Speech Translation for Low-Resource Languages: The Case of Pashto

Andreas Kathol, Kristin Precoda, Dimitra Vergyri, Wen Wang, Susanne Riehemann

SRI International, Menlo Park, CA 94025, USA
kathol@ai.sri.com

Abstract

We present a number of challenges and solutions that have arisen in the development of a speech translation system for American English and Pashto, highlighting those specific to a very low resource language. In particular, we address issues posed by Pashto in the areas of written representation, corpus creation, speech recognition, speech synthesis, and grammar development for translation.

1. Introduction

We discuss a number of challenges and solutions that have arisen in the development of a speech translation system for American English and Pashto, one of the major languages of Afghanistan, under the auspices of DARPA's CAST program. Similar systems have been built for other language pairs (e.g. [1,2,3,4]), and what we will emphasize here are specific issues and solutions required by a very low resource language. While some of these may be unique to Pashto, others illustrate issues common to low-resource languages in general.

The intended domain of application of the system is medical exchanges, in particular first encounters between a patient and a medical professional. The prototype system integrates two speech recognizers (one for each language), two parser/generators, two speech synthesizers, and a user interface that is both user friendly and flexible enough to handle various types of errors. All components have been designed for a handheld target platform. (For an overview of the system see [5].)

2. Pashto data creation and collection

By far the greatest challenge in the project has been the lack of Pashto resources of all sorts. This made it necessary for us to cover much more ground than is customary in speech/language technology projects. In particular, we needed to establish an orthography as well as create all corpora to be used for any purpose.

2.1. Orthography issues

The single greatest challenge to all aspects of the project has been the lack of any standardized writing system or spelling norms and consequent writing of one word multiple ways, and different words the same way. This complicates any computational processing that relies on string comparisons, in other words all computational processing, as any given string may not have a unique interpretation.

Before taking on the substantial task of creating needed corpora, we had to work out much linguistic analysis (in particular phonemic analysis), which usually can be taken for granted in the case of better-studied languages. Initially, we trained our language consultants in phonological awareness and tried to transcribe acoustic data directly into a phonemic representation to avoid the nonstandardization of the native orthography. We found however that awareness of the phonemes was quite difficult for speakers to achieve, and made still more difficult by the fact that our analysis was intended to cover a broad range of Pashto dialects and thus was not descriptive of any one speaker's inventory. In addition, our phonological analysis evolved somewhat over time as we became aware of additional phenomena. We therefore began transcribing in native (Arabic-based) script instead, as it was independent of the phonological analysis, and found it to be more reliable despite its difficulties. Words in native script were then phonemically transcribed. In many cases multiple phonemic representations were associated with each script form. These representations may reflect widely differing pronunciations, or entirely different words written the same way, or variant but equivalent syntactic forms. Likewise, any given phonemic representation may be associated with one or several script forms, with one or several meanings intended.

Speech recognition processing used an isomorphism of the native script, so that texts could be used for language modeling. A "primary" pronunciation for each distinct meaning was manually identified for input to the translation component and synthesis engine. Despite our best efforts at verification, maintaining consistency across multiple forms of the same word remained a challenge. In addition, even using only the primary pronunciations produced considerable ambiguity for the translation component.

2.2. Corpora for speech recognition

The only preexisting corpus of recorded Pashto we could fine was a series of untranscribed Voice of America Pashto service broadcasts, recorded by the Linguistic Data Consortium from the broadcasts. As it was readily available, we manually transcribed about 5 hours of this data. However, the data had several problems: the broadcasts were dominated by fewer than a dozen speakers, showed only a fraction of the dialect diversity of Pashto, and were not of very high audio quality, nor did they show a dialog speech style. We therefore recruited approximately 80 regionally diverse Pashtuns from a local émigré community, and asked each to record 100-200 spontaneously generated utterances, including answers to questions. These recordings totaled about 7 hours of speech. The acoustic model training data consisted of these 12 hours of speech, or approximately 100,000 words. The same data was used for language model training, excluding about a third of the VOA data that could not be used for language model purposes because it was not transcribed in Pashto script. Additional in-domain text was obtained by translating English role-played dialogs, yielding a total of 270,000 words of language modeling training data. The vocabulary totaled 6680 different forms.

2.3. Corpus for grammar development

To be most useful in the development of linguistically sound grammatical analyses, a Pashto sentence must be translated into idiomatic English and glossed with both literal meaning and basic morphosyntactic information, for example:

dzmA	pIns@l	p@kAr	day
my	pencil.S.D	need.S.D	be.3S.M

'I need a pencil.'

