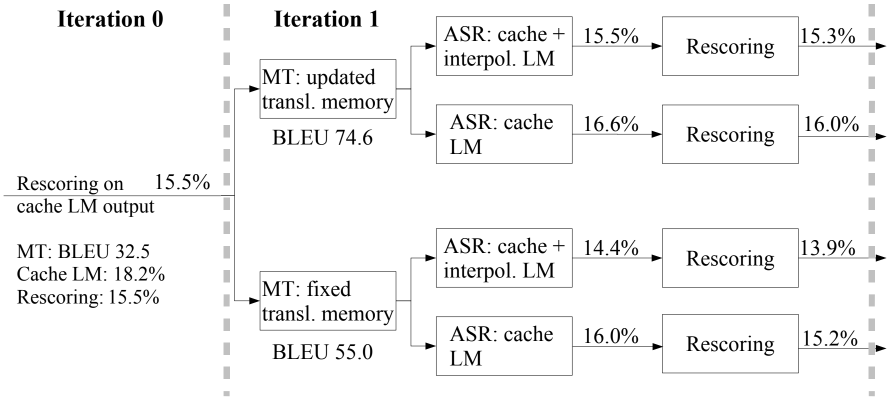


Figure 4.3. WERs of different component combinations on data set II.

in their n-best WERs. An updated translation memory forces the MTE-ASR systems in iteration 1 towards the n-best ASR hypotheses of iteration 0 which were used for updating the memory. Therefore, the nWERs for the systems based on an updated translation memory increase significantly (approaching the first best WERs), while the nWERs remain constant for the fixed translation memory systems. This nWER development is depicted in Figure 4.4 for the cache + interpolated LM systems.
The reasons for the better performances of the cache + interpolated LM systems compared to the cache LM systems can be found in the improved MT context information as well as in the higher mismatch between baseline language model and data set II. Based on its superior performance, the combination of fixed translation memory and cache + interpolated LM was picked as final document driven iterative MTE-ASR system. This final system had a WER of 13.9%, a nWER of 7.6%, a BLEU score of 78.6 and a NIST score of 10.6. A summarizing overview on the performance of the final system components is shown in Figure 4.5.

It should be kept in mind that the data was split into four parts so as not to make use of additional information provided by the fact the data was read four times overall. This means there were in fact four final systems, one subsystem per subset. The used parameter settings[2] for each subsystem were again found by manual gradient descent, but now on the 10% held-out data randomly chosen from each of the four data subsets, i.e. parameter tuning was done for each of the four subsystems separately. For this reason, there were always only up to fifty sentences used for parameter tuning. Nevertheless, the found parameter settings always yielded a good performance. This fact may be surprising, especially when looking at the relatively high number of parameters

[2] The parameter settings for the final subsystems can be found in appendix C.

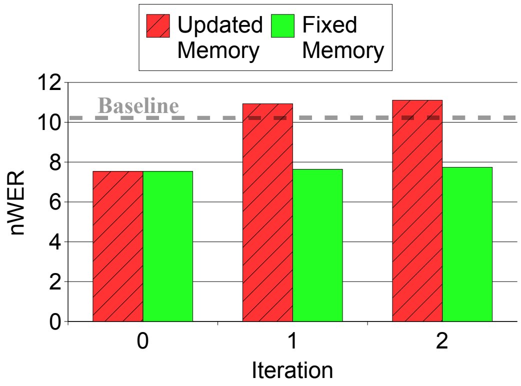
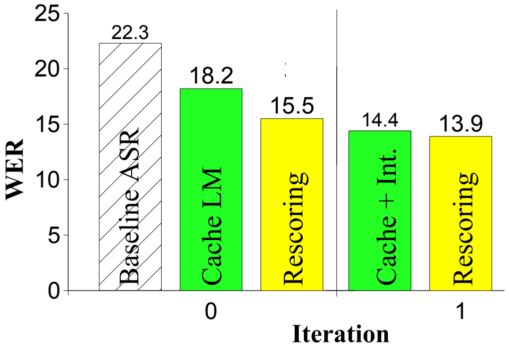


Figure 4.4. Development of ASR nWERs on data set II.



MT performance:

	NIST	BLEU
Iteration 0	6.9	32.5
Iteration 1	8.9	55.0

Figure 4.5. Final document driven iterative MTE-ASR system - results for data set II.

used for rescoring. At first, it has to be noted that all parameter settings were always searched within the ranges that turned out to be useful in the experiments done on data set I. Moreover, it has to be mentioned that the main focus for rescoring parameter tuning was on the translation model weight, as this parameter turned out to be the most important parameter when applying rescoring on output provided by an ASR system using the cache LM scheme. This may be, in part, explained by the fact that the tuning of the language model weight, the word penalty and the filler word penalty were also taken into consideration when tuning the cache LM together with the interpolated LM parameters. As for the rescoring parameters apart from the translation model weight, these were only changed from zero (zero means no rescoring in respect to this parameter) if high differences in WER could be observed and if the changes seemed ”plausible” (whereas ”plausibility” was up to the authors,

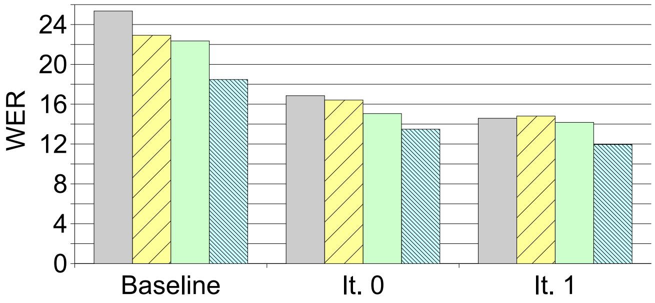


Figure 4.6. Development of WERs for the four subsystems.

certainly subjective, consideration).

A significant difference in WER could be observed for the four iterative MTE-ASR subsystems on their respective data subset. The best baseline WER was 18.5%, the worst was 25.4%. This very high difference of 11.9% absolute is to be explained by the different speakers. The data subset of the subsystem with the lower WER happened to be read only by speakers with a relatively good articulation. It could be observed that the subsystem suffering from a bad articulation profited the most from the additional knowledge provided by the MT. Its relative gain in WER was 42.4% after iteration 1, compared to a relative gain of 35.3% for the other subsystem. The maximal absolute difference in WER between the four subsystem was now only 3.4%. Figure 4.6 shows the the WERs of the four subsystem up to iteration 1.

