Statistical Error Correction Methods for Domain-Specific ASR Systems

Horia Cucu[1], Andi Buzo[1], Laurent Besacier[2], Corneliu Burileanu[1]

[1]University "Politehnica" of Bucharest, Romania
{horia.cucu, andi.buzo, corneliu.burileanu}@upb.ro
[2]LIG, University Joseph Fourier, Grenoble, France
laurent.besacier@imag.fr

Abstract. Whenever an ASR company promises to deliver error-proof transcripts to the end user, manual verification and correction of the raw ASR transcripts cannot be avoided. This manual post-editing process systematically generates new and correct domain-specific data which can be used to incrementally improve the original ASR system. This paper proposes a statistic, SMT-based ASR error correction method, which takes advantage of the past corrected ASR errors to automatically post-process its future transcripts. We show that the proposed method can bring more than 10% WER improvements using only 2000 user-corrected sentences.

Keywords: ASR, error correction, language modeling, statistical machine translation (SMT)

1 Introduction

There are many applications in which general, large vocabulary ASR (Automatic Speech Recognition) systems are the only choice, because a potentially unknown speaker can potentially speak about anything. In this kind of applications the ASR word error rate (WER) is generally around 15-20%. However, most ASR applications have a much more specific task (limited number of speakers, speaking domain, etc.) and, for these simpler scenarios, the customer asks for better performance. Several model adaptation methods have been proposed to increase the general ASR's performance for a specific task. These methods require task-specific acoustic data to adapt the ASR's acoustic model and task-specific textual data to adapt the ASR's language model. The drawback of these model adaptation methods is that they require access to the internals of the ASR system and therefore they cannot be used if the ASR system is purchased as a *black-box*.

Among the various ASR applications, there are cases in which the user can benefit from error-prone transcriptions (i.e. spoken document retrieval, on-line movie captioning, etc.), but there are also cases in which the transcription must be manual-

ly post-edited (corrected) to become useful (i.e. dictation, interviews/news transcription, etc.). In this second case, the ASR user will always post-edit the raw transcriptions to create error-proof transcripts (for the final reader), systematically generating new and correct domain-specific data.

This paper deals with the latter case, investigating the ways in which the user-generated data can be used to incrementally improve the ASR's output. Although this study also analyzes the scenario in which the ASR system itself can be improved (and shows that its language model can be adapted to obtain better transcriptions), our main focus is on the scenario in which *the ASR system is regarded as a black-box*. In this scenario, we show that the user-generated data can be used to train a statistical post-editor (SPE), which can be afterwards employed to correct the raw ASR transcriptions.

The remainder of the paper is organized in five sections. Section 2 presents the most relevant related works in ASR errors correction, section 3 describes our proposed correction methods and section 4 evaluates them. In section 5 we analyze and discuss the various types of corrections based on some particular examples and in section 6 we draw the conclusions of this work.

2 Related Work

ASR error correction methods can be broken down in two categories: a) ASR adaptation methods and b) transcripts post-editing methods. ASR adaptation methods have been proven to be very effective when the recognition task is strictly defined and when the user has access to the internals of the ASR system. In [1] the user corrected text is used to identify the most probable cause for the error: out-of-vocabulary (OOV) word, wrong pronunciation or language model (LM) probabilities. Based on the identified error, the ASR is automatically adjusted: OOVs are inserted in the lexicon, the probability of the corrected bigram or trigram is boosted, new word pronunciations are generated. In our former papers [2-3] we also showed that LM adaptation can be successfully used to adapt an ASR system to a specific domain, even if the textual data is only available in a different language.

In other scenarios the ASR system is purchased as a black-box and the user does not have any hooks to modify the ASR models. In that case the ASR error correction methods apply a post-editing block to correct the errors in the raw ASR transcripts. One of the first works on this subject [4] uses a fertility channel model to correct 1-to-1, 1-to-2 and 2-to-1 errors. The paper shows that, for a particular dataset, the post-editing correction can be combined with LM adaptation to obtain even better results (24% relative WER improvement). Jung, Jeong, and Lee [5] also apply this fertility channel model (using syllables instead of words) along with an improved LM – that incorporates statistical and other higher level linguistic knowledge. They obtain much better results (on a very domain-specific Korean speech database): 40% relative WER improvement using only 70 training utterances.

