CROSS-LINGUAL TOPIC PREDICTION FOR SPEECH USING TRANSLATIONS

Sameer Bansal[1], *Herman Kamper*[2], *Adam Lopez*[1], *Sharon Goldwater*[1]

[1]School of Informatics, University of Edinburgh, UK
[2]Dept. E&E Engineering Stellenbosch University, South Africa
{sameer.bansal, sgwater, alopez}@inf.ed.ac.uk
kamperh@sun.ac.za

ABSTRACT

Given a large amount of unannotated speech in a low-resource language, can we classify the speech utterances by topic? We consider this question in the setting where a small amount of speech in the low-resource language is paired with text translations in a high-resource language. We develop an effective cross-lingual topic classifier by training on just 20 hours of translated speech, using a recent model for direct speech-to-text translation. While the translations are poor, they are still good enough to correctly classify the topic of 1-minute speech segments over 70% of the time—a 20% improvement over a majority-class baseline. Such a system could be useful for humanitarian applications like crisis response, where incoming speech in a foreign low-resource language must be quickly assessed for further action.

Index Terms— speech translation, low-resource speech processing, speech classification, unwritten languages

1. INTRODUCTION

Quickly making sense of large amounts of linguistic data is an important application of language technology. For example, after the 2011 Japanese tsunami, natural language processing was used to quickly filter social media streams for messages about the safety of individuals, and to populate a person finder database [1]. Japanese text is high-resource, but there are many cases where it would be useful to make sense of *speech* in *low-resource* languages. For example, in Uganda, as in many parts of the world, the primary source of news is local radio stations, which is broadcast in many languages. A pilot study from the United Nations Global Pulse Lab identified these radio stations as a potentially useful source of information about a variety of urgent topics related to refugees, small-scale disasters, disease outbreaks, and healthcare [2, 3]. With many radio broadcasts coming in simultaneously, even simple classification of speech for known topics would be helpful to decision-makers working on humanitarian projects.

Speech classification systems have traditionally used automatic speech recognition (ASR) systems to first convert speech to text, which is then used as input to a classifier. However, this pipelined approach is impractical for unwritten languages, spoken by millions of people around the world. Although transcriptions cannot be obtained in these settings, translations could provide a viable alternative supervision source [4–7]. Recent research has shown that it is possible to train direct **S**peech-to-text **T**ranslation (ST) systems from speech paired only with translations [8–10]. Since no transcription is required, this is useful in very low-resource settings. However, in realistic low-resource settings where only a few hours of training data is available, these end-to-end ST systems produce poor translations [11]. But it has long been recognized that there are good uses for bad translations [12]. Could classifying the original speech be another one of these use cases?

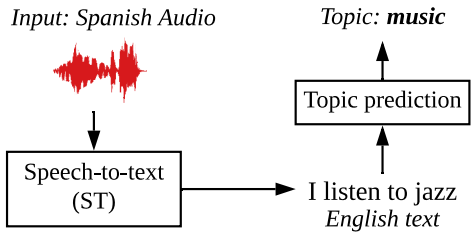

Fig. 1. Spanish speech is translated to English text, and a classifier then predicts its topic.

We answer this question affirmatively: we first use ST to translate speech to text, which we then classify by topic using supervised models (Figure 1). Although our ultimate goal is to work with truly low-resource languages, available datasets of this type are still too small to thoroughly evaluate and analyse. We therefore test our method on a corpus of conversational Spanish speech paired with English text translations that has been widely used in ST research [9, 13], enabling us to put our results in context. Using an ST model trained on 20 hours of Spanish-English data, we predict topics correctly 71% of the time, and we outperform the majority class baseline with less than 10 hours of training data. These promising results are the first we know of for this task, and open the door to future work on cross-lingual topic prediction from speech.

2. METHODS

Speech-to-text translation. We use the method of Bansal et al. [11] to train neural sequence-to-sequence Spanish-English ST models. As in that study, before training ST, we pre-train the models using English ASR data from the Switchboard Telephone speech corpus [14], which consists of around 300 hours of English speech and transcripts. In [11] this was found to substantially improve translation quality when the training set for ST was only tens of hours.