4.2.3 Conclusion

Even though a very high relative gain of 30.6% in WER compared to the baseline ASR system could be accomplished for the non-iterative approach (iteration 0), the relative gain could be further increased to 37.7% for the iterative approach (iteration 1). This shows that the iterative approach could be successfully applied in the document driven case to further increase the recognition accuracy.

Chapter 5

ASR Driven Iterative MTE-ASR

This chapter is structured in the same manner as the chapter for the document driven case. At first, the most promising system component combination will be selected for the ASR driven iterative MTE-ASR depicted in Figure 5.1. This is done in section 5.1 with the help of a first data set. The resulting final system is then re-investigated in section 5.2 using a second data set.

5.1 System Component Selection

5.1.1 Experimental Setup

Data

The data set used for these experiments corresponds to data set I of the document driven case, i.e. the same 506 parallel Spanish and English sentences were used. The data was now read only two times, each time by three Span-

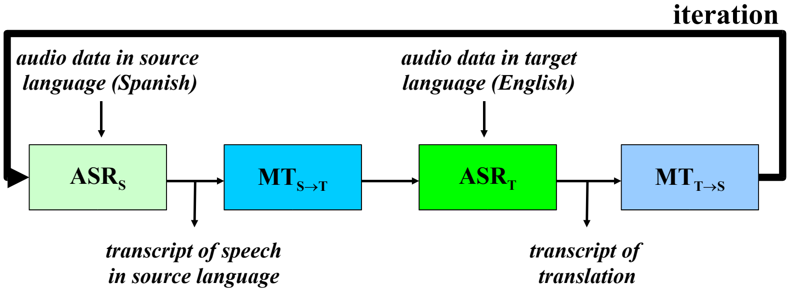

iteration
audio data in source language (Spanish)
audio data in target language (English)
$\mathbf{ASR_S}$
$\mathbf{MT_{S\rightarrow T}}$
$\mathbf{ASR_T}$
$\mathbf{MT_{T\rightarrow S}}$
transcript of speech in source language
transcript of translation

Figure 5.1. ASR driven iterative MTE-ASR.

	WER	nWER	OOV	Perplexity
English Baseline ASR	13.5	7.4	0.56%	21.9
Spanish Baseline ASR	15.1	8.4	3.20%	75.5

Table 5.1. Performance characteristics of the baseline ASR systems on data set I.

ish[1] and three English speakers. As a consequence, the data had to be split in two separate parts, and all experiments were run on two separate MTE-ASR systems. The performance values are once again computed on the complete output of both subsystems. Ten percent of the data was randomly selected as held-out data for parameter tuning of the individual subsystems. Because of some flawed recordings, the reduced Spanish data consisted of 900 sentences composed of 5,398 (1,021 different) words. The respective English data consisted of 898 sentences with 5,333 (786 different) words. The Spanish audio data equals 36 minutes, the English 32 minutes. The best possible NIST score on this data set is 12.1 for the translation direction English to Spanish and 12.2 for the translation direction Spanish to English.

Baseline ASR Systems

The same English baseline ASR system was used as in the experiments for the document driven case. It had a WER of 13.5%. The difference in WER compared to the English baseline ASR system for the document driven case, which had a WER of 12.6%, can be explained by the fact that data set I was now read only twice and that the parameter optimization was now done on a 10% held-out data set. Table 5.1 gives an overview on performance as well as OOV rate and baseline language model perplexity for the English and Spanish baseline ASR systems. The Spanish ASR system is once again based on the Janus Recognition Toolkit (JRTk) with its IBIS single pass decoder [10]. The sub-phonetically tied three-state HMM based recognition system has 2 K codebooks and 8 K distributions. All other basic characteristics are equivalent to characteristics of the English recognizer. The ASR system was trained on South American Spanish as well as Castilian Spanish, namely on 112 h South American speech data (mainly Mexican and Costa Rican dialects) and 14 h Castilian Spanish speech data. The South American corpus was composed of 70 h Broadcast News data, 30 h Globalphone data and 12 h Spanish Spontaneous Scheduling Task (SSST) data. It gave a WER of 15.1%; this higher WER compared to the WER of the English recognizer can be explained by the approximately four times higher perplexity of the Spanish language model. The higher perplexity of the Spanish LM is due to the fact that Spanish is a morphological more complex language than English.

[1] The Spanish speakers mostly had a Castilian Spanish accent. A few speakers had a South American accent

Input provided by	BLEU	NIST	WER	nWER
Spanish Transcripts	40.8	7.0	47.2	31.1
Spanish Baseline ASR	39.0	6.7	51.1	35.2
English Transcripts	34.9	6.2	56.3	38.3
English Baseline ASR	31.6	5.7	61.1	43.9

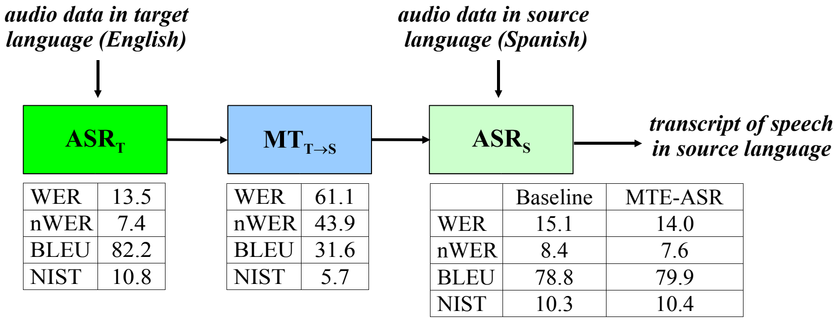
Table 5.2. Performance of baseline MT systems on data set I.

Baseline MT Systems

The same Spanish to English statistical machine translation system was used as before. The English to Spanish machine translation system is equivalent to the Spanish to English system, only that the translation direction was inverted during training. The language model was again the same as the language model of the baseline ASR system. Table 5.2 gives an overview on the performance of the MT systems when using the transcripts as input and when using the first best ASR hypotheses as input. The BLEU score of the Spanish to English MT system decreases from 40.8 to 39.0 when using the first best ASR hypotheses as input instead of the Spansih transcripts. The BLEU score of the English to Spanish MT system decreases from 34.9 to 31.6 when using the first best ASR hypotheses as input instead of the Spansih transcripts.

