ASR DOMAIN ADAPTATION METHODS FOR LOW-RESOURCED LANGUAGES: APPLICATION TO ROMANIAN LANGUAGE

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 investigates the possibility of using statistical machine translation to create domain-specific language resources. We propose a methodology that aims to create a domain-specific automatic speech recognition system for a low-resourced language when in-domain text corpora are available only in a high-resourced language. We evaluate a new semi-supervised method and compare it with previously developed semi-supervised and unsupervised approaches. Moreover, in the effort of creating an out-of-domain language model for Romanian, we introduce and experiment an effective diacritics restoration algorithm.

Index Terms— ASR domain adaptation, SMT, language modeling, diacritics restoration

1. INTRODUCTION

Adaptation methods are a very practical way of bootstrapping the development of automatic speech recognition (ASR) systems for low-resourced languages. The acquisition of speech databases and text corpora for low-resourced languages is generally a costly task, but these costs can be lowered or even avoided by using various *acoustic or language adaptation* methods.

This motivation has lately led many research groups to design and apply various adaptation techniques to low-resourced languages. For example, [1] investigates several *acoustic model adaptation* techniques for bootstrapping acoustic models for Vietnamese, while [2] is concerned with *adapting acoustic models* for multi-origin non-native speakers. Several *language adaptation* methods for spoken dialogue systems are proposed in [3] (English to Spanish) and [4] (French to Italian). These last two methods use statistical machine translation (SMT) to adapt language resources and models. A similar technique is used in [5] to create resources for Icelandic ASR.

This work presents our progress in SMT-based ASR *domain adaptation* methods. In [6] we have presented an unsupervised domain adaptation method and in [7] we have proposed two distinct semi-supervised techniques. In this paper we introduce a new domain adaptation method and show that it outperforms the previously proposed techniques. Moreover, in the context of building an out-of-domain language model (LM) for Romanian, we focus upon a mandatory language processing operation: *diacritics restoration*.

The general methodology of creating an in-domain ASR system is presented in Figure 1. For low-resourced languages some resources might be missing. This is exactly the case for Romanian: neither general text, nor in-domain text is available for language modeling. The following sections will address these problems as follows: Section 2 discusses the problem of creating *in-domain* textual data for ASR domain adaptation, while Section 3 deals with some specific issues regarding the acquisition of *general* text corpora.

To be more specific, in this paper we apply SMT-based adaptation methods to port an in-domain (tourism) French corpus to Romanian, with the final goal of creating an in-domain ASR system for Romanian.

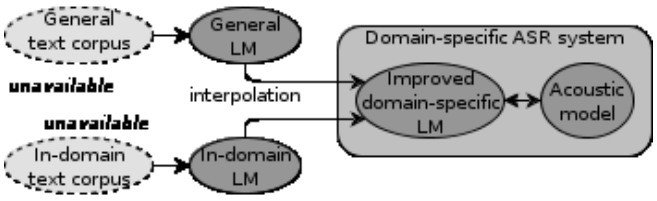

Figure 1. The general ASR domain adaptation methodology

2. SMT-BASED LANGUAGE MODEL ADAPTATION

The domain adaptation methodology we propose in this paper aims to utilize an in-domain text corpus available in a high-resourced language to create in-domain textual data for a low-resourced language. The ideal, fully-supervised scenario would imply a human expert translating the whole corpus. This process would optimize performance, but, on the other hand, it is very expensive. The least expensive scenario would imply a SMT system. In this second case the performance would be clearly influenced by the errors in the machine translated text. An efficient balance between

cost and performance can be obtained with various semi-supervised scenarios. Machine translation can be used to cut the costs and win time, while a human expert can post-edit (correct) the SMT output.

In our first work on this topic [6] we investigated the fully unsupervised scenario. We used Google online MT system to translate a French in-domain corpus to Romanian without any human intervention. Even if the resulted text was not error-proof we used it to create an in-domain language model and demonstrated that the method improves ASR performance for in-domain speech.

Going further, we developed two semi-supervised scenarios [7]. The initial in-domain French text was split into two parts denoted *partA* and *partB*. Both of them were Google-translated and we obtained two in-domain Romanian texts: *partA_GoMT* and *partB_GoMT*. The second, smaller part was manually corrected generating *partB_GoMTpp*. Finally, as illustrated in Figure 2 (method 1), *partA_GoMT* was concatenated with *partB_GoMTpp* to obtain a complete in-domain Romanian text.