Topic modeling and classification. To classify the translated documents, we first need a set of topic labels, which were not already available for our dataset. We therefore initially discover a set of topics from the target-language (English) training text using a topic model. To classify the translations of the test data, we choose the most probable topic according to the learned topic model. To train our topic model, we use Nonnegative Matrix Factorization (NMF) [15, 16].We also experimented with Latent Dirichlet Allocation [17], but manual inspection revealed that NMF produced better topics.

3. EXPERIMENTAL SETUP

Data. We use the Fisher Spanish speech corpus [18], which consists of 819 phone calls, with an average duration of 12 minutes, giving a total of 160 hours of data. We discard the associated transcripts and pair the speech with English translations [19]. To simulate a low-resource scenario, we sampled 90 calls (20h) of data (*train20h*) to train both ST and topic models, reserving 450 calls (100h) to evaluate topic models (*eval100h*). We investigate ST models of varying quality, so we also trained models with decreasing amounts of data: *ST-10h*, *ST-5h*, and *ST-2.5h* are trained on 10, 5, and 2.5 hours of data, respectively, sampled from *train20h*. To evaluate ST only, we use the designated Fisher test set, as in previous work.

Fine-grained topic analysis. In the Fisher protocol, callers were prompted with one of 25 possible topics. It would seem appealing to use the prompts as topic labels, but we observed that many conversations quickly departed from the initial prompt and meandered from topic to topic. For example, one call starts: "Ok today's topic is marriage or we can talk about anything else ...." Within minutes, the topic shifts to jobs: "I'm working oh I do tattoos." To isolate different topics within a single call, we split each call into 1-minute long segments to use as 'documents'. This gives 1K training and 5.5K test segments, but leaves us with no human-annotated topic labels for them.

Obtaining gold topic labels for our data would require substantial manual annotation, so we instead use the human translations from the 1K (*train20h*) training set utterances to train the NMF topic model with *scikit-learn* [20], and then use this model to infer topics on the evaluation set. These *silver* topics act as an oracle: they tell us what a topic model would infer if it had perfect translations.

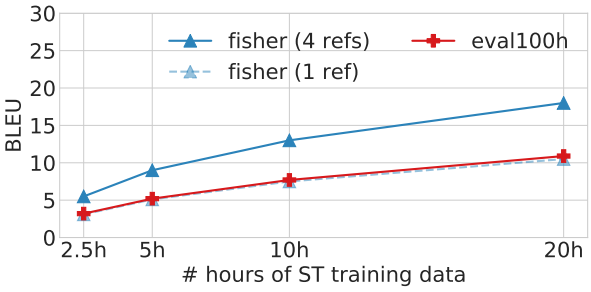

Fig. 2. BLEU scores for Spanish-English ST models computed on Fisher test set, using all 4 human references available, and using only 1 reference, and on *eval100h*, for which we have only 1 human reference.

To evaluate our ST models, we apply our ST model to test audio, and then predict topics from the translations using the NMF model trained on the human translations of the training data (Figure 1). To report accuracy we compare the predicted labels and silver labels, i.e., we ask whether the topic inferred from our predicted translation (ST) agrees with one inferred from a gold translation (human).

4. RESULTS

Spanish-English ST. To put our topic modeling results in context, we first report ST results. Figure 2 plots the BLEU scores on the Fisher test set and on *eval100h* for Spanish-English ST models. The scores are very similar for both sets when computed using a single human reference; scores are 8 points higher on the Fisher test set if all 4 of its available references are used. The state-of-the-art BLEU score on the Fisher test set is 47.3 (using 4 references), reported by [9], who trained an ST model on the entire 160 hours of data in the Fisher training corpus. By contrast, our 20 hour model (*ST-20h*) achieves a BLEU score of 18.1. Examining the translations (Table 1), we see that while they are mediocre, they contain words that might enable correct topic classification.

Topic modeling on training data. Turning to our main task of classification, we first review the set of topics discovered from the human translations of *train20h* (Table 2). We explored different numbers of topics, and chose 10 after reviewing the results. We assigned a name to each topic after manually reviewing the most informative terms; for topics with less coherent informative terms, we include *misc* in their names.