5.1.2 Baseline MTE-ASR Systems

The ASR driven iterative system design provides not only transcription hypotheses for the target language (English) translation but also transcription hypotheses for the source language (Spanish) speech. The iterative design automatically combines the improvement of the source language ASR and the target language ASR. In particular, it would have been possible to start the iteration cycle with improving the Spanish ASR with knowledge gained by automatically translating the hypotheses of the English baseline system first. Depending on the performance of the respective baseline ASR systems this may be desirable. This work only concentrates on the case where the target ASR system is improved first within the iteration cycle. As a consequence, the first improvement of the source ASR system is done with the help of the already improved target ASR system. For an accurate comparison of the iterative approach with the non-iterative MTE-ASR approach in regard to the improvement of the source language side ASR system, it is, therefore, necessary to consider a separate non-iterative source language side MTE-ASR system. The non-iterative target language side MTE-ASR system is implicitly given in iteration 0 of the iterative system design.
Figure 5.2 shows the results for the best non-iterative MTE-ASR approach on the source language side (Spanish). Once again, it was better to use the combination of rescoring on cache LM output instead of rescoring on cache + interpolated LM output. Comparing these results with the non-iterative improvement

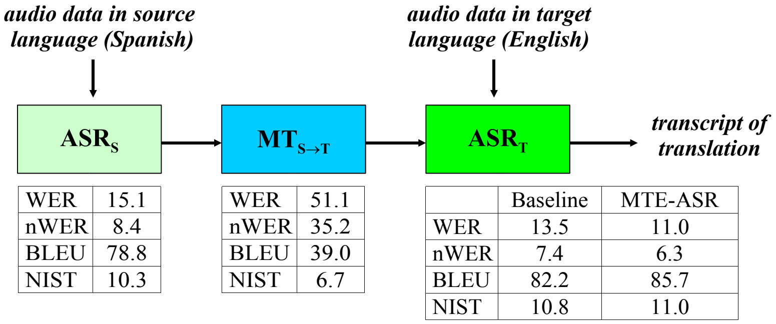

audio data in target language (English)
audio data in source language (Spanish)
ASR$_T$
MT$_{T\to S}$
ASR$_S$
transcript of speech in source language

WER	13.5
nWER	7.4
BLEU	82.2
NIST	10.8

WER	61.1
nWER	43.9
BLEU	31.6
NIST	5.7

	Baseline	MTE-ASR
WER	15.1	14.0
nWER	8.4	7.6
BLEU	78.8	79.9
NIST	10.3	10.4

Figure 5.2. Source side baseline MTE-ASR: Results on data set I.


audio data in source language (Spanish)
audio data in target language (English)
ASR$_S$
MT$_{S\to T}$
ASR$_T$
transcript of translation

WER	15.1
nWER	8.4
BLEU	78.8
NIST	10.3

WER	51.1
nWER	35.2
BLEU	39.0
NIST	6.7

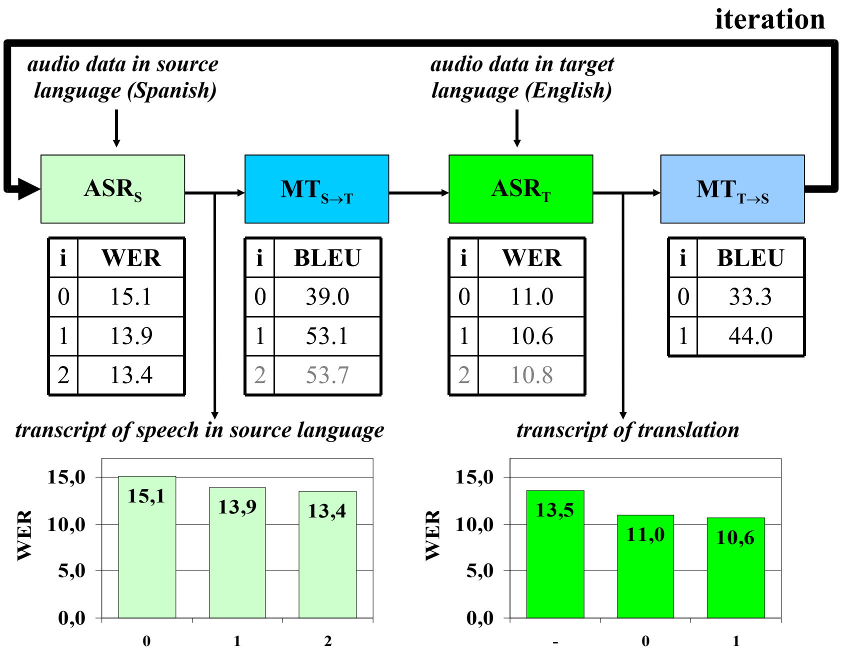
	Baseline	MTE-ASR
WER	13.5	11.0
nWER	7.4	6.3
BLEU	82.2	85.7
NIST	10.8	11.0

Figure 5.3. Target side baseline MTE-ASR: Results on data set I.

of the target language side (English) it becomes apparent that the decrease in WER for the English ASR system is higher than for the Spanish ASR system. This can be explained by the fact that Spanish is a morphological more complicated language than English, which is also the reason for why the on the test data set computed perplexity of the Spanish BTEC language model is approximately 3.5 times higher than the perplexity of the English BTEC language model.

5.1.3 Iteration Results

Figure 5.3 shows the results for iteration 0. For iterations > 0, only the combined MT improvement technique with a fixed translation memory was taken into consideration based on the results for the document driven case. For ASR improvement, rescoring on cache LM output and rescoring on cache + interpo-


iteration
audio data in source language (Spanish)
audio data in target language (English)
ASR_S
$MT_{S\rightarrow T}$
ASR_T
$MT_{T\rightarrow S}$
transcript of speech in source language
transcript of translation
WER
15,0
10,0
5,0
0,0
15,1
13,9
13,4
0
1
2
WER
15,0
10,0
5,0
0,0
13,5
11,0
10,6
-
0
1

i	WER
0	15.1
1	13.9
2	13.4

i	BLEU
0	39.0
1	53.1
2	53.7

i	WER
0	11.0
1	10.6
2	10.8

i	BLEU
0	33.3
1	44.0

Figure 5.4. ASR driven iterative MTE-ASR: Results on data set I.

lated LM output were examined. The additional use of an interpolated language model for the English ASR resulted in a slightly worse WER (the difference was statistically insignificant). This was true for all examined iterations (0-2) and can be explained by the already very good match of the English baseline LM with the used data set (the perplexity was only 21.9). For the Spanish ASR system a small gain in WER (again statistically insignificant) could be accomplished when using an interpolated language model based on the output of the improved English to Spanish translation component, i.e. when applying an interpolated LM in iteration 2. This small gain can be explained by the higher mismatch between the Spanish baseline LM and the given data (the perplexity was 75.5). The fact that the gain was only minimal may be due to the still relatively moderate performance of the improved Spanish MT component. Overall, no significant changes in performance could be observed for iteration 2 compared to iteration 1; therefore, no further iterations have been carried out. Figure 5.4 gives a summarizing overview on the performance of the best found system component combination on data set I.

