IMPROVING LANGUAGE MODELS FOR ASR USING TRANSLATED IN-DOMAIN DATA

Stefan Kombrink, Tomáš Mikolov, Martin Karafiát, Lukáš Burget

Brno University of Technology

{kombrink,imikolov,karafiat,burget}@fit.vutbr.cz

ABSTRACT

Acquisition of in-domain training data to build speech recognition systems for under-resourced languages can be a costly, time-demanding and tedious process. In this work, we propose the use of machine translation to translate English transcripts of telephone speech into Czech language in order to improve a Czech CTS speech recognition system. The translated transcripts are used as additional language model training data in a scenario where the baseline language model is trained on off- and close-domain data only. We report perplexities, OOV and word error rates and examine different data sets and translators on their suitability for the described task.

Index Terms— Low Resource ASR, Language Modeling, Machine Translation

1. INTRODUCTION

There are more than 6000 actively spoken languages all across the world, and only a small fraction of them have enough data available to build ASR systems well-adapted to specific domains. Rapid development of low-resource language ASR systems often focus on dealing with a lack of acoustic modeling data. Many approaches have in common, that acoustic data is shared across languages to provide more training data [1, 2]. Often untranscribed audio data is used a semiautomatic way [3]. As far as language modeling is concerned, people have deployed automated techniques to retrieve large amounts of data from the web [4, 5]. A recent work was reported in [6] which improved language models for ASR systems by exploiting the information about structure in bilingual text data. In this work we translated English telephone speech transcripts into Czech using three common translators and used this data in our language models. Because he had only English CTS data available we used English-Czech as language pair which is considered to be one of the more difficult pairs in SMT.

When we started to develop our Czech CTS ASR system we first acquired Czech web data and some subtitle data. Even though at that time Czech could not be regarded an under-resourced language any more, a huge effort was taken to manually transcribe more than 2000 hours of telephone calls. Thus, for many under-resourced languages the lack of specific in-domain data could still be even more severe. On the other hand, the language in question may be related to at least one language with more resources available, e.g. Slovak to Czech, Dutch to English or Catalan to Spanish. Nowadays, there is a variety of online translation services available in the internet. Those make use of bilingual data resources some of which are even available for free: The OPUS corpus collection offers e.g. 27.0 million sentence pairs for English-Dutch and 3.9 million sentence pairs for Czech-Slovak. Although these translation systems are not designed to translate spontaneous speech per se, they might complement the existing off-domain language modeling data with their approximate translation results in our desired target language.

In the following section we briefly compare the role of language modeling in ASR and MT. In section 3 a description of the baseline ASR system and language models is given. Section 4 provides information about the utilized translation systems. Next, we explain how the translated CTS data was used. We report results on perplexities, OOV and word error rates. Finally, we conclude our findings in section 6.

2. LANGUAGE MODELS IN ASR AND SMT

Automated speech recognition and machine translation are related in some sense. In both tasks, a

1. hypothetical search space S is being built
2. decoding (with pruning) is performed to find the most likely hypotheses S
3. best hypothesis $\hat{s} \in S$ is found as maximizing argument of the combined estimate of two independent models:

$$\hat{s} = \arg\max_{s} P_M(X|s)P_{LM}(s), M \in \{AM, TM\}$$

In case of speech recognition X is an acoustic feature vector. The acoustic likelihood P_{AM} is estimated using the acoustic model AM. In case of machine translation X is a string sequence in the source language. The translation likelihood P_{TM} is estimated using a translation model TM which is in state-of-the art systems usually factored by several models each of which is specialized on one of the complex relationships existing between source and target language.

978-1-4673-0046-9/12/$26.00 ©2012 IEEE

In both tasks, P_{LM} is the likelihood of string s as estimated by a language model LM. However, it has to satisfy different claims. Since in SMT many errors can be fixed by reordering words, the LM is mainly used to maintain syntactical integrity. These are no properties of a single domain exclusively but an entire language and can be transferred across domains to a large extent. This may be also the reason why grammar-based systems in machine translation are still common. In large vocabulary speech recognition, on the other hand, many errors occur due to confusions between words sharing acoustically similar pronunciations. Those are defined by the utilized vocabulary which is domain dependent. Attempts to build better language models using syntactical information were not very successful up to now.