For evaluation, silver labels are obtained by applying this topic model to human translations on the test data. We argued above that the silver labels are sensible for evaluation despite not always matching the assigned call topic prompts, since they indicate what an automatic topic classifier would predict given correct translations and they capture finer-grained changes in topic. Table 3 shows a few examples where the silver

audio	yo eh oigo la música en inglés o americana
human	i eh listen to music in english or american
ST	i eh listen to the music in english
topic	*music*
audio	soy católica pero no en realidad casi no voy a la iglesia
human	i am catholic but actually i hardly go to church
ST	i'm catholics but reality i don't go to the church
topic	*religion*

Table 1. Examples of Spanish **audio** shown as Spanish text. An **ST** system translates the audio into English text, and we give the **human** reference. Our task is to predict the **topic** of discussion in the audio, which are potentially signaled by the underlined words.

Topic	Most informative terms
family-misc	married, kids, huh, love, three
music	music, listen, dance, listening, hear
intro-misc	hello, fine, name, hi, york
religion	religion, god, religions, believe, bible
movies-tv	movies, movie, watch, theater
welfare	insurance, money, pay, expensive
languages-misc	english, spanish, speak, learn
tech-marketing	phone, cell, computer, call, number
dating	internet, met, old, dating, someone
politics	power, world, positive, china, agree

Table 2. Topics discovered using human translated text from *train20h*, with manually-assigned topic names.

labels differ from the assigned call topic prompts. In the first example, the topic model was arguably incorrect, failing to pick up the prompt *juries*, and instead focusing on the other words, predicting *intro-misc*. But in the other examples the topic model is reasonable, correctly identifying the topic in the third example where the transcripts indicate that the annotation was wrong (specifying the topic prompt as *music*). In general, the topic model classifies a large proportion of discussions as *intro-misc* (typically at the start of the call) and *family-misc* (often where the callers stray from their assigned topic).

Our analysis also supports our observation that discussed topics stray from the prompted topic in most speech segments. For example, among segments in the 17 training data calls with the prompt *religion*, only 36% have the silver label *religion*, and the most frequently assigned label is *family-misc* (46%).

Topic classification on test data. We have four ST model translations: *ST-2.5h, 5h, 10h, 20h* (in increasing order of quality). We feed each each of the audio utterences in *eval100h* into the topic model from Table 2 to get the topic distribution and use the highest scoring topic as the predicted label.

human translation	Assigned	Silver
hello good afternoon have you ever been in a jury in a trial	juries	intro-misc
i also receive many letters of life insurance from banks	spam	welfare
they tell us we have to talk about marriage	music	family-misc

Table 3. Example audio utterances from *eval100h*. We show a part of the human translation here. **Assigned** is the topic assigned to speakers in the current call to prompt discussion. **Silver** is topic inferred by feeding the human translation through the topic model.

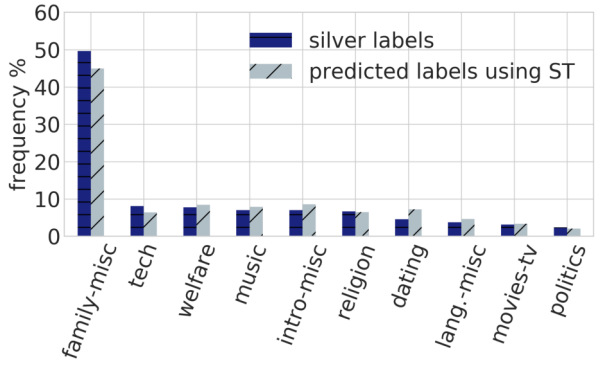

Fig. 3. Distribution of topics predicted for the 5K audio utterances in *eval100h*. **silver** labels are predicted using human translations. The **ST** model has been trained on 20 hours of Spanish-English data.