5.1.4 Conclusion

In the context of a subsequent rescoring, it seems that the use of an interpolated language model in addition to the cache LM scheme can only be helpful if the data provided for interpolation came from an already improved MT component. Even if based on an improved MT component, gains in WER may only be expected if a certain mismatch between baseline language model and data is given. Furthermore, no significant gains in recognition accuracy are to be expected by recursively applying knowledge provided by the improved MT components. This means, improving the involved MT systems once is sufficient. As a consequence, the iteration should be aborted before an involved MT component would be improved a second time, namely during iteration 2. Since we started the iterative process by improving the target side ASR, we should, therefore, abort the iterative process after rescoring the source side ASR output in iteration 2.

5.2 Final System

5.2.1 Experimental Setup

Final System Design

The final ASR driven system design is shown in Figure 5.5. Based on the results for the document driven case and the results for the ASR driven designs examined so far, language model interpolation for the involved ASR components is only applied after improvement of their respective MT component. The iterative process is aborted in iteration 2 so that no involved MT component is improved twice.

Data

The data set used for re-investigating the final system design corresponds to data set II of the document driven case, i.e. the same 500 parallel Spanish and English sentences were used. The data was read two times, each time by three Spanish and five English speakers. As a consequence, the data was split in two separate parts, and all experiments were run on two separate MTE-ASR systems. As before, the performance values are computed on the complete output of both subsystems. Ten percent of the data was randomly selected as held-out data for parameter tuning of the individual subsystems. Because of some flawed recordings, the reduced Spanish data set has 904 sentences composed of 6,395 (1,089 different) words. The respective English data set has 880 sentences with 6,751 (946 different) words. The Spanish audio data equals 45 minutes, the English 33 minutes. The best possible NIST score on this data set is 12.2 for the translation direction English to Spanish and 12.4 for the translation direction Spanish to English.

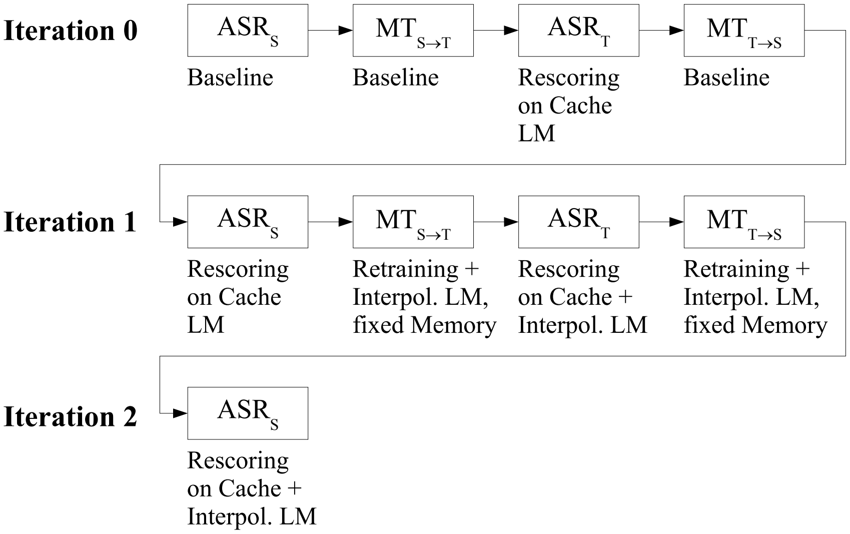


Figure 5.5. Final ASR driven iterative system design.

	WER	nWER	OOV	Perplexity
English Baseline ASR	20.4	9.0	0.53%	86.0
Spanish Baseline ASR	17.2	8.9	2.04%	130.2

Table 5.3. Performance characteristics of the baseline ASR systems on data set II.

Baseline System Components

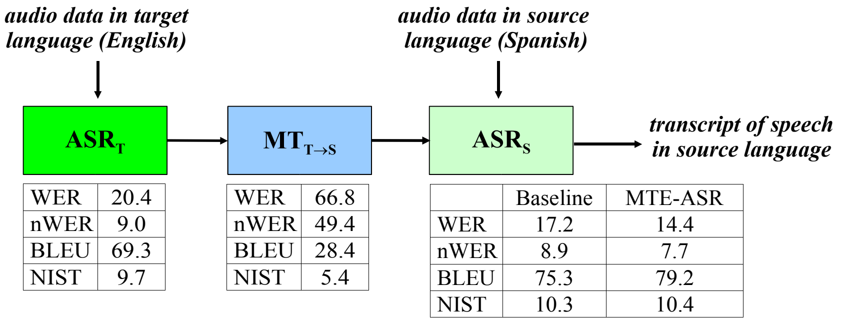
The same baseline ASR and baseline MT systems were used as before. Table 5.3 gives an overview on performance as well as OOV rate and baseline language model perplexity for both ASR systems. The performance for the baseline MT systems can be found in the following description of the baseline MTE-ASR systems.

5.2.2 Baseline MTE-ASR Systems

Figure 5.6 shows the non-iterative source side MTE-ASR system performance. The non-iterative target side MTE-ASR system (refer to Figure 5.7) is once again equivalent to iteration 0 of the iterative system design.

5.2.3 Iteration Results

A summarizing overview on the performance of the final ASR driven iterative MTE-ASR system is shown in Figure 5.8. The final target side output had a

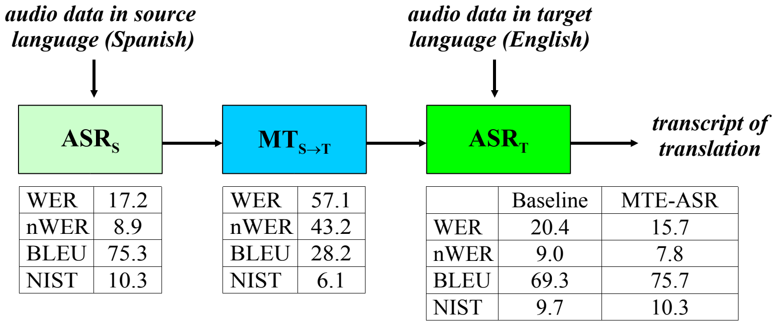

audio data in target language (English)
audio data in source language (Spanish)
ASR_T
$MT_{T \to S}$
ASR_S
transcript of speech in source language