The second semi-supervised adaptation method regards *partB* of the in-domain French text and the Romanian *partB_GoMTpp* text as parallel corpora and uses them to train a domain-specific SMT system. Undoubtedly, the resulted SMT system will be worse than Google's when *partB* is very small, but it may out-perform Google's as more text is manually corrected. The trained SMT system was afterwards used to translate *partA* of the in-domain text, generating the *partA_dsMT*. Finally, as shown in Figure 2 (method 2), *partA_dsMT* was concatenated with *partB_GoMTpp* to obtain a complete in-domain Romanian text.

In our previous work [7] we concluded that the second method out-performs the first one when more than 5% (500 phrases) of the text is manually corrected. Nevertheless, the first method had its own advantages even when the second one out-performed it (see in-depth analysis). Consequently, in this paper we propose the usage of both the Google-translated text (*partA_GoMT*) and the text translated by our domain-specific SMT system (*partA_dsMT*) to create the in-domain language model. This third method of creating in-domain Romanian text is shown in Figure 2 (method 3).

3. GENERAL LANGUAGE MODEL CONSTRUCTION – SPECIFIC ISSUES

There are some specific issues regarding the construction of an out-of-domain language model for Romanian. First, there are no large enough publicly-accessible text corpora and second, most of the text corpora which can be collected via the Web need in-depth preprocessing. These issues are addressed in the next sections.

3.1. General text corpora acquisition

Romanian is a low-resourced language from the point of view of plain text corpora. In 2010, Macoveiciuc [8] reported on the acquisition of RoWaC, a 50M words Romanian corpus, and its commercial availability within the Sketch Engine. The authors state that prior to their work in 2010, there were no large, publicly-accessible, general-language corpora for Romanian.

In conclusion, with no data or little data for language modeling, we were required to create our own general-Romanian corpus. The best solution we found was to use the Web-as-Corpus concept and collect textual data from various on-line sources [6].

All the large Romanian news corpora we collected via the Web lacked diacritics. Consequently, among other, simpler preprocessing operations we have addressed, we were also required to construct a diacritics restoration tool for Romanian.

3.2. Diacritics restoration

Romanian is a language that makes intensive use of diacritics. Even though it uses only 5 diacritical characters (ă, â, î, ş, ţ), their occurrence frequency is vey high: about 30% to 40% of the words in a general text contain at least one diacritical character. A text that lacks diacritics would generally have these characters substituted by their non-diacritical forms: a, a, i, s, t. Even though for a human reader the meaning of a text without diacritics is most of the times obvious (given the paragraph context), the diacritics restoration task is not trivial for a computer.

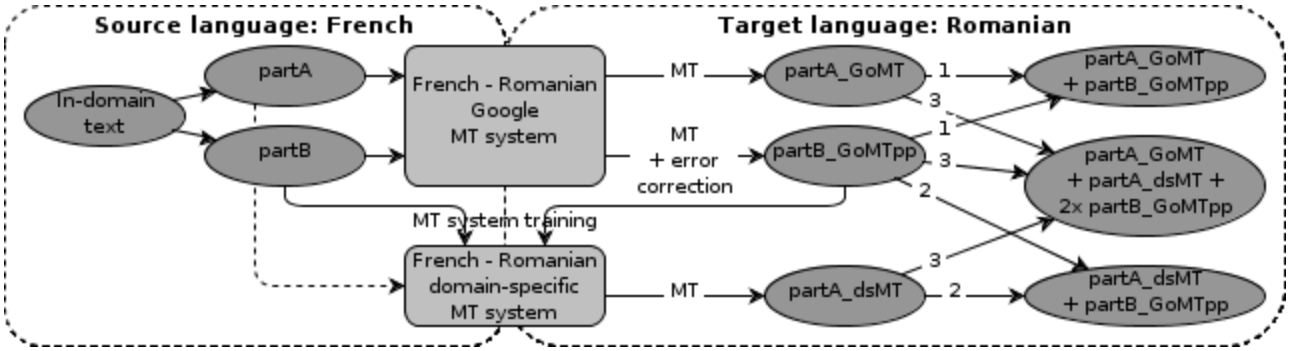

Figure 2. Semi-supervised, SMT-based language model adaptation methods

All the Romanian news corpora which were acquired using the WaC (Web-as-Corpus) approach come without diacritics. For a news article, the diacritics are not very important since any reader has access to the paragraph-level context and ambiguities seldom appear.

On the other hand, for an ASR task the lack of diacritics in the output text is not acceptable, because the output text could be very short and consequently ambiguous. Therefore, an automatic diacritics restoration system is definitely needed. It can be used to restore the diacritics on the output text of a diacritics-lacking ASR system or to restore the diacritics on all the corpora required for language modeling and thus to create an ASR system which directly outputs texts with diacritics.