Figure 3 compares the frequencies of the silver labels with the predictions from the *ST-20h* model. The *family-misc* topic is predicted most often—almost 50% of the time. This is reasonable since this topic includes words associated with small-talk. Other topics such as *music*, *religion* and *welfare* also occur with a high enough frequency to allow for a reasonable evaluation.

Figure 4 shows the accuracy for all ST models, treating the silver topic labels as the correct topics. We use the *family-misc* topic as a majority class *naive baseline*, giving an accuracy of 49.6%. We observe that ST models trained on 10 hours or more of data outperform the *naive-baseline* by more than 10% absolute, with *ST-20h* scoring 71.8% and *ST-10h* scoring 61.6%. Those trained on less than 5 hours of data score close to or below that of the naive baseline: 51% for *ST-5h* and 48% for *ST-2.5h*.

Since topics vary in frequency, we look at label-specific accuracy to see if the ST models are simply predicting frequent topics correctly. Figure 5 shows a normalized confusion matrix for the *ST-20h* model. Each row sums to 100%, repre-

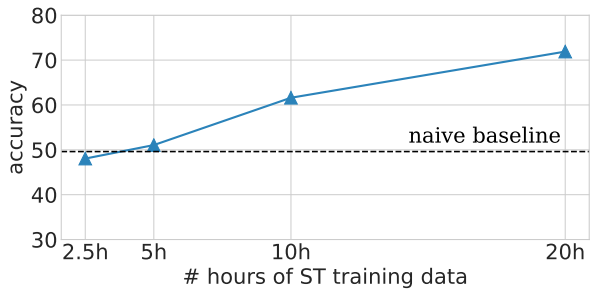

Fig. 4. Accuracy of topic prediction using ST model output. The **naive baseline** is calculated using majority class prediction, which is the topic *family-misc*.

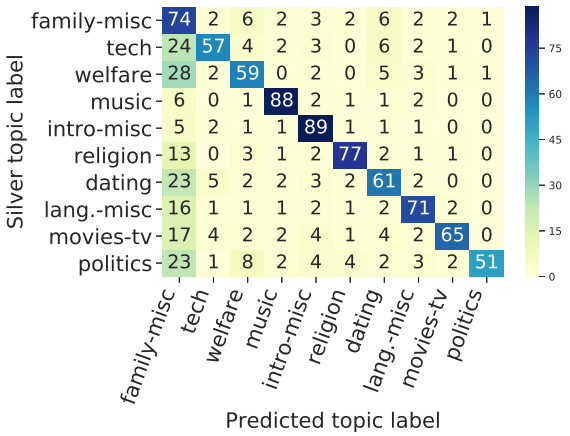

Fig. 5. Confusion matrix for ST model trained on 20 hours of Spanish-English data. Each cell represents the percentage of the silver topic labels predicted as the x-axis label, with each row summing to 100%.

senting the distribution of predicted topics for any given silver topic, so the numbers on the diagonal can be interpreted as the topic-wise recall. For example, a prediction of *music* recalls 88% of the relevant speech segments. We see that the model has a recall of more than 50% for all 10 topics, making it quite effective for our motivating task. The *family-misc* topic (capturing small-talk) is often predicted when other silver topics are present, with, for instance, 23% of the silver *dating* topics predicted as *family-misc*.

5. RELATED WORK

We have shown that low-quality ST can be useful for speech classification. Previous work has also looked at speech analysis without high-quality ASR. In a task quite related to ours, [21] showed how to cluster speech segments in a completely unsupervised way. In contrast, we learn to classify speech using supervision, but what is important about our result is it shows that a small amount of supervision goes a long way.

A slightly different approach to quickly analyse speech, is the established task of *keyword spotting*, which asks whether any of a specific set of keywords appears in a segment [22, 23]. Recent studies have extended the early work to end-to-end keyword spotting [24,25] and to semantic keyword retrieval, where non-exact but relevant keyword matches are retrieved [26–28]. In all these studies, the query and search languages are the same, while we consider the cross-lingual case.