3. SETUP

3.1. Czech recognizer

Our Czech LVCSR used one lattice decoding pass and speaker adaptations. MLLR transformations were estimated from neural net based phoneme posterior decoding using critical-band features. A fast VTLN estimation using MFCC features was used to obtain per-speaker warping factors. Lattice decoding was performed using a pruned bigram language model and phoneme posterior features. Subsequently, the lattices were expanded using an unpruned trigram language model.

The acoustic models were trained discriminatively (fMPE) on almost 100 hours of Czech telephone speech data. Half of the data consisted of spontaneous speech and the other half was read speech. The test set used 2.2 hours of spontaneous speech (2606 utterances) from various telephone recordings. We used NIST word error rate scoring which maps frequent words sharing similar meaning to a unique word form. Furthermore, it neglects hesitational words and certain repetitions. Word insertion penalties and language model scales were tuned on the test set.

3.2. Baseline language models

We defined two off-domain baseline language models: The first one (*Seznam*) is using 780M words of web-data.[1] This kind of data is usually among the first being used when building an ASR system for an under-resourced language. Second, 3.8M words of movie transcripts from the OPUS corpus were added (*Seznam+Subtitles*). They contained some amount of colloquial speech. Nowadays, these data can be obtained freely for various languages.[2] Hesitations as commonly used in telephone speech transcripts did not occur.

For the interpolation of language models we used a validation set consisting of 2870 utterances of Czech spontaneous speech. Although the subtitle data is little compared to the web-data it obtained a high weight and the perplexity on both valid and test data decreased substantially (see 3). As in-domain language modeling data we used 1M words of English Switchboard transcripts (*SWB*) and 10M words of English Fisher 1+2 transcripts (*Fisher*) available from LDC.[3]

3.3. Dictionaries

In Czech language word pronunciations can be derived robustly by using a small set of rules or G2P models estimated on few example pronunciations. Thus, the creation of a pronunciation dictionary was not studied in the context of this work. Instead, 167k existing pronunciations have been used to create dictionaries for the *Seznam* and *Seznam+Subtitles* language models, respectively. The dictionary size of the *Seznam* LM was 129k and for the *Seznam+Subtitles* LM 131k pronunciations.

4. TRANSLATORS

We compared three different translators available online to produce our Czech in-domain text data from English telephone speech transcriptions. There exists a variety of publicly available web translation services, but only a few of them support less common languages like Czech.

4.1. Google Translate

Google Translate is a free online translation service. It supports 59 target and source languages and is able to translate pairwise between all of them. It is possible to upload entire documents and translate them. For spontaneous speech, we found translations of long documents incomplete and rendered into an unusable mix of English and Czech phrases and paragraphs. Well written and formatted text did not cause such troubles. The issue was solved by translating only small portions and concatenate the results, but this led to tedious manual work and allowed us to process just the Switchboard part our textual data. The Google Translate API claims to have support for 365 languages i.e. 100000 language pairs and might be more convenient to use for such a task.

4.2. Bing Translator

This web translation service offered by Microsoft supports 37 target and source languages and is able to translate between all pairs. The target group are localization customers who use the provided translation API and common users who want to translate documents and web pages which is for free. The free online translator appears to have a limit on the maximum number of words, and after a few hundred lines the translation

[1] With permissions: `http://www.seznam.cz`

[2] OPUS bilingual corpora: `http://opus.lingfil.uu.se/`

[3] Switchboard-1 transcripts: `http://www.isip.piconepress.com/projects/switchboard/`

will stop and the remaining text is kept in the source language. However, it is possible to upload text in form of HTML pages and translate all off the reel. A reasonably higher word limit is applied in that case as well such that larger documents can be split manually into smaller chunks and translated as uploaded HTML pages.

