Enhancing Automatic Speech Recognition for Romanian by Using Machine Translated and Web-based Text Corpora

Horia Cucu[12]*, Laurent Besacier*[2]*, Corneliu Burileanu*[1]*, Andi Buzo*[1]
[1]University “Politehnica” of Bucharest, Romania
[2]LIG, University “Joseph Fourier” Grenoble, France

Abstract

This study addresses the lack of general and domain-specific text resources for Romanian Automatic Speech Recognition (ASR) systems. To overcome this problem, we propose to use Machine Translated (MT) text and the Web, as language modeling resources. The domain-specific ASR system built is eventually evaluated in terms of word error rate and significant improvements are being reported using MT-text and data extracted from the Web. The paper also describes and evaluates a diacritics restoration system for Romanian, which is mandatory to exploit Web Romanian data that comes generally without diacritics. With the methodology presented here, a decent large vocabulary, speaker independent ASR system for Romanian was obtained in a relatively short period of time.

1. Introduction

The field of ASR for high-resourced languages (such as English or French, for example) has reached a peak thanks, notably, to the fact that speech resources needed to build robust acoustic models and text resources used to create general or domain-specific language models are widely available. On the opposite side, for under-resourced languages, the performance of ASR systems is acceptable only in very particular cases: small vocabulary or isolated words recognition or closed task grammar, etc.

For the Romanian language, in particular, researchers have tried to overcome the lack of speech resources by applying innovative training strategies for Hidden Markov Models (HMM) [1, 2, 3] or by using mixed recognition methods (HMMs + neural networks + fuzzy systems) [4]. The results presented by all the latest studies address only isolated words recognition [1, 3] or small vocabulary speech recognition [2, 4]. As for the language models, these studies report only the usage of word loop grammars [1, 4], specific task grammars [2] or small vocabulary bigram models [2]. Moreover, a common conclusion and planned future work of all the above studies is to create a better (general or domain-specific) language model for Romanian. Given this, it is clear that the acquisition and preprocessing of large text corpora for language modeling is still an open topic for the Romanian language. This topic is still open for other low-resourced languages as well, and it was lately addressed by several papers which propose the Web as a resource [5, 6, 7].

On the other hand, the task of porting and adapting language or acoustic resources or even models from high-resourced languages to low-resourced languages makes perfect sense and is expected to produce better ASR systems. Several algorithms and methods of adaptation have been proposed and experimented lately. Word decomposition algorithms for language modeling applied to languages such as Somali, Amharic and Hungarian are described in [8, 9, 10], while [11] presents two techniques (cross-lingual and grapheme-based acoustic modeling) for bootstrapping acoustic models for Vietnamese.

This paper proposes a method for building domain specific ASR systems for under-resourced languages using domain specific text corpora in high-resourced languages and machine translation (MT) systems. This method is applied and evaluated in terms of perplexity (PPL), out of vocabulary (OOV) rate and word error rate (WER) for an ASR system constructed with a tourism specific text corpus (machine) translated from French to Romanian. Preliminary results using a similar method were also reported in [12] using an English – Icelandic MT system, while [13] evaluates a similar method using an English – Japanese MT system, but reports only perplexity improvements, not ASR results. A second novelty introduced by this paper addresses the topic of Romanian text corpora acquisition and diacritics restoration using the Web. Finally, to the best of our knowledge, this is the first study that reports large vocabulary ASR results for Romanian.

2. Adapting general LM to specific domain using MT corpus

The methodology for creating a machine translated enhanced LM is depicted in Figure 1.