Another approach in ASR error correction uses statistical replacement rules extracted from a corpus of raw transcripts along with their manual corrections. Kaki,

Sumita and Iida [6] report on using character co-occurrence rules and obtain an improvement of 8.5% relative WER, when 4300 utterances (from a Japanese travel specific speech database) were used for training. Brandow and Strzalkowski [7] propose a method based on word sequences replacement rules, but do not provide any evaluation figures. Mangu and Padmanabhan [8] use ASR confusion networks composed of several word confusion sets to define rules that specify when the second candidate in a confusion set should be preferred over the first one. This method's reported improvement is only 4.2% relative WER, when 4000 speech utterances (from the Switchboard database) are used for training. Sarma and Palmer [9] regard ASR error correction only from an information retrieval point of view. They compile co-occurrence statistics for each word in the vocabulary and finally use these statistics to identify possible ASR errors, but do not provide an ASR overall evaluation.

In this paper we first analyze the potential of (unsupervised and semi-supervised) LM adaptation to correct ASR errors and propose a *statistical post-editing error correction method* inspired from statistical machine translation (SMT). Our proposed method goes beyond the state-of-the art by employing *a new strategy for error detection and correction* which is based on SMT principles and tools (Moses SMT Toolkit [10]). The method resembles in some aspects with the works presented in [4] and [5] because the SMT system also incorporates a fertility model, but extends this idea by generalizing this type of model. Moreover, we are the first to use a regular SMT system for ASR transcription post-editing and in the end we describe how the SMT system's *phrase translation table* can be further used to deeply analyze the ASR errors. We evaluate the method on a Romanian domain-specific speech database (weather news) and obtain an improvement of 10.5% relative WER when 2000 utterances are used for training.

3 ASR Output Error Correction

Consider the scenario in which a *general ASR system* is used *daily* to transcribe domain specific speech, for example broadcast weather news. Due to the mismatch between the general ASR's language model and the recognition domain (weather news), the recognition accuracy will be relatively poor. The ideal case would be to have a domain-specific ASR system, but this is not always possible (due to the lack of domain-specific data). In our previous studies [2-3] we considered the same problem and showed that if we have domain-specific text data in a different language we can translate it to the target language and successfully use it to adapt the ASR system. Now the premises have changed: we have *domain-specific audio data* in the target language, but not *domain-specific text data* (the required media type is not available). Consequently, we propose employing a general ASR system to transform the audio data into text and then use it to incrementally improve the ASR's output.

3.1 The Unsupervised Scenario

In the fully unsupervised scenario the raw ASR transcriptions are directly used for LM adaptation. Apparently, the system would have no means of learning from its previous errors, because it does not know when it makes mistakes. However, adapting the language model with the raw transcriptions might be beneficial, because the LM probabilities for domain-specific words and word sequences will be boosted up.

3.2 The Semi-supervised Scenarios

In the scenario considered in this study (daily transcriptions of broadcasted weather news) the ASR user (i.e. a media company offering transcription services) cannot deliver the raw transcripts to the final user due to the fact that they contain many errors. This is an application where it is mandatory to have error-proof transcriptions; therefore the ASR output must be manually verified and post-edited (corrected) to become useful. In this case the ASR user will always post-edit the raw transcriptions, systematically generating new and useful domain-specific data. This user-generated data can be used in various ways to incrementally improve the ASR's output and the improvement would obviously return to the user as he will have to do less and less corrections over time.

If the user has the possibility to modify the internals of the ASR system, then a good choice would be to *adapt the language model* with the manually generated text. Doing so, the user will create a domain-specific ASR system, which will be much more adequate at recognizing domain-specific speech. This is the first semi-supervised scenario discussed in this paper.