4.3. Babylon Translation Software

We downloaded the free trial of Babylon 9 translator and installed it under Windows XP. The translation progress is considerably faster than Bing. More than 800 language pairs are officially supported. After translating one larger web document our license had already expired. But the vendor offers monthly, annual and lifetime subscriptions which can be bought cheaply through the internet.

5. EXPERIMENTS

5.1. Czech CTS data

Translator	%OOV	Processed Data
Bing	1.8%	SWB and Fisher
Babylon	1.9%	SWB
Google	5.0%	SWB

Table 1. OOV rates of translators on English CTS data

In Table 1 we show how well the examined translators were able to process the Czech CTS data. Hesitations contained in the original data did not get translated and remained in their original English notation. This also happened occasionally to short phrases, proper English names and rare words for all translators, worst of all performing Google Translate. Hence, we mapped 11 frequently occurring hesitations manually to their Czech counterparts. The measured OOV rates give an impression of how "clean" the resulting translations were after the mapping was applied.

5.2. Dictionary extension

The size of the vocabulary shared between all translated data sets was around 15k words. Amongst those we were looking for new words which could possibly lower the OOV rates on the validation data. We found, that both test and validation data contained a high number of hesitations (other words could not significantly lower the OOV rate). Since the Czech CTS transcripts now contained hesitations it just seemed reasonable to add these to the dictionary as well. We extended the *Seznam+Subtitles* dictionary manually into a dictionary that could be used for all language models built upon the translated CTS transcriptions (*Seznam+Subtitles+CTS*). Table 2 shows the decrease in OOV rate using the extended dictionary.

Dictionary	%OOV Valid	%OOV Test
Seznam	7.9%	8.7%
Seznam+Subtitles	7.7%	8.6%
Seznam+Subtitles+CTS	5.3%	4.1%

Table 2. OOV rates using different dictionaries

5.3. Language Models

In Table 3 we compare eight language models. The first two listed are dedicated to the baseline language models (*Seznam, Seznam+Subtitles*) whereas the remaining ones used additional data translated from Czech CTS. Interpolation weights for all text corpuses were obtained by optimizing perplexity on the validation data set. All models were smoothed using GT except the two baseline models which used KN. In the first column, we show the weight for the added translated CTS data. The second and third column shows the validation and test data perplexities (PPL). As the weight for the successively added Czech in-domain data increases, both testing and validation data show consistent improvements up to 10% relative on the test set. The Switchboard data proved to be most effective gaining almost maximum improvement and weight. The combination of several translators (SWB All) seemed to provide slightly complementary information.

LM Data	Weight	PPL Valid	PPL Test
Seznam	-	642	650
+Subtitles	-	454	479
+SWB Google	0.25	431	449
+SWB Bing	0.28	425	440
+SWB Babylon	0.29	424	439
+SWB All	0.32	418	433
+Fisher	0.24	440	449
+SWB+Fisher	0.34	417	432

Table 3. Language model perplexities sorted by the weight of translated CTS data

5.4. Speech Recognition Results

We ran lattice decoding using a pruned bigram language model trained on all available data and the full dictionary (*Seznam+Subtitles+CTS*). By doing so we obtained a performance of 55.5% WER. Next, we rescored using the unpruned trigram versions of all LMs and obtained WER numbers.

Table 4 shows the eight language models ordered by decreasing WER. By adding the Czech CTS data we improved between 0.9% and 1.5% absolute depending on which CTS data the combined LM was using. The degradation for *Seznam* is not surprising since the rescoring LM uses less data and a smaller vocabulary.

Newly added words were hesitations without exception, and those got ignored in WER scoring. That being said, a de-

LM	Valid PPL	%WER / Change
Seznam	642	55.8 /-0.8
+Subtitles	454	55.0 / 0.0
+SWB+Fisher	417	54.1 / **0.9**
+SWB All	418	54.0 / **1.0**
+SWB Bing	425	54.0 / **1.0**
+SWB Babylon	424	53.9 / **1.1**
+SWB Google	431	53.6 / **1.4**
+Fisher	440	53.5 / **1.5**

Table 4. Language models - Recognition performance on test

creased WER could just be caused by modeling newly seen bi- and trigram contexts of existing words. Usually decreasing PPL numbers of language models on validation or test data should also indicate decreasing WER numbers. But surprisingly, correlation between WER and PPL (or amount of used training data) can not be found. Although the translated *SWB Google* data contained the highest OOV rate its LM was performing better than all other models using SWB data.