A statistical language modeling method which has been successfully used for several disambiguation tasks (including true-casing [9]) is proposed in this paper for diacritics restoration. The only resource needed by this method is a text corpus with correct diacritics. Based on this corpus, two higher-level structures are built: an n-gram language model and a probabilistic map (which links all non-diacritical word forms to all their possible diacritical word forms). The probabilities for the various diacritical word forms were estimated from a text corpus with correct diacritics. An excerpt of the map is shown in Figure 3.

```
…
dacia: dacia 1.0
fabricand: fabricând 1.0
pana: pana 0.005, pană 0.008, până 0.987
sarmana: sărmana 0.847, sărmană 0.153
tari: tari 0.047, ţari 0.002, ţări 0.942, târî 0.008
…
```

Figure 3. Probabilistic map excerpt

Given a text corpus in which the diacritics are partly or entirely missing we estimate the diacritical form of every word in the corpus in a word-by-word manner. If the non-diacritical word form *ndw* is not found in the probabilistic map we leave it unchanged. Otherwise, we estimate the diacritical form dw^*, given the preceding sequence of N diacritical words *dws*, by finding the diacritical word form dw_i that maximizes this formula:

$$dw^* = \arg\max_{dw_i} p(dw_i \mid dws) \times p(dw_i \mid ndw)$$

The first factor in the equation is estimated by the n-gram language model, while the second one is estimated by the probabilistic map.

4. DIACRITICS RESTORATION EXPERIMENTS

4.1. Experimental setup and results

To build up the system (n-gram LM and probabilistic map) we have used two plain text corpora: a 5.3M words corpus collected via the Web and a 9.8M words corpus which was prior available.

In order to find the best setup for the diacritics restoration system we have varied the LM order N from 2 to 5 and we have also tried to use a plain probabilistic map (which assigns equal likelihoods to every diacritical word forms of a non-diacritical word form). The SRI-LM Toolkit [10] was used to create the LMs and to perform the disambiguation (disambig tool) of the various existing non-diacritical word forms.

The various versions of the system were evaluated in terms of word error rate (WER) and character error rate (ChER), on a large text corpus (1.2M words), using the NIST Scoring Toolkit. The various experimental results are presented in Table 1.

LM	*ndw-dw* map	WER [%]	ChER [%]
2-gram	probabilistic	2.07	0.50
3-gram	probabilistic	1.99	0.48
4-gram	probabilistic	2.00	0.49
5-gram	probabilistic	2.01	0.49
3-gram	plain	2.24	0.54

Table 1. Diacritics restoration results

As noted from Table 1, the LM-order (N) variations do not bring important improvements. However, the results obtained using a probabilistic map, instead of a plain map, are much better.

4.2. Related work

Besides this method, several, fundamentally different, diacritics restoration methods were developed for the Romanian language.

An elaborate, knowledge-based diacritics restoration method, using *part-of-speech tagging* to disambiguate the different diacritical words hypotheses, is introduced in [11]. Nevertheless, this method was reported to have lower performance figures than our proposed algorithm: a 2.25% WER and a 0.60% ChER. These results are obtained on a different test set than ours (there is no standard evaluation corpus for Romanian diacritics restoration).

In [12] the diacritics restoration system is regarded as a sequential filtering process based on *unigrams and bigrams of diacritical words* and *trigrams of diacritical word-suffixes*. The authors insist on the fact that this method is adapted to Romanian thanks to the usage of *word-suffixes trigrams*. In 2008, the authors reported a 2.13% WER (on the same test set as ours), but after various refinements [13] they reported even better results: a 1.4% WER and a 0.4% ChER (on a different test set).

In conclusion, we assert that the diacritics restoration system we have developed is one of the best available for Romanian, and can be considered as sufficient for our ASR experiments.

4.3. Diacritics restoration in the context of ASR

The reason we developed a diacritics restoration system was to correct the large text materials (169M words) we

collected via the Web, with the final goal of creating a general language model for the Romanian ASR system. Consequently, it makes perfect sense to evaluate the diacritics restoration system in the context of ASR.

For this, we used the speech resources and model presented in Section 5.1. For language modeling we used the text without diacritics (exp 1 and 2), the text with diacritics restored using our method (exp 3) and the text with diacritics restored using the system presented in [13] (exp 4). This last text was provided by the authors of [13]. In exp 2 we have restored the diacritics on the hypotheses text outputted by the ASR system.