However, in some cases, the ASR system is purchased as a *black-box* and the user can only attach an automatic post-editing correction block to the baseline transcription system. In this second scenario, we propose using a statistical post-editor (SPE) that is trained to correct the *systematic ASR errors*, just as Figure 1 illustrates. In statistical machine translation (SMT), a parallel corpus (composed of sentences in the source language aligned with sentences in the target language) is used to train a translation model. After the training phase, the SMT system is able to translate source language sentences into the target language. The post-editing error correction method we propose in this paper uses an SMT system that regards the raw ASR transcripts as text in the source language and the manually corrected ASR transcripts as text in the target language. Consequently, this statistical post-editor is trained on a parallel corpus (composed of raw and corrected transcripts), thus learning how to *"translate" raw transcripts into corrected transcripts*. This error correction method is based on the observation that the post-editing task has quite a repetitive nature (the ASR usually makes systematical errors). The idea of using an SMT system to correct text was used before by Simard [11-12] and Lagarda [13], but only in the context where the raw text came from a rule-based SMT system and not from an ASR system (as in our scenario).

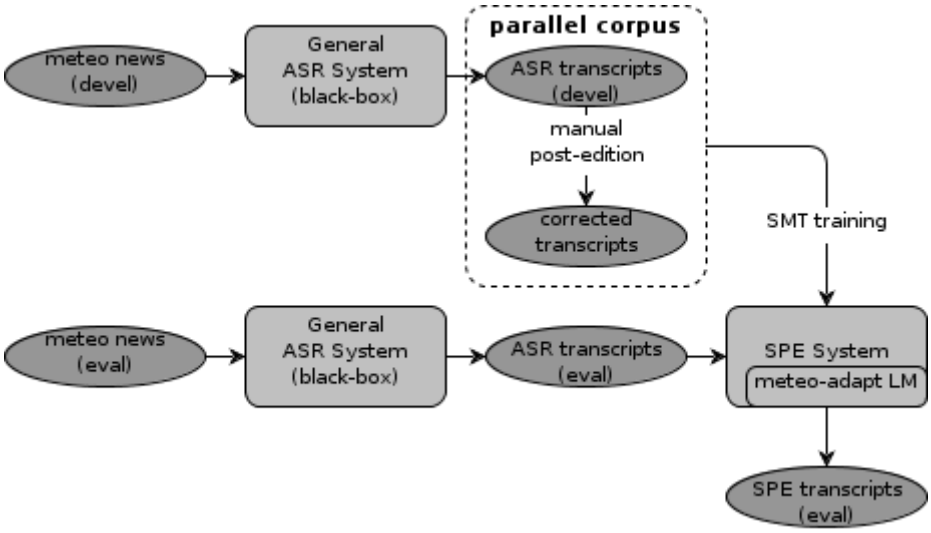

Fig. 1. Training and using the SPE system to correct ASR transcripts

4 ASR Error Correction Experiments

4.1 Experimental Setup

For all the ASR experiments presented in this work we have used the same HMM-based acoustic model [2]. This system models the 36 phonemes in Romanian in a context-dependent manner with 4000 HMM senones and 16 Gaussian mixtures per senone state [14]. The acoustic model was previously created and optimized (using the CMU Sphinx Toolkit [15]) 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 [14] and covers all the words in the LMs.

The general language model was previously created using a corpus of about 169M words collected from the Internet [2]. The domain specific language models were created using only the post-edited weather news transcripts. The domain adapted language models were created by interpolating the general language model with the domain specific language models. SRI-LM Toolkit [16] was used to create all the language models and to eventually interpolate them. The statistical machine translation system was trained using Moses SMT Toolkit [10]. The same toolkit was also used during system evaluation to automatically post-edit the raw transcriptions.

For the weather news experiments, we developed a new speech database by recording text news using three new speakers. One of these speakers recorded 2000

speech utterances to be further used for development (LM adaptation and SPE training) and all the speakers recorded 200 speech utterances each (a total of 600 utterances) to be further used for ASR evaluation. Although we admit that the evaluation database is quite small, we consider that the experimental results are conclusive and intend back them up with further experiments on a larger database.

4.2 Experimental Results

The general ASR system obtains a word error rate (WER) of 11.4% on the weather news evaluation database. This is considered to be the baseline which we want to improve by language model adaptation and statistical post-editing.