Such glosses are an invaluable source of information about the morphological classes that are actually used in speech, as there are many words that can belong to different morphological classes, depending on the speaker (and can even vary within a speaker).

The glossed translation corpus includes about 4800 sentences. The Pashto lexicon, built primarily from these sentences, comprises approximately 5000 citation forms.

3. Pashto speech recognition

The development of the Pashto automatic speech recognizer was guided by three factors: the amount of available training data, problems associated with Pashto orthography and the requirements for close to real-time performance on a platform with limited memory.

3.1. Test sets

Two test sets were defined. Test set 1 comprised 5 dialectally diverse speakers and 5128 words, and was not a close match to the medical dialogs that formed the bulk of the language model data. 4% of the words in this test set were out of vocabulary (OOV). Test set 2, with 4 Eastern Afghan Pashto speakers and 3409 words, was a fairly good match to the medical dialog domain and had a 1.6% OOV rate.

3.2. Acoustic modeling

Our system used a front end with a 16 kHz sampling rate, 10 ms frame advance rate, and mel frequency cepstral coefficients (13 coefficients plus first- and second-order differences). Using an inventory of 43 phonemes, we trained 3-state triphone hidden Markov models (HMMs) of fairly small size to fit the limitations of the anticipated small-footprint platform. We trained phone-state tied mixture models (129 phone state Gaussian clusters) with 32 Gaussians each. We also compared the model of this size with a bigger model which used decision tree state clustering with 350 state clusters and 64 Gaussians per cluster, and we found this model to be worse, with a 10% relative performance reduction on test set 1. The models were trained using maximum likelihood estimation followed by discriminative maximum mutual information estimation (MMIE, [6]). Discriminative training provided a 4-7% relative word error rate (WER) improvement when using the small models on the different test sets.

3.3. Language modeling

Because of the morphological complexity of Pashto and the small amount of available training data, language modeling posed a serious challenge. We addressed the problem by adapting the algorithm presented in [7], and built a language model that had more fine-grained backoff layers than a traditional word n-gram language model. To achieve this, we first generated a clustering tree for the vocabulary with the root of the tree representing the whole vocabulary and every node representing a class that includes all words in its descendant nodes. The tree is generated using the minimum discriminative information clustering algorithm using a similarity metric based on the left and right contexts of a word. When estimating the conditional probability of a word based on its n-gram prefix, we first back off to its context with the most distant word replaced by its class, from the most specific to the most general, and if none of these backoffs could guarantee a minimum number of occurrences then back off to the normal lower-order (n-1)-gram prefix. The resulting language model achieves a relative perplexity reduction of over 10% and a significant word error rate reduction on the different test sets as shown in Table 1.

	Test set 1	Test set 2
Standard trigram	43.5	35.1
Hierarchical trigram	37.9	31.2

Table 1: Word error rate, in percent. Comparison of standard and hierarchical language model used in combination with MMIE trained acoustic models.

3.4. Evaluation of speech recognition accuracy

Because of the nonstandardized Pashto orthography the evaluation of recognition accuracy has been problematic. A word may be written in different ways and certain word boundaries are not well defined. The results in Table 1 are computed using the standard definition of WER where the recognition hypothesis is compared to a single reference. To address the problem of variable orthographies, we modified the WER calculation procedure by changing the reference to assign word boundaries flexibly, and by counting different spellings of the same word as equally correct. We then calculated possible WER values for a small test set including only a single, Eastern Afghan speaker and 272 words. With the standard WER calculation which assumes an uncontroversial orthography, the WER was approximately 21%. Given exactly the same recognizer output and changing only the purely orthographic points described just above, the error rate ranges from 11.4% to 29.7%. This demonstrates the need to reflect carefully on the evaluation procedure and, if appropriate, adapt it to the challenges presented by essentially unwritten languages.

3.5. Use of untranscribed data

Since we had very little transcribed training data, we began to explore ways of using untranscribed data (see also [8]). We used our recognizer to automatically transcribe 25 hours of newly collected data, then discarded about 7 hours of data recognized with low confidence and used the remaining data to retrain the acoustic model (cf. [9]). The results were not significantly different from the original model. This approach needs further investigation to overcome the problem of the high error rate on the new data, or to explore ways of selecting new data to be manually transcribed in order to benefit such a limited resources system.

4. Speech Synthesis

The development of the Pashto speech synthesis component has proceeded in collaboration with Cepstral LLC, and has largely used the same methods as other languages. Worth mentioning is our initial difficulty finding a suitable speaker, as many speakers (of Pashto and of other languages) are sensitive to the possible uses to which a synthetic voice sounding like their own voice could be put. We therefore first planned to base the Pashto voice on an English speaker mimicking native-speaker Pashto recordings, to create a voice that could be used without any restrictions.