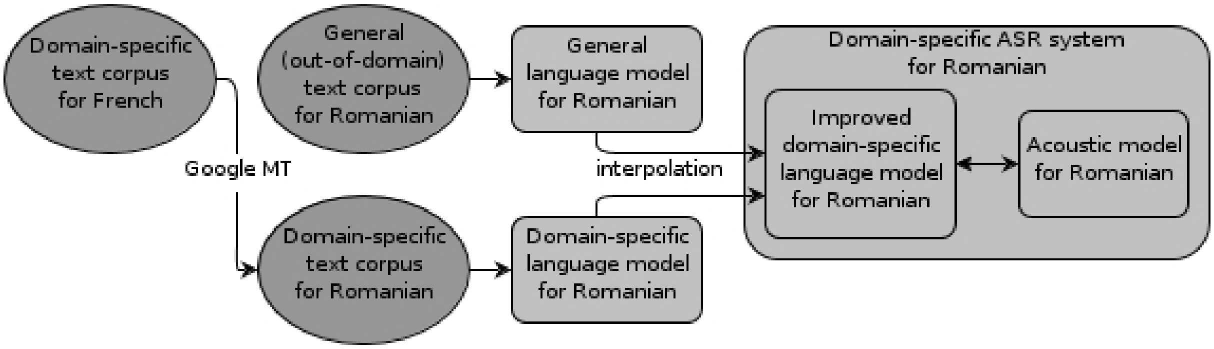
Domain-specific text corpus for French
General (out-of-domain) text corpus for Romanian
General language model for Romanian
Domain-specific ASR system for Romanian
Google MT
interpolation
Improved domain-specific language model for Romanian
Acoustic model for Romanian
Domain-specific text corpus for Romanian
Domain-specific language model for Romanian

Fig. 1. Machine-translated enhanced language model

The method implies the existence of a general (out of domain) text corpus for the low-resourced language, a domain-specific text corpus in a high-resourced language and an MT system for the two languages. As seen in Figure 1, the domain-specific text corpus in the high-resourced language is firstly translated into the low-resourced language. Second, a domain-specific trigram LM and a general trigram LM are constructed using the two corpora. And finally, the two language models are interpolated using an appropriate weight. The idea behind this method is to be able to use the most frequent general language structures that are probably well defined in the out-of-domain LM and, simultaneously, to take advantage of the domain-specific words and structures in the MT LM.

In most of the cases, a general text corpus for the low-resourced language is unavailable, but one can use various methods [5, 6, 7] to acquire it using the Web. The advantage here is that the domain is not relevant; the only important thing is to have correct texts that are representative for the various lexical constructions in the low-resourced language.

Of course, text corpora obtained via the Web is mandatory to be further preprocessed before it can be used for language modeling. The various preprocessing operations include, but are not limited to: html tag stripping, numbers to text conversion, abbreviation expansion, brackets handling, punctuation marks handling, case conversion. A more complex preprocessing operation that might be needed for some languages (and that is mandatory for Romanian) is diacritics restoration. The state-of-the-art methodology for diacritics restoration is depicted in Figure 2.

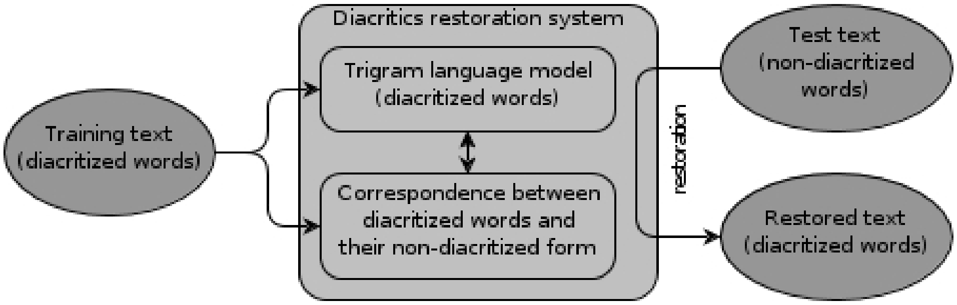
Diacritics restoration system
Trigram language model (diacritized words)
Correspondence between diacritized words and their non-diacritized form
Training text (diacritized words)
Test text (non-diacritized words)
restoration
Restored text (diacritized words)

Fig. 2. Diacritics restoration system

It is compulsory to have a training corpus with correct diacritized words which is needed to create diacritized words bigram or trigram LM. Using the same corpus or, alternatively, a larger vocabulary of diacritized words, a correspondence table between the diacritized words and the non-diacritized words vocabularies must be constructed. In the end, the language model and the mapping table are used to disambiguate the larger, non-diacritics corpus acquired via the Web.