WER	20.4
nWER	9.0
BLEU	69.3
NIST	9.7

WER	66.8
nWER	49.4
BLEU	28.4
NIST	5.4

	Baseline	MTE-ASR
WER	17.2	14.4
nWER	8.9	7.7
BLEU	75.3	79.2
NIST	10.3	10.4

Figure 5.6. Source side baseline MTE-ASR: Results on data set II.


audio data in source language (Spanish)
audio data in target language (English)
ASR_S
$MT_{S \to T}$
ASR_T
transcript of translation

WER	17.2
nWER	8.9
BLEU	75.3
NIST	10.3

WER	57.1
nWER	43.2
BLEU	28.2
NIST	6.1

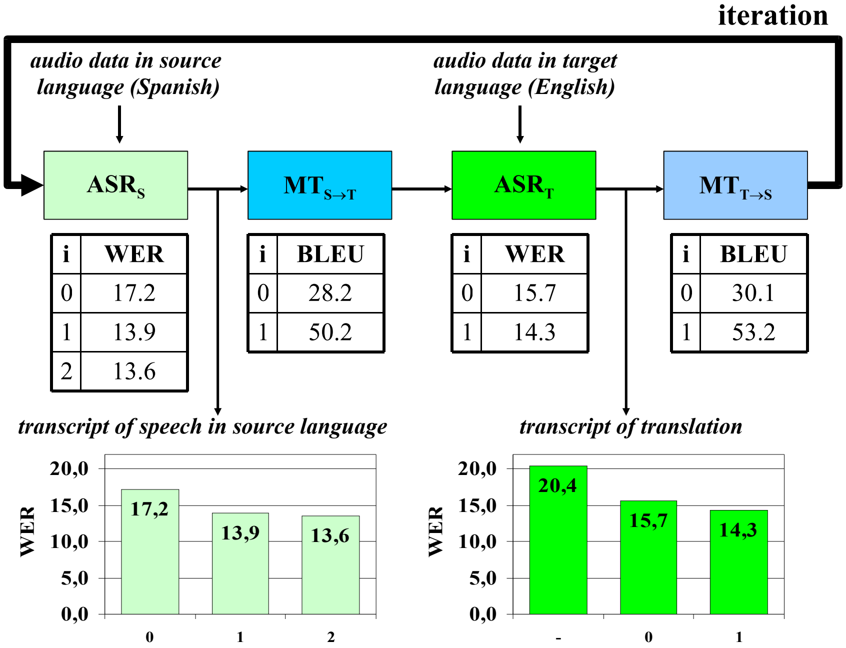
	Baseline	MTE-ASR
WER	20.4	15.7
nWER	9.0	7.8
BLEU	69.3	75.7
NIST	9.7	10.3

Figure 5.7. Target side baseline MTE-ASR: Results on data set II.

WER of 14.3% (and a nWER of 7.5%, a BLEU score of 77.7 and a NIST score of 10.5).

5.2.4 Conclusion

The non-iterative ASR driven MTE-ASR design yielded a relative gain of 23.2% in WER on the target language side (English) and a relative gain of 16.2% on the source language side (Spanish). This already relatively high gains could be further increased to 29.9% on the target side and to 21.3% on the source side by applying the iterative scheme. Similiar results have been gained for the document driven case in chapter 4. The iterative system design, therefore, constitutes a feasible and promising approach to Machine Translation Enhanced Automatic Speech Recognition.


iteration
audio data in source language (Spanish)
audio data in target language (English)
ASR$_S$
MT$_{S\to T}$
ASR$_T$
MT$_{T\to S}$
transcript of speech in source language
transcript of translation
WER
20,0
15,0
10,0
5,0
0,0
17,2
13,9
13,6
0
1
2
WER
20,0
15,0
10,0
5,0
0,0
20,4
15,7
14,3
-
0
1

i	WER
0	17.2
1	13.9
2	13.6

i	BLEU
0	28.2
1	50.2

i	WER
0	15.7
1	14.3

i	BLEU
0	30.1
1	53.2

Figure 5.8. ASR driven iterative MTE-ASR: Results on data set II.

Chapter 6

Conclusion

6.1 Summary

In this work, several approaches were examined for improving the ASR performance on the target language speech for human mediated translation scenarios by incorporating information which became available through automatically translating transcripts of the source language speech, hence the name Machine Translation Enhanced Automatic Speech Recognition (MTE-ASR). The source language transcripts were either given (document driven case) or had, at first, to be created on the source language speech with the help of a source side ASR system (ASR driven case).
Starting from the document driven case and based on ideas found in related work, several basic non-iterative MTE-ASR approaches were developed. The successful basic techniques were:

- Language model interpolation: interpolating the baseline language model with a small language model computed on the MT n-best lists.

- Applying a cache language model scheme: enhancing the language model probabilities of words found within the MT n-best lists.

- Selecting hypotheses from ASR n-best lists with the help of the available MT knowledge. The ASR n-best lists were enriched with the first best MT hypotheses and either provided by the baseline ASR system or by a, with one or both of the above mentioned techniques, improved ASR system.

The best results among these basic, non-iterative MTE-ASR techniques could be gained by hypothesis selection from n-best lists provided by an ASR system applying the cache LM scheme. This was true for the document driven case as well as the ASR driven case. For the document driven case, a relative gain of 30.6% in word error rate compared to the baseline system WER of 22.3% was accomplished on the used test data set (data set II). For the ASR driven case, a relative gain of 23.2% compared to the baseline system WER of

20.4% was accomplished.