This sensitivity around voice usage may hold for many relatively small language communities. For this system, however, we were eventually able to locate a native Pashto speaker of the targeted accent region who was willing to allow his voice to be used for the program's purposes. Crucially, better-quality synthesis was achieved more quickly when using a native Pashto speaker for the diphone database, than using the mimicked productions.

5. Pashto grammar component

Despite the current popularity of statistical approaches to machine translation, we opted for a knowledge-based approach based on hand-coded grammatical rules and lexical entries. There are several reasons for this decision. One of these is that the creation of Pashto data sufficient for the application of statistical approaches is prohibitively difficult. Another is that rule-based translation can be ported relatively easily to new domains, as only new vocabulary items should need to be added and the basic grammatical framework should remain largely constant.

5.1. Gemini-based translation

As the particular framework for the development of the translation component we used SRI's Gemini natural language engine ([10]). Among the features that make Gemini attractive for a bidirectional translation application is its ability to use the same grammar for either parsing or generation. In parsing mode, Gemini takes an input string and produces a semantic representation in quasi-logical form (QLF) format representing the denoted event (or state) with its major participants and modificational information. To generate, it applies the grammar "in reverse", producing a surface string from a QLF representation. While the QLF format used here was originally designed for automated reasoning tasks, we have successfully used it as an interlingua representation with only minor modifications. Since the QLF format is language independent, the translation process simply requires the successful passing of QLF information between two Gemini processes, one for each language. As there has been extensive prior work on a large Gemini grammar of English, the development of the translation component has consisted for the most part of creating a Gemini grammar of Pashto, including a lexicon.

5.2. Some challenges of Pashto grammar

The conceptual simplicity of the approach contrasts with the reality of developing a phrase structure grammar for a language whose morphological and syntactic structure is considerably different from English.

5.2.1. Morphology

Major categories in Pashto have many inflectional forms. Nouns and adjectives are distinguished for case, number, and (in the case of adjectives) gender. The sets of inflectional forms are organized into declensional classes ("paradigms") and a variety of subclasses, which must be reflected in Gemini's lexicon. Of note is the degree to which available linguistic resources failed to capture the full amount of observed morphological variation. In addition, nouns are not necessarily the same class or gender for different speakers, and occasionally there is even variability within a speaker.

While native speakers have intuitions about the forms of particular words, identifying the appropriate declensional class is very difficult. We attempted to automate this process by prompting the native speaker with particular forms that help identify the applicable paradigm. As long as the forms offered by the native speaker consultant are consistent and the set of paradigms is complete, the paradigm identification task can be partially automated. Cases in which the proposed forms do not coincide well with speakers' intuitions often point to the need to postulate previously unrecognized morphological declension (sub)classes. In some cases, the difficulty seems to lie in variable intuitions provided on different occasions by a single native speaker. We interpret this phenomenon as providing some counterevidence to the notion of paradigms as well defined and static, or at least to the reliability of native speaker intuitions. This variability holds across speakers as well. Verifying all paradigm information against observed data and not relying more heavily than necessary on native speaker intuitions is probably the best path to pursue, even in the face of extremely limited corpora.

5.2.2. Syntax

Typologically, Pashto is a head-final, split-ergative language with some degree of word order freedom. These and other properties required a considerable amount of effort to capture in Gemini's phrase structure based rule format. The initial approach of taking English rules as starting points and adapting them to the requirements of Pashto often proved to be of limited use. For instance, there is no equivalent in English to the phenomenon of "second-position clitics" ([11]), shown in boldface below:

z@	**ba**	z@r	rAS@m
I	FUT	soon	PFX-PERF.go-1S

'I *will* come soon.'

What makes second-position clitics especially challenging is their semantic heterogeneity. Clitics can provide information about tense, modality, and verbal arguments (object or subject, depending on tense) or denote the possessor of a preceding or following phrase. Finally, multiple clitics may co-occur, each making a different semantic contribution. Our approach to such complexities was driven by practical considerations and favored a series of simple robust solutions for a number of frequent subcases over a possibly more elegant but less robust single-rule analysis.

It further became clear that the initial model of working out both grammars in isolation was not very practical. In our approach, translatability requires that the target grammar can generate a string from the QLF produced by the source grammar. Any QLF mismatches can be resolved by revising the rules on the target grammar side. On the other hand, for practical purposes, it is often much simpler to adjust the source language grammar so that its QLFs are such that the target grammar can successfully generate from them. Thus, in developing the Pashto grammar numerous changes were also made to the English grammar.