Several methods were proposed in the literature for Romanian diacritics restoration. Among these, [14] and [15] seem to be the most effective. The diacritics restoration system presented in [14] (a system that we will also use in our experiments) uses a slightly different methodology than the one presented above. The main difference is that the diacritics restoration is regarded as a sequential filtering process based on unigrams and bigrams of diacritized words and trigrams of diacritized word-suffixes.

3. Experimental setup

3.1. Speech database and the acoustic model

For all our experiments we have used a HMM based acoustic model that we have previously created and optimized using the CMU Sphinx toolkit[1] default training strategy. The number of senones and the number of Gaussian mixtures per senone state have been changed to 4000 and respectively 16 in order to better adapt the system to the training speech database size and variability. A large (600 thousands words) phonetic dictionary for Romanian was already available [1], but it had to be manually updated with some proper names (hotels, places, etc) that were needed for the domain-specific experiments. The training speech database consists of 54 hours of read speech recorded by 17 speakers and the testing speech database consists of approximately 3.7 hours of read speech recorded by 12 of the previous 17 speakers. The test phrases are divided into four groups based on their subject (Table 1).

Table 1. Testing database summary

	Text			Speech	
Group	**Phrases**	**Running words**	**Subject**	**Speakers**	**Total size**
01	100	1703	news, articles	11	2.07 hours
02	244	1903	dialogue (library)	4	0.71 hours
03	150	913	dialogue (tourism)	3	0.40 hours
04	150	960	dialogue (tourism)	3	0.51 hours

Groups *01* and *02* were part of the database described in [1], while groups *03* and *04* have been especially created for the experiments concerning this study. 300 phrases were randomly selected out of the original French version of the *media* corpus (see Section 3.2). They were manually translated to Romanian and recorded by three speakers. These groups of phrases were removed from the machine translated *media* corpus.

3.2. Text corpora

This section describes the various text corpora that have been used in our experiments. Note that most of the corpora and especially the larger ones have been acquired using the Web.

The *media* corpus is a machine translated[2] corpus. It was translated from its original, already available, French version and comprises tourism specific transcriptions of spontaneous speech. After automatic translation, it consists of about 10 thousand phrases summing up to a total of 64 thousand words.

The *europarl* corpus is a free corpus available on-line[3] [16] for all the European Union's languages and comprises the discussions in the European Parliament. The English or French *europarl* corpora are larger as these countries were part of the EU from its birth, while the Romanian corpus consists of 225 thousand phrases summing up to a total of 5.3 million words with correct diacritics.

The *9am* and *hotnews* corpora have been obtained by automatically downloading and preprocessing all the articles from two on-line[4] newspapers. The texts address all types of news. See Table 2 for further details.

The *misc* corpus was already available before this study began and it comprises newspaper articles, PhD thesis, and literature. It consists of 300 thousand phrases summing up to a total of 9.8 million words with correct diacritics.

Table 2. Romanian text corpora summary

Name	Phrases	Running words	Diacritized
media (MT)	9.8k	64k	most words
europarl	225k	5.3M	yes
9am	3.5M	63M	no
hotnews	6.0M	100M	no
misc	300k	9.8M	yes

[1] CMU Sphinx Speech Recognition Toolkit (http://cmusphinx.sourceforge.net)

[2] Google on-line MT system (http://translate.google.com)

[3] European Parliament Proceedings Parallel Corpus (http://www.statmt.org/europarl)

[4] http://www.9am.ro and http://www.hotnews.ro

3.3. Diacritics restoration systems

As seen in section 3.2, there are some corpora that require diacritics restoration. For this preprocessing operation we had two choices and we have decided to use them both: we were given the opportunity to use the diacritics restoration system described in [14] and we had the resources (diacritized words corpora for training: *misc* and *europarl*) to develop our own diacritics restoration system following the general methodology presented in Section 2.