Exp	Diacritics restoration	WER
1	no diacritics restoration	64.5%
2	on hypotheses text, after ASR (using this method)	30.5%
3	on text corpus, before LM (using this method)	29.7%
4	on text corpus, before LM (using [13])	29.4%

Table 2. Diacritics restoration in the context of ASR

Table 2 presents the results in terms of automatic speech recognition WER. In exp 1 we compared hypotheses texts without diacritics with reference texts with diacritics. The high WER argues for the necessity of diacritics restoration for Romanian. Comparing exp 2 and exp 3 we can conclude that better results are obtained if the diacritics restoration is done on the text corpus, before language modeling, as opposed to the hypotheses texts. Experiments 3 and 4 show the difference in ASR performance between the best diacritics restoration system for Romanian [13] and the method we have proposed in this paper.

5. DOMAIN ADAPTATION EXPERIMENTS

5.1. Experimental setup

For all ASR experiments presented in this work we have used the same HMM-based acoustic model [7]. The 36 phonemes in Romanian were contextually modeled with 4000 HMM senones and 16 Gaussian mixtures per senone state [6]. The acoustic model was previously created and optimized (using the CMU Sphinx Toolkit) with a training database of about 54 hours of Romanian read speech. This speech database was progressively developed by our research group and now comprises isolated words, general newspaper articles and domain-specific (library) dialogues. The texts were recorded by 17 speakers (7 males and 10 females). The phonetic dictionary used in the experiments was created using a graphemes-to-phonemes conversion tool [7] and covers all the words in the language models.

Using the Web-as-Corpus approach we have collected a large text corpus (169M words), whose diacritics were restored as discussed in the previous section (exp 3). This corpus was used to create the out-of-domain language model for Romanian. The domain-specific language model was obtained using a French tourism-specific corpus (64k words) translated to Romanian with the various SMT-based methods presented in Section 2. The domain-specific SMT system required by methods 2 and 3 was developed using the Moses Toolkit. The SRI-LM Toolkit was used to create all the language models and to eventually interpolate them.

The evaluation of all the language models was done in terms of perplexity (PPL), out-of-vocabulary (OOV) rate, trigram hits and ASR word error rate (WER), on a test database which contains 55 minutes of tourism-specific Romanian read speech.

5.2. Experimental results

Table 3 presents the ASR results obtained for the in-domain language models after the interpolation with the out-of-domain language model. The results for the three semi-supervised methods are grouped in three main columns. The first results line is repeated in every column: these are, in fact, the results obtained for the unsupervised adaptation method (in this case no error correction was performed on the Google translated corpus). The other lines show the results improvements as 5%, 10%,… 40% of this corpus was corrected.

The comparison between the first two semi-supervised methods has been presented in [7]. We will now focus on the improvements brought by the third method.

First of all we observe that all the three semi-supervised methods exhibit significant better performance figures than the basic unsupervised method even when only a small amount of corrected data (5%) is used. These methods issue better and better ASR systems as more machine translated text is being corrected, but the growth in performance eventually saturates.

Secondly, we see that all the performance figures for the third method are better than those for the other two, regardless of the amount of corrected data. In particular, for

	method 1				method 2				method 3			
partB size	PPL	OOV [%]	3gram hits [%]	WER [%]	PPL	OOV [%]	3gram hits [%]	WER [%]	PPL	OOV [%]	3gram hits [%]	WER [%]
00%	42.5	0.80	55.4	16.2	42.5	0.80	55.4	16.2	42.5	0.80	55.4	16.2
05%	34.4	0.80	56.0	14.6	36.3	0.80	58.8	14.2	28.8	0.75	60.1	13.1
10%	32.4	0.53	56.8	13.9	30.1	0.53	58.6	12.7	27.1	0.48	60.3	12.5
20%	29.0	0.48	57.7	13.1	26.7	0.48	59.5	12.6	24.8	0.48	61.3	11.8
30%	26.6	0.48	58.2	12.4	24.3	0.48	59.9	11.6	23.9	0.48	61.9	11.3
40%	25.2	0.48	59.1	12.2	23.8	0.48	60.2	11.5	23.8	0.48	62.2	11.2

Table 3. Improved in-domain language model results (after interpolation with out-of-domain LM)

small amounts of corrected data (5-10%), the third method is significantly better than the other two. Thanks to its construction methodology, this method benefits from in-domain words and sequences of words from both the Google translation and the domain-specific translation. This explains the better perplexity and 3-gram hits and consequently the lower WER. The OOV rate is most of the times equal, regardless of the method, thanks to the broad coverage of the out-of-domain language model.