In the first experiment we consider the fully unsupervised scenario:

- the baseline ASR system is used to decode the 2000 utterances in the development database;
- the *raw transcripts* are used to adapt the general language model;
- the adapted ASR system is evaluated on the 600 utterances in the evaluation database.

Table 1 presents the adapted-ASR system results when 500, 1000, 1500 and respectively 2000 transcriptions are used for adaptation. The conclusion that emerges from this experiment is that even *unprocessed ASR data* can be successfully used to adapt the general ASR system. A relative improvement of 10.5% is significant and it is very encouraging given the relatively small amount of data used for adaptation (2000 raw sentences).

Table 1. Unsupervised LM adaptation

	# adaptation transcripts	**WER [%]**	**relative gain**
baseline ASR	0	11.4%	n/a
unsupervised adapted ASR	500	10.7%	6.1%
	1000	10.5%	7.9%
	1500	10.4%	8.8%
	2000	10.2%	10.5%

For the second experiment we exploit the errors made by the baseline system (on the development database) to create a statistical post-editing system, as follows:

- the baseline ASR system is used to decode the 2000 utterances in the development database;
- the raw transcripts are manually post-edited to create a set of corrected transcripts;
- the set of raw transcripts and the set of corrected transcripts are used to train a SPE system (as in Figure 1);
- the baseline ASR system + the additional SPE correction block is evaluated on the 600 utterances in the evaluation database (as in Figure 1).

Table 2 presents the results for this second experiment. The conclusion that emerges from this experiment is that we can obtain an important WER improvement even if we do not have access to the ASR internals, by attaching an automatic correction block as proposed above.

Table 2. Black-box ASR + SPE

	# corrected transcripts	**WER [%]**	**relative gain**
baseline ASR	0	11.4%	n/a
baseline ASR (black-box) + SPE block	500	10.7%	6.1%
	1000	10.4%	8.8%
	1500	10.2%	10.5%
	2000	10.2%	10.5%

In our third experiment we use the manually corrected transcripts to adapt the language model in the baseline ASR system. These transcripts do not contain any errors (as opposed to the unsupervised scenario) and consequently the adaptation is much more effective (see Table 3). The ASR improvement brought by this method is also more significant than the one obtained in the second experiment, but this error correction mechanism *can be applied only if the user has the means of changing the LM (the ASR is not a black-box).*

Table 3. Semi-supervised LM adaptation

	# adaptation transcripts	**WER [%]**	**relative gain**
baseline ASR	0	11.4%	n/a
semi-supervised adapted ASR	500	6.8%	40%
	1000	6.0%	47%
	1500	5.4%	53%
	2000	4.9%	57%

Finally, provided that we have access to the ASR internals, we combined the two semi-supervised methods presented above (LM adaptation and SPE correction). In this case, our experiments showed that the SPE block is left with almost nothing to correct and cannot bring any further improvements, but does not degrade the ASR performance either (same WER as in 3rd experiment).

5 Error Correction Analysis and Discussion

The ASR results obtained through unsupervised and semi-supervised LM adaption are quite easy to understand. *Unsupervised LM adaptation* boosts the probabilities of some domain-specific words and word sequences *which were correctly recognized* by the baseline ASR system. With increased LM probabilities, these items have a higher

chance to be outputted in the future (and also in the ASR evaluation phase). This is in concordance with the reality: future weather news will also contain many weather terms and phrases.

Besides the above advantage, *semi-supervised LM adaptation* benefits from several other key features, deriving from the fact that the adaptation is done with correct ASR transcripts:

- all the words in the manually corrected transcripts get a LM probability boost,
- the wrongly recognized words and phrases do not get a LM probability boost,
- many OOV words for the baseline ASR can be detected in the development phase and recovered.

Our analysis showed that the baseline ASR system lacked 315 words among the ones *uttered in the development database* (315 OOVs). A few examples are: *climatologice* (climatologically), *burniță* (drizzle), *se înnorează* (it's getting cloudy), *tunete* (thunders), *lapovița* (sleet), *consistenți* (consistent), *aversele* (the showers), etc. The OOVs detected in the development transcripts can be automatically recovered: inserted in the adapted LM and in the ASR vocabulary. This improves the overall ASR system, because many of these words were also *uttered in the evaluation database*: the 600 evaluation utterances initially had 48 OOVs, among which almost 60% were recovered through the adaptation process.