Given the two diacritics restoration systems we have denoted every restored diacritics corpus in the following manner: *<corpus name>.DRS1* (if the system described in [14] has been used) or *<corpus name>.DRS2* (if the diacritics restoration system we have developed has been employed).

3.4. Language models

For this study, several language models have been developed using the various corpora described in section 3.2 and/or their restored diacritics version(s). All the language models are trigram LMs and have been developed using the SRI-LM toolkit[5]. This toolkit was also used for the interpolation of the general LMs with the domain-specific LM and for the diacritics restoration process (*disambig* command used).

The interpolation of the general LMs with the domain-specific LM has been done with the weights 0.1/0.9 because our goal is to create a domain-specific ASR so the domain-specific LM should prevail. Another thing to be noted is that for the larger corpora and for the experiments that use more corpora, the number of unigrams had to be limited to the most frequent 64000 due to an ASR decoder (Sphinx3) limitation.

4. Diacritics restoration results

Diacritics restoration evaluation will not be done using the standard performance figures (word level accuracy, f-measure, etc) as in [14] and [15], because our only interest is the impact of diacritics restoration on the performance of the ASR system. Consequently, all results tables that will be presented further on compare the LMs in terms of perplexity (PPL), out-of-vocabulary (OOV) rate[6] and ASR performance (WER). All WERs are obtained by comparing the diacritized-words reference phrases with diacritized-words hypotheses (the only exception is Exp 6 in Table 3).

Table 3 compares several ASR systems that use LMs constructed with: a) the original corpus (Exp 1, 3, 6 and 7), b) the original corpus with removed and restored diacritics (Exp 2, 4 and 5) c) the original corpus with restored diacritics (Exp 8 and 9). Note that for the *9am* corpus, which had no diacritics in its original form, the following experiments were performed: i) Exp 6 (WER evaluation with original corpus that does not have diacritics) and ii) Exp 7 (same experiment but WER is evaluated after having restored diacritics on the ASR hypotheses). Note that for the *media* corpus only the tourism phrases were used for evaluation because the results on the other test files groups are irrelevant (the *media* corpus is a small domain specific corpus and test files groups *01* and *02* do not include texts from the same domain).

Table 3. Diacritics restoration evaluation

Exp	LM built with corpus	Test Files Groups: 01 – 02			Test Files Groups: tourism only (03 – 04)		
		PPL	OOV	WER	PPL	OOV	WER
1	europarl	686.4	4.1%	27.0%	532.1	4.8%	34.0%
2	europarl.DRS1	686.5	4.2%	27.3%	532.9	4.8%	33.8%
3	media	-	-	-	42.0	3.4%	18.7%
4	media.DRS1	-	-	-	40.7	3.5%	18.9%
5	media.DRS2	-	-	-	40.4	3.6%	18.7%
6	9am (hyp w/o diacritics)	632.0	26.7%	47.8%	350.0	22.9%	62.3%
7	9am (hyp.DRS2)			22.0%			29.5%
8	9am.DRS1	235.9	1.9%	20.2%	188.2	4.5%	28.4%
9	9am.DRS2	237.8	1.8%	20.3%	189.8	4.2%	28.7%

[5] The SRI Language Modeling Toolkit (http://www-speech.sri.com/projects/srilm)

[6] The OOV rate is related to the LM vocabulary (64k words), while the phonetic dictionary is actually smaller (~50k words), because only a closed-set phonetic dictionary is available (and no phonetizer).

The results of the first five and last two experiments can be used to evaluate the two diacritics restoration systems. As they clearly show, the performance of the ASR system is almost identical when using DRS1, DRS2 or the original corpus, thus we can conclude that for an ASR task both diacritics restoration systems are efficient enough.