When a large amount (30-40%) of corrected data is used, the second and the third semi-supervised methods exhibit similar performance figures. In conclusion, the semi-supervised domain adaptation method proposed in this paper is much more adequate than the previously proposed methods when only a small part of corrected text is available.

5.3. Related work

Unsupervised language model domain adaptation using SMT (English to Japanese) text was proposed back in 2002 by [14]. However, This paper only reports perplexity results and does not make any investigations on semi-supervised approaches.

In 2008, [5] proposed a similar unsupervised SMT (English to Icelandic) method, but used it for creating the out-of-domain language model. This paper is also focused on the impact on ASR, reporting WER improvements obtained thanks to the SMT text, but the analysis is still limited to the basic unsupervised scenario.

A more recent paper [3] extends the analysis to several domains in the effort of porting an English ASR system to Spanish. The translation is also done in an unsupervised fashion.

In conclusion, the unsupervised methodology is not new, but its semi-supervised extensions, which were reported to bring valuable improvements, represent a real novelty in this field.

6. CONCLUSION

In this study we have dealt with two main problems involved in creating a domain-specific ASR system for Romanian: the acquisition and pre-processing of a general text corpus and the development of an in-domain text corpus. We have collected the largest text corpus available for Romanian and we have proposed an efficient method to overcome the problem of missing diacritics.

The in-domain textual data was obtained using an innovative SMT-based semi-supervised methodology. The method proposed in this paper needs only a small amount of manual text corrections (which implies only a small amount of extra work) to significantly improve ASR performance for in-domain speech utterances.

7. REFERENCES

[1] V.B. Le, L. Besacier, "Automatic Speech Recognition for Under-Resourced Languages: App. to Vietnamese Language," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 17, no. 8, pp. 1471-1482, 2009.

[2] S. Sam, E. Castelli, L. Besacier, "Unsupervised Acoustic Model Adaptation for Multi-Origin Non Native ASR," *Interspeech 2010*, Tokyo, 2010.

[3] K. Suenderman, J. Liscombe, "Localization of speech recognition in spoken dialog systems: How machine translation can make our lives," *Interspeech 2009*, Brighton, U.K.

[4] B. Jabaian, L. Besacier, F. Lefevre, "Combination of Stochastic Understanding and Machine Translation Systems for Language Portability of Dialogue Systems," *ICASSP 2011*, Prague, p. 5612-5616, 2011.

[5] A. Jensson, K. Iwano, S. Furui, "Development of a speech recognition system for Icelandic using machine translated text," *SLTU 2008*, Hanoi, Vietnam, 2008.

[6] H. Cucu, L. Besacier, C. Burileanu, A. Buzo, "Enhancing Automatic Speech Recognition for Romanian by Using Machine Translated and Web-based Text Corpora," *SPECOM 2011*, Kazan, Russia, 2011.

[7] H. Cucu, L. Besacier, C. Burileanu, A. Buzo, "Investigating the Role of Machine Translated Text in ASR Domain Adaptation: Unsupervised and Semi-supervised Methods," *ASRU 2011*, Hawai, USA, to appear 2011.

[8] M. Macoveiciuc, A. Kilgarriff, "The RoWaC Corpus and Romanian Word Sketches", *Multilinguality and Interoperability in Language Processing with Emphasis on Romanian*, Romanian Academy Publishing House, Bucharest, pp. 151-168, 2010.

[9] L. Liţă, A. Ittycheriah, S. Roukos, N. Kambhatla, "tRuEcasIng," *ACL 2003*, Sapporo, Japan, p.152-159, 2003.

[10] A. Stolcke, "SRILM - an extensible language modeling toolkit," *ICSLP 2002*, Colorado, USA, 2002.

[11] D. Tufiş, D. Ceauşu, "DIAC+: A Professional Diacritics Recovering System," *LREC 2008*, Marrakech, Morocco, 2008.

[12] C. Ungurean et al., "Automatic Diacritics Restoration for a TTS-based E-mail Reader Application", *UPB Scientific Bulletin – Series C*, vol. 70, no. 4, Bucharest, 2008.

[13] C. Ungurean, D. Burileanu, "An advanced NLP framework for high-quality Text-to-Speech synthesis," *SPED 2011*, Braşov, Romania, 2011.

[14] H. Nakajima, H. Yamamoto, T. Watanabe, "Language Model Adaptation with Additional Text Generated by Machine Translation," *COLING 2002*, vol. 2, pp. 716-722, 2002.