There has been some limited work on cross-lingual keyword spotting. [29] introduced a baseline system which combined ASR and text translation to build a German speech retrieval system using French text queries. But source language transcriptions to train ASR are unlikely to be available in our scenarios of interest. Some recent studies have attempted to use vision as a complementary modality to do cross-lingual retrieval [30,31]. However, to the best of our knowledge, cross-lingual topic classification for speech has not been considered elsewhere.

6. CONCLUSIONS AND FUTURE WORK

Our results show that poor speech translation can still be useful for speech classification in low-resource settings. By varying the amount of training data, we found that ST systems trained on as little as 10 hours (around 8K parallel utterances) of Spanish-English data produce translations which still allow topics to be correctly classified in 61% of input speech segments, outperforming a majority baseline. With 20 hours of parallel data, accuracy is more than 70%.

Since this is the first work in cross-lingual topic classification, there are a number of interesting avenues for future work. We used our ST model as an off-the-shelf system, and did not tune its performance for the topic prediction task. We hope future work will improve accuracy further. We used silver labels to evaluate our approach—this allowed us to compare several different settings using an objective metric. However, human annotations of topics will be the next step. We also used a pipelined approach of ST followed by classification. An alternative would be to train a topic classifier on input speech directly, but we speculate that this would require more substantial resources. Cross-lingual topic modeling may also be useful when the target language is high-resource; we learned target topics just from the 20 hours of translations, but in future work, we could use a larger text corpus in the high-resource language to learn a more general topic model covering a wider set of topics, and/or combine it with keyword lists curated for specific scenarios like disaster recovery [32].

7. ACKNOWLEDGEMENTS

This work was supported in part by a James S McDonnell Foundation Scholar Award for SG and a Google Faculty Research Award for HK. We thank Ida Szubert, Marco Damonte, and Clara Vania for helpful comments on drafts of this paper.

8. REFERENCES

[1] G. Neubig, Y. Matsubayashi, M. Hagiwara, and K. Murakami, "Safety information mining—what can NLP do in a disaster—," in *Proc. IJCNLP*, 2011.

[2] J. Quinn and P. Hidalgo-Sanchis, "Using machine learning to analyse radio content in Uganda: Opportunities for sustainable development and humanitarian action," United Nations Global Pulse Lab Kampala, Tech. Rep., 2017. [Online]. Available: http://air.ug/~jquinn/papers/UNGP_radio_analysis_report_2017.pdf

[3] R. Menon, H. Kamper, E. Van Der Westhuizen, J. Quinn, and T. R. Niesler, "Feature exploration for almost zero-resource asr-free keyword spotting using a multilingual bottleneck extractor and correspondence autoencoders," in *Proc. Interspeech*, 2019.

[4] S. Bird, L. Gawne, K. Gelbart, and I. McAlister, "Collecting bilingual audio in remote indigenous communities," in *Proc. COLING*, 2014.

[5] D. Blachon, E. Gauthier, L. Besacier, G.-N. Kouarata, M. Adda-Decker, and A. Rialland, "Parallel speech collection for under-resourced language studies using the Lig-Aikuma mobile device app," *Procedia Computer Science*, vol. 81, pp. 61–66, 2016.

[6] G. Adda, S. Stüker, M. Adda-Decker, O. Ambouroue, L. Besacier, D. Blachon, H. Bonneau-Maynard, P. Godard, F. Hamlaoui, D. Idiatov *et al.*, "Breaking the Unwritten Language Barrier: The BULB project," *Procedia Computer Science*, vol. 81, pp. 8–14, 2016.

[7] L. Besacier, B. Zhou, and Y. Gao, "Towards speech translation of non written languages," in *Proc. SLT*, 2006.

[8] A. Berard, O. Pietquin, C. Servan, and L. Besacier, "Listen and translate: A proof of concept for end-to-end speech-to-text translation," in *NIPS Workshop on end-to-end learning for speech and audio processing.*, 2016.

[9] R. J. Weiss, J. Chorowski, N. Jaitly, Y. Wu, and Z. Chen, "Sequence-to-sequence models can directly transcribe foreign speech," in *Proc. Interspeech*, 2017.

[10] S. Bansal, H. Kamper, A. Lopez, and S. Goldwater, "Towards speech-to-text translation without speech recognition," in *Proc. EACL*, 2017.