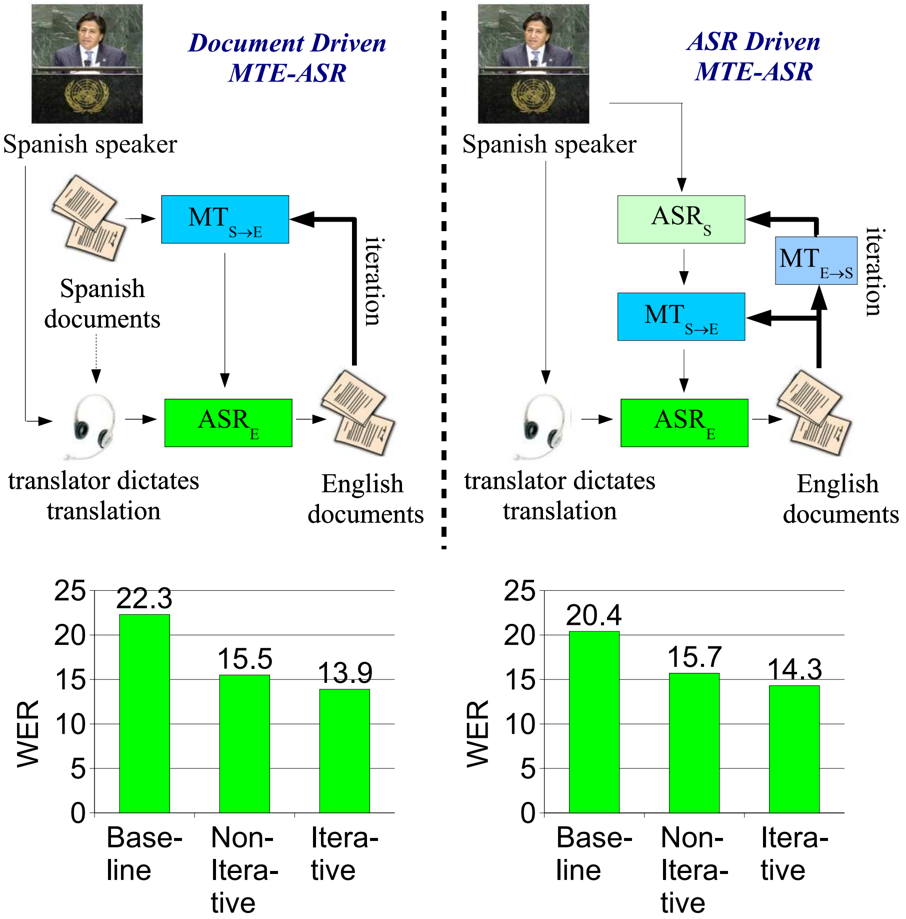


Figure 6.1. Results for improving the target language side ASR (English, data set II).

After developing the basic MTE-ASR techniques, their integration into an iterative system design was examined. The basic idea behind this iterative design was not only to make use of the available source language information for ASR enhancement, but to also make additional use of the available target language information for MT enhancement in the hope to further improve the speech recognition accuracy with the help of such an improved MT component. As a consequence of this examination, different MT improvement techniques had to be considered, namely retraining the MT system with the ASR translation hypotheses as additional training data and interpolating the MT target

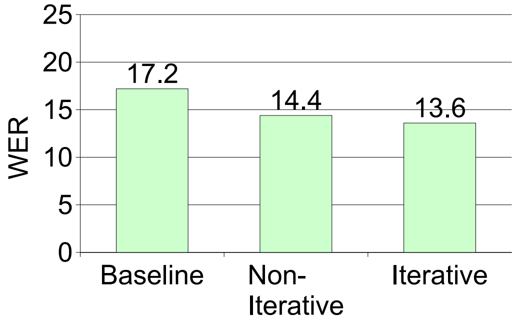


Figure 6.2. Results for improving the source language side ASR (Spanish, ASR driven case, data set II).

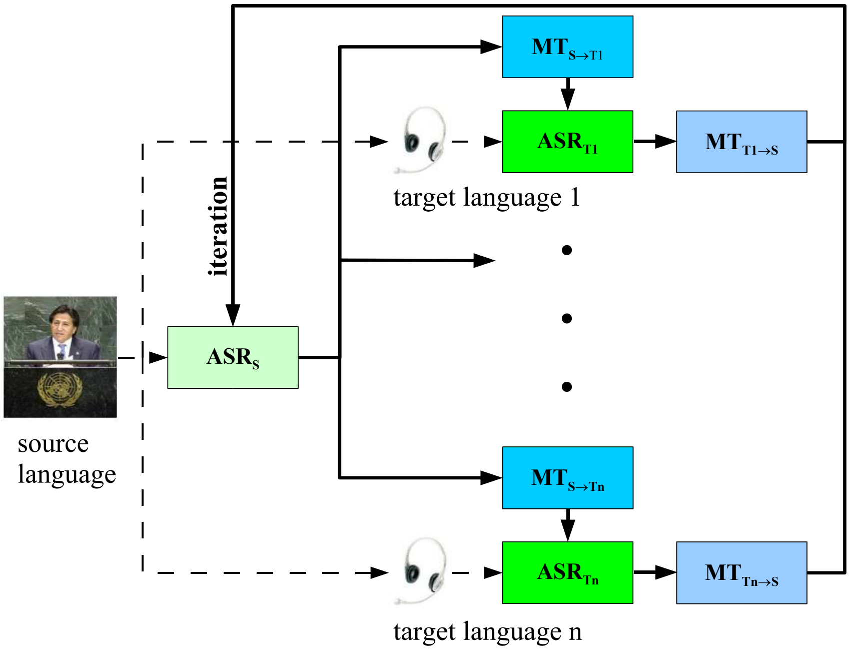
language model with a small language model computed on the ASR n-best lists. It turned out that combining those two techniques yielded the best results. However, in the context of further improving the speech recognition accuracy, it was necessary to constrain the retraining in a way that the translation memory component of the MT system was not updated, i.e. the translation memory was kept fixed to the original training data.
The best results within this iterative framework could be accomplished by integrating language model interpolation into the above described best basic MTE-ASR approach after an improved MT becomes available. Furthermore, it could be observed that improving the involved MT component(s) just once is sufficient. This means that the iterative process should be aborted right before an involved MT component would be improved a second time. Figure 6.1 gives a summarizing overview on the performance of the baseline ASR, the non-iterative MTE-ASR and the iterative MTE-ASR for the document driven case and the ASR driven case on data set II. Because the ASR driven iterative system design automatically combines the improvement of the source language ASR (in our case Spanish), an according overview is given in Figure 6.2. The results show that the examined non-iterative approaches and especially the iterative approach constitute a feasible and promising way for Machine Translation Enhanced Automatic Speech Recognition.

6.2 Future Work

Only the first best hypotheses of the source language ASR system were translated in the ASR driven case. One possible future development would therefore be to translate complete ASR lattices. Furthermore, system parameter tuning was done by manual gradient descent throughout this work. This should be automated. Another important issue is the use of an updated translation memory. Using a reliable confidence measure for updating the translation memory with only those ASR translation hypotheses that are most likely correct, it should be possible to further increase the MT performance without losing helpful complementary MT knowledge. Moreover, having a reliable confidence measure at hand, it can be hoped that the automatically generated translation and source speech transcripts can be successfully applied to create an improved MT component that will perform better on new, unseen data (of the same domain)[1]. An important next research step would be the testing of the applied MTE-ASR approaches on a more complex, and in regard to a possible tangible use case, more realistic data set. Bilingual data from European Parliament debates is being considered for this at the moment. Given more realistic data, different new use case specific problems will have to be addressed. For example, the assumption made so far that for every spoken target sentence the respective source sentence (audio) data is known and fully available will not be maintainable any longer. Furthermore, self-corrections of the human translator, as they are likely to be seen in the case of simultaneous translations, have to be considered.