In some cases expressions in the two languages are so far apart that adapting the source language grammar was impractical for achieving translation compatibility. In such cases a transfer component maps the original source language QLF onto a target language QLF that has the desired generation behavior. One example of such a highly divergent translation pair is the following:

zrr@	de	rAjigigxi
heart-S.D	2POSS	PFX-rise-3S

‘Are you nauseous?’ (literally ‘Is your heart rising?’)

While the phrase structure-based format of Gemini rules may perhaps be less well suited to Pashto than to other languages, it has offered enough flexibility for approximate solutions to many grammatical issues. We are therefore optimistic that such an approach is also applicable to other languages and that rule-based grammars remain a viable alternative to more data-intensive approaches.

6. Concluding remarks

Many of the challenges posed by Pashto are similar to those of other low-resource languages (and high-resource languages, for that matter). Two strategies employed in this project were resource creation and exploitation of knowledge resources, such as reference grammars, dictionaries, and native speaker insights. Some of the most basic, often overlooked kinds of fundamental resources upon which much is built include though certainly are not limited to:

- consistent orthographic conventions
- existence of a language standard (e.g. as used in broadcasting or education)
- previous experience with the language, including speakers who are technologists, native speaker expertise or access to native speaker insights
- depth, breadth, quality and reliability of existing linguistic knowledge resources

When some of these are missing, it becomes increasingly difficult to build a speech technology system using standard techniques, and innovative, situation-specific approaches may be required.

7. Acknowledgments

This material is based upon work supported by the Defense Advanced Research Projects Agency (DARPA) under Contract Nos. NBCHD040058 and N66001-99-D-8504. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the U.S. Government. Many thanks are due to Chris Culy, Ascander Dost, Horacio Franco, Mike Frandsen, John Fry, Mark Gawron, Mohammad Shahabuddin Khan, Robert J. Podesva, Colleen Richey, Valerie Wagner, and the many Pashto speakers who supported us.

8. References

[1] Black, A., Brown, R., Frederking, R., Singh R., Moody, J., and Steinbrecher, E. “TONGUES: Rapid development of a speech-to-speech translation system,” *Proc. HLT 2002*, pp. 24–27, 2002.

[2] Schultz. T., Alexander, D., Black, A., Petersen, K., Suebvisai, S., and Waibel, A. “A Thai speech translation system for medical dialogs,” *Proc. HLT/NAACL 2004 Demonstrations*, pp. 34–35, 2004.

[3] Zhou, B., Dechelotte, D., and Gao, Y. “Two-way speech-to-speech translation on handheld devices,” *Proc. INTERSPEECH 2004*, pp. 1637–1640, 2004.

[4] Narayanan, S., Ananthakrishnan, S., Belvin, R., Ettaile, E., Gandhe, S., Ganjavi, S., P. G. Georgiou, G., Hein, C. M., Kadambe, S., Knight, K., Marcu, D., Neely, H. E., Srinivasamurthy, N., Traum, D., and Wang, D. “The Transonics Spoken Dialogue Translator: An aid for English-Persian doctor-patient interviews,” *Working Notes of the AAAI Fall symposium on Dialogue Systems for Health Communication*, 2004.

[5] Precoda, K., Franco, H., Dost, A., Frandsen, M., Fry, J., Kathol, A., Richey, C., Riehemann, S., Vergyri, D., Zheng, J., and Culy, C. “Limited-domain speech-to-speech translation between English and Pashto,” *Proc. HLT/NAACL 2004 Demonstrations*, pp. 9–12, 2004.

[6] Zheng, J., Butzberger, J., Franco, H., and Stolcke, A. “Improved maximum mutual information estimation training of continuous density HMMs,” *Proc. EUROSPEECH 2001*, pp. 679–682, 2001.

[7] Zitouni, I. and Kuo. H. J. “Effectiveness of the backoff hierarchical class n-gram language models to model unseen events in speech recognition,” *Proc. of IEEE Automatic Speech Recognition and Understanding*, 2003.

[8] Wessel, F. and Ney, H. “Unsupervised training of acoustic models for large vocabulary continuous speech recognition,” *IEEE Transactions on Speech and Audio Processing*, pp. 23–31, 2005.

[9] Hakkani-Tür, D., Ricardi, G. and Gorin, A. “Active learning for automatic speech recognition,” *Proc. ICASSP 2002*, pp. 3904–3907, 2002.

[10] Dowding, J., Gawron, J. M., Appelt, D. E., Bear, J., Cherny, L., Moore, R., and Moran, D. B. “Gemini: A natural language system for spoken language understanding,” *Proc. ACL 1993*, pp. 54–61, 1993.

[11] Roberts, T. *Clitics and Agreement.* Doctoral dissertation, MIT, 2000.