The SPE correction block manages to improve the raw ASR transcription by replacing erroneous words and word sequences with their correct counterparts. We analyzed the replacements made by this SPE block and reached several interesting conclusions. First, some of the most frequent OOVs in the evaluation database were in part corrected (1-to-1 replacements):

- *masive muntoase* (mountains) → *masivele muntoase* (the mountains): 2 corrections out of 3,
- *averse* (showers) → *aversele* (the showers): 1 correction out of 3, etc.

In the same manner (1-to-1 replacements), many other wrongly recognized words (not only OOVs) were corrected by the SPE block:

- *ceața* (the fog) → *ceață* (fog): 6 corrections,
- *noi* (we) → *norii* (the clouds): 6 corrections,
- *continua* (will continue) → *continuă* (continues): 5 corrections,
- *sînt* ([they] are, old form) → *sunt* ([they] are): 7 corrections, etc.

A second conclusion that emerged from analyzing the replacements was that, besides the 1-to-1 replacements, the SPE block also performs many-to-many replacements, as follows:

- *climatul logica/logice* (logical climate) → *climatologice* (climatologically): 6 corrections,
- *va sta la dispoziție* (will be available)→ *vă stă la dispoziție* (is available): 7 corrections,

- *nori consistent şi* (consistent clouds and) → *nori consistenți* (consistents clouds): 3 corrections.

We further analyzed the SPE translation table and found that it has learnt many other replacement rules that are potentially useful for the weather news domain, but were not needed (and consequently not applied) in this evaluation:

- *bun găsit dantelă şi domnilor* (welcome, lace and gentlemen) → *bun găsit doamnelor şi domnilor* (welcome, ladies and gentlemen)
- *vor găsi tuturor* ([they] will find everyone) → *bun găsit tuturor* (welcome everyone)
- *teama de peste o vreme la această* (the fear over a while at this [time]) → *cam atât despre vreme la această* (that's all about the weather at this [time])
- *valori la prânz între opt şi* (values at noon between eight and...) → *valori cuprinse între opt şi* (values in the interval eigth and...)
- *noi taxe vreme* (new taxes weather) → *nu uitați vremea* (don't forget the weather)

One last important conclusion is that the SPE usually learns replacement rules for phrases which have similar pronunciations. This is important, because the SPE should not change the acoustical context. However the phrase translation table also contains (wrong) rules, for which the acoustical context is not preserved at all:

- *ce* (what) → *aceste* (these)
- *cinci* (five) → *în jur de* (around)

The above observation leaves room for an important improvement for this system: the SPE rules could be filtered based on the acoustical similarity of the replacement pairs.

6 Conclusion and Future Work

Several ASR error correction methods have been proposed in the past 10-15 years and were shown to improve the speech-to-text transcription process. Most of these methods regard the ASR system as a *black-box* and propose a correction block to post-process the raw transcripts.

This paper also proposed such an ASR correction block which uses SMT principles and tools to "translate" the raw transcripts into corrected transcripts. This SPE block takes advantage of the user generated corrections, in a scenario in which the ASR user must verify and correct the transcripts to be able to actually use them. We showed that the proposed method is scalable and gets better WER results as more user-generated data is used. We also deeply analyzed the ASR errors and the SPE corrections based on the output transcriptions and the SMT phrase translation table. A key conclusion which emerged from this analysis was that the correction block makes 1-to-1 replacements in particular word contexts, but also *many-to-many replacements*.

In the near future we plan to use the N-best raw transcripts along with the manually corrected transcripts to train the SPE system (increasing the size of the training parallel corpus). Moreover, we intend to introduce another factor in the factored translation model (of the SPE system) in order to weight the translation rules based on the *acoustical similarity* of the replacement pairs. This could be done by aligning the phonetic transcription of the replacement pairs and will provide an acoustical basis for our particular "translation" scenario.