A realistic application for the introduced iterative ASR driven MTE-ASR would be for example an offline working transcription system to assist the publication of European Parliament or United Nations speeches in different languages (including the source language). Looking at the fact that there are six official United Nations languages and twenty official European Parliament languages, the possible benefit becomes easily apparent. It has to be noted, especially, that the iterative approach directly allows an incorporating of knowledge provided not just by one additional audio stream in another language but by many. An according scenario that shows this is depicted in Figure 6.3. Further in the future an on-line system is imaginable for providing high quality transcripts in real time, to be used for example as closed captioning for TV broadcasts of debates.

[1] To what extend the MT improvement techniques used so far are suitable to positively influence the MT performance on unseen data has not been investigated within this work.



iteration
$\mathrm{MT}_{S\rightarrow T1}$
ASR_{T1}
$\mathrm{MT}_{T1\rightarrow S}$
target language 1
source
language
ASR_{S}
$\mathrm{MT}_{S\rightarrow Tn}$
ASR_{Tn}
$\mathrm{MT}_{Tn\rightarrow S}$
target language n

Figure 6.3. Iterative ASR driven MTE-ASR in the case of n target languages.

Appendix A

Additional Cache LM Experiments

A.1 Differentiated Increasing of LM Probabilities

The method applied in the cache LM experiments described in chapter 3 was quite simple: the LM probability of all MT mono-grams was increased by a constant value. A more sophisticated approach would be, for example, to increase the probabilities of words that occur very often in the respective n-best list by a greater value than the probabilities of words that occur less often.
At first the MT n-best lists were analyzed more closely, to see if there is a correlation between the amount of occurrence of a word in the n-best list and the "correctness" of that word, where a word is defined as correct if it is part of the English transcript of the respective sentence. For this the MT n-best list words were separated into four partitions:

- Partition I: all words that occurred in at least 66% of the translations found in the n-best list
- Partition II: all words that occurred in at least 33% of the translations found in the n-best list and not in partition I
- Partition III: all words that occurred in at least 10% of the translations found in the n-best list and not in partition I or II
- Partition IV: all words that occurred at least once and not in one of the other partitions

Table A.1 shows the number of words in the respective partition together with the rate of correct words in percent for different MT n-best list sizes.

After performing some first experiments, it became clear that increasing the LM probabilities for the words found in the translations by a factor greater

n	I	II	III	IV
10	2439 74%	651 38%	749 16%	634 9%
50	2187 76%	954 42%	1649 13%	2771 5%
100	2153 76%	1004 43%	1878 12%	4496 4%
150	2130 76%	1041 42%	2018 12%	5550 3%

Table A.1. Number of words and "word correct" rate for n-best list word partitions.

n	0	0.2	0.4	0.6	0.8	1.0	1.2	1.3
20	10.39	10.41	10.35	10.32	10.33	10.35	10.37	10.41
40	10.46	10.51	10.43	10.37	10.40	10.44	10.48	10.46
60	10.48	10.51	10.41	10.37	10.41	10.51	10.53	10.56
80	10.51	10.53	10.47	10.42	10.50	10.62	10.68	10.64

Table A.2. WERs for a differentiated increasing of word probabilities.

than 1.3 will inevitably lead to a decline in recognition accuracy, even if only the words found in partition I were increased by a greater value, but increasing the LM probabilities for the words found in partition IV by a smaller value than for the words found in the other partitions will lead to a small decrease in WER of the cache LM system. Table A.2 shows the WERs for systems where the probabilities of words found in partition IV were increased by different values in the range from 0 to 1.3. The probabilities for all other words found in the MT n-best lists were increased by the value 1.3. This approach was not further pursued as the observed gain in performance was only minimal.

A.2 Considering Synonyms

Another possibility to improve the original cache LM approach would be to not only increase the LM probabilities of the words found in the MT translations but also of all their synonyms. For this reason the vocabulary on the 20-best MT hypotheses was computed for all sentences and extended it by all synonyms found in the WORDNET database. Through this the vocabulary was increased by approximately 60% without increasing the coverage of the test set vocabulary at all. This approach was, therefore, not further pursued.

Appendix B

Document Driven MTE-ASR: Parameter Settings

	System I	System II	System III	System IV
ASR	lz=30, lp=-15 fp=30, n=10, d=1.2	lz=26, lp=5 fp=35, n=20, d=1.4	lz=32, lp=10 fp=30, n=30, d=1.3	lz=30, lp=-15 fp=5, n=30, d=1.2
Resc.	w_{TM}=0.25, lp'=20, fp'=-25, md=5	w_{TM}=0.15	w_{TM}=0.15, fp'=5	w_{TM}=0.25
MT	n_{LM}=1, i=0.9 n_{RT}=3, x=4	n_{LM}=1, i=0.8 n_{RT}=3, x=4	n_{LM}=5, i=0.9 n_{RT}=1, x=2	n_{LM}=1, i=0.9 n_{RT}=1, x=2
ASR	lz=32, lp=-5, fp=10, n=20, d=1.3, n_{LM}=5 i=0.05	lz=30, lp=-5, fp=15, n=20, d=1.3, n_{LM}=1, i=0.1	lz=32, lp=5, fp=30, n=10, d=1.3 n_{LM}=1, i=0.1	lz=30, lp=0, fp=5, n=20, d=1.3, n_{LM}=1 i=0.05
Resc.	w_{TM}=0.125	w_{TM}=0.125, w_{LM}=0.075	w_{TM} = 0.1, fp'=10	w_{TM}=0.175, w_{LM}=0.025

Table B.1. Parameter settings for the final document driven system on data set II. Unlisted parameters were set to zero.

Appendix C

ASR Driven MTE-ASR: Parameter Settings

	System I	System II
E. ASR	lz=30, lp=10, fp=30	lz=30, lp=-5, fp=40
MT	-	-

S. ASR	lz=30, lp=-5, fp=40
	n=20, d=0.8
		lz=28, lp=-5, fp=40
		n=10, d=0.5

Resc.	w_{TM}=0.075	w_{TM}=0.1

Table C.1. Parameter settings for the Spanish non iterative ASR driven system on data set I. Unlisted parameters were set to zero.