[11] S. Bansal, H. Kamper, K. Livescu, A. Lopez, and S. Goldwater, "Pre-training on high-resource speech recognition improves low-resource speech-to-text translation," in *Proc. NAACL*, 2019.

[12] K. W. Church and E. H. Hovy, "Good applications for crummy machine translation," *Machine Translation*, vol. 8, no. 4, pp. 239–258, 1993.

[13] S. Bansal, H. Kamper, K. Livescu, A. Lopez, and S. Goldwater, "Low-resource speech-to-text translation," in *Proc. Interspeech*, 2018.

[14] J. Godfrey and E. Holliman, "Switchboard-1 Release 2 (LDC97S62)," 1993.

[15] M. W. Berry, M. Browne, A. N. Langville, V. P. Pauca, and R. J. Plemmons, "Algorithms and applications for approximate nonnegative matrix factorization," *Computational statistics & data analysis*, 2007.

[16] S. Arora, R. Ge, and A. Moitra, "Learning topic models–going beyond svd," in *Proc. FOCS*, 2012.

[17] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of Machine Learning Research*, 2003.

[18] D. Graff, S. Huang, I. Cartagena, K. Walker, and C. Cieri, "Fisher Spanish Speech (LDC2010S01)," 2010.

[19] M. Post, G. Kumar, A. Lopez, D. Karakos, C. Callison-Burch, and S. Khudanpur, "Improved speech-to-text translation with the Fisher and Callhome Spanish-English speech translation corpus," in *Proc. IWSLT*, 2013.

[20] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, 2011.

[21] M. Dredze, A. Jansen, G. Coppersmith, and K. Church, "NLP on spoken documents without ASR," in *Proc. EMNLP*, 2010.

[22] J. G. Wilpon, L. R. Rabiner, C.-H. Lee, and E. R. Goldman, "Automatic recognition of keywords in unconstrained speech using hidden markov models," *IEEE Trans. Acoust. Speech Signal Process.*, vol. 38, no. 11, pp. 1870–1878, 1990.

[23] A. Garcia and H. Gish, "Keyword spotting of arbitrary words using minimal speech resources," in *Proc. ICASSP*, 2006.

[24] D. Palaz, G. Synnaeve, and R. Collobert, "Jointly learning to locate and classify words using convolutional networks," in *Proc. Interspeech*, 2016.

[25] R. Menon, H. Kamper, J. Quinn, and T. Niesler, "Fast ASR-free and almost zero-resource keyword spotting using DTW and CNNs for humanitarian monitoring," in *Proc. Interspeech*, 2018.

[26] C. Chelba, T. J. Hazen, and M. Saraclar, "Retrieval and browsing of spoken content," *IEEE Signal Proc. Mag.*, vol. 25, no. 3, 2008.

[27] Y.-C. Li, H.-y. Lee, C.-T. Chung, C.-a. Chan, and L.-s. Lee, "Towards unsupervised semantic retrieval of spoken content with query expansion based on automatically discovered acoustic patterns," in *Proc. ASRU*, 2013.

[28] L.-s. Lee, J. Glass, H.-y. Lee, and C.-a. Chan, "Spoken content retrieval—beyond cascading speech recognition with text retrieval," *IEEE Trans. Audio, Speech, Language Process.*, vol. 23, no. 9, pp. 1389–1420, 2015.

[29] P. Sheridan, M. Wechsler, and P. Schäuble, "Cross-language speech retrieval: Establishing a baseline performance," in *Proc. SIGIR*, 1997.

[30] H. Kamper and M. Roth, "Visually grounded cross-lingual keyword spotting in speech," in *Proc. SLTU*, 2018.

[31] D. Harwath, G. Chuang, and J. Glass, "Vision as an interlingua: Learning multilingual semantic embeddings of untranscribed speech," in *Proc. ICASSP*, 2018.

[32] A. Olteanu, C. Castillo, F. Diaz, and S. Vieweg, "CrisisLex: A lexicon for collecting and filtering microblogged communications in crises," in *Proc. ICWSM*, 2014.