Going forward, one can also note the significantly worse results obtained in experiment 6. For this experiment, the ASR hypotheses are w/o diacritics, because the LM is trained with the original *9am* corpus (w/o diacritics). Consequently, this experiment reveals the importance of diacritics restoration for Romanian ASR. Experiment 7 shows the performance of the same ASR system for which a diacritics restoration operation (DRS2) is applied on the ASR hypotheses. The results are significantly better than in experiment 6, but a little worse than in experiments 8 and 9 where the diacritics restoration is performed on the LM training data instead of on the ASR hypotheses. To conclude, for an ASR task, the diacritics restoration operation is mandatory for Romanian and it is more effective if it is applied on the training side than on the test side. Such techniques allows us to exploit more LM data extracted from the Web, which eventually improve out-of-domain ASR performance (compare Exp 1 to Exp 8 and 9).

In the next section we investigate portability to the tourism domain using machine-translated text.

4. Domain adaptation results using MT enhanced LMs

Given the conclusions in the previous section regarding diacritics restoration, we have selected only DRS2 for this next task and we present below the results for the domain adaptation experiments. Table 4 compares four ASR systems that use general language models constructed with out-of-domain corpora. As a reminder, *europarl* corpus is the smallest (5.3 million words), *9am* corpus is medium sized (63 million words) and *hotnews* corpus is the largest (100 million words).

Table 4. General (out-of-domain) language models evaluation

Exp	LM built with corpus	Test Files Groups: all (01 – 04)		
		PPL	OOV	WER
1	europarl	669.1	4.1%	28.3%
2	9am.DRS2	232.5	2.0%	21.8%
3	hotnews.DRS2	205.9	2.2%	21.1%
4	europarl + 9am.DRS2 + hotnews.DRS2	190.6	2.0%	20.7%

The conclusion that can be drawn based on the results presented in this table is that the LM and ASR system performance gradually increases as larger corpora are used, but at some point this growth saturates. For example, the performance of the LM built with all the corpora (*europarl + 9am.DRS2 + hotnews.DRS2*) is only a little better than the performance of the LM built with the *hotnews.DRS2* corpus, although a large amount of extra data (*europarl + 9am.DRS2*) is used to create the LM.

The evaluation of the domain-specific language models is summarized in Table 5. Experiment 1 presents the performance of the best general LM (the one built with *europarl + 9am.DRS2 + hotnews.DRS2* corpora), experiment 2 lists the results obtained for the domain-specific MT-based-LM (built with the *media.DRS2* corpus) and experiment 3 shows the improvements obtained in ASR by interpolating the general and the domain-specific LMs as described in Section 2.

Table 5. Domain-specific language models evaluation

Exp	ASR with LM	Test Files Groups: 03			Test Files Groups: 04			Test Files Groups: 03 – 04		
		PPL	OOV	WER	PPL	OOV	WER	PPL	OOV	WER
1	best general LM	180.7	3.6%	26.0%	168.3	4.1%	29.4%	174.3	3.8%	27.8%
2	domain specific LM	52.2	3.1%	18.3%	31.5	4.2%	19.1%	40.4	3.6%	18.7%
3	domain specific LM interpolated with best general LM	52.4	0.4%	14.2%	38.0	0.9%	18.0%	44.5	0.7%	16.1%

As Table 5 states, although the perplexity of the interpolated LM is higher that the perplexity of the machine translated LM, the ASR results are better in both the evaluated cases thanks to the smaller out-of-vocabulary rate. As expected, the ASR system in experiment 3 has the lowest out of vocabulary rate. In conclusion the machine translated corpus plays a significant role in the improvement of the general LM and consequently the improvement of ASR for a domain-specific task.

5. Conclusions

This paper has presented a new method of improving ASR for a low-resourced language (Romanian) using machine translated text originally available in a high-resourced language (French). The study was based on a small amount of initial resources: a very small size domain-specific French corpus and a small size out-of-domain Romanian corpus and provides methods to obtain and preprocess broader resources using the Web. Moreover the machine translation is regarded as a Web-based free service.

A general diacritics restoration method was also described and compared with the state-of-the-art diacritics restoration system for Romanian. The study shows that from the ASR point of view the two diacritics restoration system are equally efficient.