A second research direction is to evaluate the ASR correction method on various domains with different characteristics in order to see if the method is effective for broader domains, how the vocabulary richness of the domain correlates with the amount of transcriptions that need to be post-processed (corrected), etc. In the same context, changing the speech domain could also imply changing the language, because the proposed approach could be easily adapted to other languages by changing the black-box ASR system.

Finally, we also intend to investigate whether the selection of the corrected transcriptions used for SPE training has any effect on the system performance. There might be transcription subsets which generate a more effective SPE and consequently, our research goal would be to find the best selection procedure.

References

1. Yu, D., Hwang, M., Mau, P., Acero, A., Deng, L.: Unsupervised Learning from Users' Error Correction in Speech Dictation. In: 8th International Conference on Spoken Language Processing (Interspeech), pp.1969-1972, Jeju Island, Korea (2004).
2. Cucu, H., Besacier, L., Burileanu, C., Buzo, A.: Investigating the Role of Machine Translated Text in ASR Domain Adaptation: Unsupervised and Semi-supervised Methods. In: The 2011 Automatic Speech Recognition and Understanding Workshop (ASRU 2011), pp.260-265, Hawaii, USA (2011).
3. Cucu, H., Besacier, L., Burileanu, C., Buzo, A.: ASR Domain Adaptation Methods for Low-Resourced Languages: Application to Romanian Language. In: 20th European Signal Processing Conference (EUSIPCO), Bucharest, Romania (2012).
4. Ringger, E.K., and Allen, J.F.: A Fertility Channel Model for Post-Correction of Continuous Speech Recognition. In: 4th International Conference on Spoken Language Processing, vol.2, pp. 897-900, Philadelphia, USA (1996).
5. Jung, S., Jeong, M., Lee, G.G.: Speech recognition error correction using maximum entropy language model. In: 8th International Conference on Spoken Language Processing (Interspeech), pp.2137–2140, Jeju Island, Korea (2004).
6. Kaki, S., Sumita, E., Iida, H.: A method for correcting errors in speech recognition, using the statistical features of character co-occurrence. In: 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics (COLING-ACL), pp.653–657, Montreal, Canada (1998).
7. Brandow, R.L., Strzalkowski, T.: Improving speech recognition through text-based linguistic post-processing. United States Patent 6064957 (2000).
8. Mangu, L., Padmanabhan, M.: Error corrective mechanisms for speech recognition. In: 27th International Conference on Acoustics, Speech, and Signal Processing (ICASSP), vol. 1, pp.29–32, Salt Lake City, USA (2001).

9. Sarma, A., Palmer, D.D.: Context-based speech recognition error detection and correction. In: Human Language Technologies: The Annual Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL 2004), pp.85–88, Boston, USA (2004).
10. Moses Statistical Machine Translation Toolkit, http://www.statmt.org/moses
11. Simard, M., Ueffing, N., Isabelle, P., Kuhn, R.: Rule-Based Translation with Statistical Phrase-Based Post-Editing. In: 2nd Workshop on Statistical Machine Translation, Prague, Czech Republic, pp. 203–206 (2007).
12. Simard, M., Goutte, C., Isabelle, P.: Statistical Phrase-based Post-editing. In: Human Language Technologies: The Annual Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL 2007), pp. 508–515, Rochester, NY, USA (2007).
13. Lagarda, A.L., Alabau, V., Casacuberta, F., Silva R., Díaz-de-Liaño, E.: Statistical Post-Editing of a Rule-Based Machine Translation System. In: Human Language Technologies: The Annual Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL 2009), pp. 217–220, Boulder, CO, USA (2009).
14. Cucu, H., Besacier, L., Burileanu, C., Buzo, A.: Enhancing Automatic Speech Recognition for Romanian by Using Machine Translated and Web-based Text Corpora. In: The 15th International Conference SPECOM, Kazan, Russia (2011).
15. CMU-Sphinx Speech Recognition Toolkit, http://cmusphinx.sourceforge.net
16. SRI Language Modeling Toolkit, http://www.speech.sri.com/projects/srilm