Hence we examined, how often models needed to backoff in PPL evaluation. Out of all models, the *SWB Google* LM showed the highest number of backoffs to bigrams for word tokens of test and validation data. At the same time, the *Fisher* LM used trigram estimates (no backoff) for word tokens 10% more often than all other models. Both models obviously performed better in terms of both perplexity and word accuracy. A possible explanation for *SWB Google* is that due to the limited amount of data bigram estimates have been considerably more reliable than the trigram ones as opposed to *Fisher* which used ten times more data and provides more reliable trigram estimates.

Both models do not perform well in combination, and we observed, that the number of back-offs for *SWB+Fisher* is higher than for any of the two underlying models. Also mixing SWB data translated by different translators did not help to decrease WER. One possible explanation is that some n-gram estimates are spoiled by repeated reordering errors which occurred across translators.

6. CONCLUSION

We showed that translated transcripts can actually not only improve PPL of our baseline language models but also decrease WER in speech recognition. We did so by using translators on English in-domain data equipped with external language models which supposedly were trained on much more Czech target language data than what was available to us. Yet it is probable that no in-domain data was used at all for building the models of the translators.

Despite perplexities lacking correlation with word error rates, using the translated in-domain data got high weight and decreased WER by 1.5% absolute. Furthermore, the following reasons keep us optimistic about the approach:

- English-Czech is one of the harder language pairs in machine translation. Other language pairs may yield better improvement.
- Improvement came from newly added bi- and trigram contexts. No new words got modeled, just hesitations, and those got neglected in WER scoring.
- This experiment can be repeated easily if a suitable translator for the desired target language is found.

Thus, we propose to compare the performance of different language pairs in future research. Adding new words (not just hesitations) should also lead to more improvement. The issue that WER did not correlate with PPL should get examined further and eventually be resolved. Deploying a self-made SMT system (where possibly even the LM data is shared with the baseline ASR) could provide a more constrained setting, lead to more insight and bring up a more sophisticated solution.

7. ACKNOWLEDGEMENTS

This work was partly supported by Czech Ministry of Trade and Commerce project No. FR-TI1/034, Technology Agency of the Czech Republic grant No. TA01011328, and Czech Ministry of Education project No. MSM0021630528. M. Karafiat was supported by Grant Agency of the Czech Republic post-doctoral project No. P202/12/P400

8. REFERENCES

[1] Ngoc Thang Vu, Franziska Kraus, and Tanja Schultz, “Cross-language bootstrapping based on completely unsupervised training using multilingual a-stabil.,” in *Proc. ICASSP*. 2011, pp. 5000–5003, IEEE.

[2] Lukáš Burget et al, “Multilingual acoustic modeling for speech recognition based on subspace gaussian mixture models,” in *Proc. ICASSP*. 2010, vol. 2010, pp. 4334–4337, IEEE Signal Processing Society.

[3] Scott Novotney, Rich Schwartz, and Sanjeev Khudanpur, “Unsupervised arabic dialect adaptation with self-training,” in *INTERSPEECH*, 2011.

[4] R. Sarikaya, Hong-Kwang Kuo, and Yuqing Gao, “Impact of web based language modeling on speech understanding,” in *Proc. ASRU*, Nov. 2005, pp. 268 –271.

[5] Ngoc Thang Vu, Tim Schlippe, Franziska Kraus, and Tanja Schultz, “Rapid bootstrapping of five eastern european languages using the rapid language adaptation toolkit,” in *INTERSPEECH*, 2010, pp. 865–868.

[6] Sebastian Stüker, Laurent Besacier, and Alex Waibel, “Human translations guided language discovery for ASR systems,” in *INTERSPEECH*, 2009, pp. 3023–3026.