	System I	System II
E. ASR	lz=26, lp=0, fp=35	lz=30, lp=-15, fp=5
MT	-	-

S. ASR	lz=30, lp=5, fp=30
	n=30, d=0.6
		lz=32, lp=10, fp=35
		n=10, d=1.0

Resc.	w_{TM}=0.15	w_{TM}=0.125

Table C.2. Parameter settings for the Spanish non iterative ASR driven system on data set II. Unlisted parameters were set to zero.

	System I	System II
S ASR	lz=30, lp=-5, fp=40	lz=28, lp=-5, fp=40
MT S to E	-	-

E ASR	lz=30, lp=10, fp=40,
	n=20, d=1.4
		lz=28, lp=-5, fp=40,
		n=20, d=1.3

Resc.	w_{TM}=0.075, w_{LM}=0.175 fp'=-5	w_{TM}=0.075, w_{TM}=0.05
MT E to S	-	-

S ASR	lz=30, lp=-5, fp=40,
	n=20, d=0.8
		lz=28, lp=-5, fp=40,
		n=10, d=0.5

Resc.	w_{TM}=0.075	w_{TM}=0.1
MT S to E	n_{LM}=20, i=0.8, n_{RT}=1, x=2	n_{LM}=10, i=0.8, n_{RT}=1, x=2

E ASR	lz=30, lp=10, fp=40,
	n=20, d=1.4
		lz=28, lp=10, fp=40,
		n=20, d=1.5

Resc.	w_{TM}=0.1, w_{LM}=0.025, lp'=15	w_{TM}=0.15, w_{LM}=0.05
MT E to S	n_{LM}=10, i=0.8, n_{RT}=1, x=2	n_{LM}=10, i=0.8, n_{RT}=1, x=2

S ASR	lz=30, lp=-5, fp=40,
	n=1, d=0.8
	n_{LM}=20, i=0.05
		lz=28, lp=-5, fp=40,
		n=1, d=0.6,
		n_{LM}=20, i=0.075

Resc.	wlp'=5	w_{TM}=0.1

Table C.3. Parameter settings for the final iterative asr driven system on data set I. Unlisted parameters were set to zero.

	System I	System II
S ASR	lz=26, lp=0, fp=20	lz=28, lp=10, fp=40
MT S to E	-	-

E ASR	lz=30, lp=-10, fp=25,
	n=20, d=1.3
		lz=30, lp=-15, fp=5,
		n=10, d=1.2

Resc.	w_{TM}=0.15, lp'=25	w_{TM}=0.15
MT E to S	-	-

S ASR	lz=30, lp=5, fp=30,
	n=30, d=0.6
		lz=32, lp=10, fp=35,
		n=10, d=1.0

Resc.	w_{TM}=0.15	w_{TM}=0.125
MT S to E	n_{LM}=20, i=0.8, n_{RT}=1, x=2	n_{LM}=20, i=0.8, n_{RT}=1, x=2

E ASR	lz=30, lp=-10, fp=25,
	n=10, d=1.2,
	n_{LM}=10, i=0.025
		lz=30, lp=-15, fp=5,
		n=10, d=1.2,
		n_{LM}=10, i=0.05

Resc.	w_{TM}=0.075, lp'=10,	w_{TM}=0.075
MT E to S	n_{LM}=5, i=0.8, n_{RT}=1, x=2	n_{LM}=5, i=0.9, n_{RT}=1, x=2

S ASR	lz=30, lp=5, fp=30,
	n=10, d=0.8
	n_{LM}=1, i=0.05
		lz=32, lp=10, fp=35,
		n=10, d=0.8,
		n_{LM}=1, i=0.05

Resc.	w_{TM}=0.15	w_{TM}=0.075, lp'=5

Table C.4. Parameter settings for the final iterative asr driven system on data set II. Unlisted parameters were set to zero.

List of Figures

List of Tables

Bibliography

[1] P. Brown, S. Della Pietra S. Chen, V. Della Pietra, S. Kehler, and R. Mercer, “Automatic Speech Recognition in Machine Aided Translation”, *Computer Speech and Language*, 8, 1994.

[2] M. Dymetman, J. Brousseaux, G. Foster, P. Isabelle, Y. Normandin, and P. Plamondon, “Towards an Automatic Dictation System for Translators: the TransTalk Project”, *Proceedings of ICSLP*, Yokohama, Japan, 1994.

[3] Steve Young, “Large Vocabulary Continuous Speech Recognition: a Review”, *IEEE Signal Processing Magazine*, 13(5): 45-57, 1996.

[4] P. Brown, S. Della Pietra, V. Della Pietra, and R. Mercer, “The Mathematics of Statistical Machine Translation: Parameter Estimation”, *Computational Linguistics*, 19(2), pp. 263–311, 1993.

[5] J. Brousseaux, G. Foster, P. Isabelle R. Kuhn, Y. Normandin, and P. Plamondon, “French Speech Recognition in an Automatic Dictation System for Translators: the TransTalk Project”, *Proceedings of Eurospeech*, Madrid, Spain, 1995.

[6] P. Placeway and J. Lafferty, “Cheating with Imperfect Transcripts”, *Proceedings of ICSLP*, Philadelphia, PA, USA, 1996.

[7] Y. Ludovik and R. Zacharski, “MT and Topic-Based Techniques to Enhance Speech Recognition Systems for Professional Translators”, *Proceedings of CoLing*, Saarbrücken, Germany, 2000.

[8] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu, “BLEU: a Method for Automatic Evaluation of Machine Translation”, *Proceedings of ACL*, Philadelphia, PA, USA, 2002.

[9] NIST Report, “Automatic Evaluation of Machine Translation Quality Using N-gram Co-Occurrence Statistics”, *http://www.nist.gov/speech/tests/mt/doc/ngramstudy.pdf*, 2002.

[10] H. Soltau, F. Metze, C. Fügen, and A. Waibel, “A one-pass decoder based on polymorphic linguistic context assignment”, *Proceedings of ASRU*, Madonna di Campiglio, Italy, 2001.

[11] F. Metze, Q. Jin, C. Fügen, K. Laskowski, Y. Pan, and T. Schultz, "Issues in Meeting Transcription - The ISL Meeting Transcription System", *Proceedings of ICSLP*, Jeju Island, Korea, 2004.

[12] S. Vogel, S. Hewavitharana, M. Kolss, and A. Waibel, "The ISL Statistical Machine Translation System for Spoken Language Translation", *Proceedings of IWSLT*, Kyoto, Japan, 2004.

[13] Andreas Stolcke, "SRILM - An Extensible Language Modeling Toolkit", *Proceedings of ICSLP*, Denver, Colorado, USA, 2002.