The final results demonstrate that adapting a general LM to a specific task using a machine translated corpus significantly improves the ASR performance for that task. On the other hand the final section presents the first large vocabulary, speaker independent ASR results for Romanian.

Acknowledgements. The research reported in this paper was funded by the Sectoral Operational Program Human Resources Development 2007-2013 of the Romanian Ministry of Labor, Family and Social Protection through the Financial Agreement POSDRU/6/1.5/S/16.

References

1. *Cucu, H., Buzo, A., Burileanu, C.*: Optimization Methods for Large Vocabulary, Isolated Words Recognition in Romanian Language, In: UPB Scientific Bulletin – Series C, No. 2, Bucharest, Romania, to appear (2011)

2. *Militaru, D., Gavăt, I., Dumitru, O., Zaharia, T., Segărceanu, S.*: Protologos, System for Romanian Language Automatic Speech Recognition and Understanding, In: Proc. of SPED 2009, pp. 21 – 32, Constanta, Romania (2009)

3. *Popescu, V., Burileanu, C., Rafaila, M., Calimanescu, R.*: Parallel Training Algorithms for Continuous Speech Recognition, Implemented in a Message Passing Framework, In: Proc. of the 14th European Signal Processing Conference "EUSIPCO", Florence, Italy (2006)

4. *Oancea, E., Gavăt, I., Dumitru, O., Munteanu, D.*: Continuous Speech Recognition for Romanian Language Based on Context Dependent Modeling, In: Proc. of International IEEE Conference „COMMUNICATIONS 2004", pp. 221-224, (2004)

5. *Le, V.B., Bigi, B., Besacier, L., Castelli, E.*: Using the Web for fast language model construction in minority languages, In: Proc. of Eurospeech2003, pp. 3117-3120, Geneva, Switzerland (2003)

6. *Draxler, C.*: On Web-based Speech Resource Creation for Less-Resourced Languages, In: Proc. of Interspeech2007, Antwerp, Belgium, (2007)

7. Cai, J.: Transcribing southern min speech corpora with a web-based language learning system, In: Proc. of SLTU2008, Hanoi, Vietnam (2008)

8. *Abdillahi, N., Nocera, P., Bonastre, J.F.*: Automatic transcription of Somali language, In: Proc. of ICSLP2006, pp. 289-292, Pittsburgh, PA, USA (2006)

9. *Pellegrini, T., Lamel, L.*: Investigating Automatic Decomposition for ASR in Less Represented Languages, In: Proc. of ICSLP2006, Pittsburgh, PA, USA (2006)

10. *Mihajlik, P., Fegyó, T., Tüske, Z., Ircing, P.*: A Morpho-graphemic Approach for the Recognition of Spontaneous Speech in Agglutinative Languages – like Hungarian, In: Proc. of Interspeech2007, Antwerp, Belgium, (2007)

11. *Le, V.B., Besacier, L.*: Automatic Speech Recognition for Under-Resourced Languages: App. to Vietnamese Language, In: IEEE Transactions on Audio, Speech and Language Processing (2009)

12. *Jensson, A., Iwano, K., Furui, S.*: Development of a speech recognition system for Icelandic using machine translated text, In: Proc. of SLTU2008, Hanoi, Vietnam (2008)

13. *Nakajima, H., Yamamoto, H., Watanabe, T.*: Language Model Adaptation with Additional Text Generated by Machine Translation, In Proc. of COLING, vol. 2, pp. 716-722 (2002)

14. *Ungurean, C., Burileanu, D., Popescu, V., Negrescu, C., Dervis, A.*: Automatic Diacritics Restoration for a TTS-based E-mail Reader Application, In: UPB Scientific Bulletin – Series C, Vol. 70, No. 4, Bucharest, Romania (2008)

15. *Tufis, D., Chitu, A.*: Automatic Diacritics Insertion in Romanian Texts, In: Proc. of COMPLEX'99, pp. 185-194, Pecs, Hungary (1999)

16. *Koehn P.*: Europarl: A Parallel Corpus for Statistical Machine Translation, In: Proc. of MT Summit, Phuket, Thailand (2005)