SOME INSIGHTS FROM TRANSLATING CONVERSATIONAL TELEPHONE SPEECH

Gaurav Kumar, Matt Post, Daniel Povey, Sanjeev Khudanpur

Center for Language and Speech Processing & Human Language Technology Center of Excellence
The Johns Hopkins University, Baltimore, MD 21218, USA
{gkumar6,post,dpovey1,khudanpur}@jhu.edu

ABSTRACT

We report insights from translating Spanish conversational telephone speech into English text by cascading an automatic speech recognition (ASR) system with a statistical machine translation (SMT) system. The key new insight is that the informal register of conversational speech is a greater challenge for ASR than for SMT: the BLEU score for translating the reference transcript is 64%, but drops to 32% for translating automatic transcripts, whose word error rate (WER) is 40%. Several strategies are examined to mitigate the impact of ASR errors on the SMT output: (i) providing the ASR lattice, instead of the 1-best output, as input to the SMT system, (ii) training the SMT system on Spanish ASR output paired with English text, instead of Spanish reference transcripts, and (iii) improving the core ASR system. Each leads to consistent and complementary improvements in the SMT output. Compared to translating the 1-best output of an ASR system with 40% WER using an SMT system trained on Spanish reference transcripts, translating the output lattice of a better ASR system with 35% WER using an SMT system trained on ASR output improves BLEU from 32% to 38%.

Index Terms— Speech Recognition, Natural Language Processing, Machine Translation, Human Language Technology, Spoken Language Translation

1. INTRODUCTION

As component technologies for automatic speech recognition (ASR) and statistical machine translation (SMT) independently become more effective, capabilities such as automatic spoken language translation (SLT) are starting to be feasible. The VERBMOBIL project [1] led to some of the early work in SLT, investigating speech translation for thematically constrained tasks such as travel planning and appointment scheduling. The NESPOLE! project, inspired by its predecessor C-STAR, focused on e-commerce applications of SLT. The DARPA TRANSTAC program broadened the SLT task to a wider range of tactical human-human communication, albeit with the constraint of real-time performance on handheld computing platforms. Examples of forays into "open domain" SLT include the DARPA GALE program [2] and the European Quaero project,[1] whose numerous participants investigated translation of broadcast news and of slightly less formal speech, which was labeled *broadcast conversations*.

A characteristic of the SLT tasks tackled over the last two decades is the increasing spontaneity of the speech, and technology has evolved to cope with the concomitant increase in difficulty. Many VERBMOBIL and NESPOLE! systems used task-specific grammars for ASR and interlingua-based approaches for translation. In contrast, all the GALE and Quaero systems required large vocabulary continuous speech recognition systems with generic language models for ASR, and wide coverage SMT systems for translation.

Said differently, the speech in tasks like travel planning and e-commerce is machine-directed and limited-domain. Broadcast news is human-directed but very well enunciated. Broadcast conversations are more spontaneous, but still intended for a public audience by the speaker. Along the axis of spontaneity, the next challenge is to develop SLT systems for conversational human-human speech. This is one of the goals of the DARPA BOLT program,[2] and the focus of this paper.

We investigate the task of translating conversational telephone speech from the Spanish FISHER corpus (LDC2010S01 and LDC2010T04) into English text. We employ a cascade of an ASR and an SMT system, and quantify the relative effectiveness of the ASR and SMT components in dealing with the informal register of conversational speech. We measure the impact of ASR errors on SMT performance, and investigate several ways to mitigate the impact by tighter coupling of the ASR and SMT systems. The key findings we present are that (i) the ASR component bears the brunt of the difficulty in translating conversational speech, not the SMT component, and as a result (ii) ASR improvement and improved coupling of the ASR-SMT components can significantly improve SLT for conversational speech. The three mitigation strategies we investigate cumulatively improve the BLEU score from 31.8% to 38.0% (or reduce TER resp. from 60.4% to 53.8%).

This work was partially supported by DARPA via IBM's BOLT Contract No HR0011-12-C-0015, and by NSF via IIS Award No 0963898.

[1] http://www.quaero.org/

[2] http://www.darpa.mil/Our_Work/I2O/Programs/

2. PRIOR WORK IN SPEECH TRANSLATION

A finite state transducer (FST) based approach to SLT is proposed in [3], wherein FST-based ASR and SMT systems are composed to form a source-language speech to target-language text transducer. Given a spoken utterance in the source language, Viterbi decoding is used to find the most likely sentence in the target language. The approach is tested on a limited-domain task with a small vocabulary. A Bayes optimal tight coupling of ASR and SMT is described in [4], and assumptions (monotone alignment) and approximations (for summing over the *hidden* source-language transcript) are proposed that lead to a solution similar to [3].

Several papers note that lexical reordering is what makes full integration of ASR and SMT difficult in the FST framework. So an ASR lattice with acoustic- and language-model scores is passed in [5, 6] to an SMT system capable of lexical reordering (cf. e.g., [7]). Confusion networks are used as an alternative ASR-SMT interface by numerous researchers, including [8, 9, 10, 11]. Tighter integration of ASR and SMT is achieved in [12] by translating the n-best ASR outputs, consolidating the resulting translations, and rescoring them using a log-linear model with additional features from the ASR output that are not easy to utilize in the SMT system. The difficulty of the speech in these papers ranges from limited-domain tasks (e.g., travel reservations) to broadcast news.

Research to improve core ASR for SMT has also received considerable attention [13, 14, 15]. A comprehensive survey of speech translation research appears in the recent review article [16].

This paper contributes to this considerable body of work by investigating translation of conversational telephone speech (CTS) and studies the impact of the informal register of conversational speech on the component systems.

Some findings reported here are as expected, e.g., translating ASR lattices instead of the 1-best output usually improves the SMT output. Other findings are somewhat surprising, e.g., the imperfect ASR output differs systematically enough from the perfect transcripts that an SMT system trained on the ASR output compensates for some of the imperfections.

Perhaps the most noteworthy insight is that the informal register of CTS does not impact the SMT task as hard as it impacts the ASR task — the SMT output is of respectable quality when the reference transcripts are translated (64% BLEU), but degrades considerably (32%-38% BLEU) when ASR output is translated. This suggests that *improving core ASR quality is as critical to improving translation of CTS as tighter ASR-SMT integration, if not more critical*.

3. CORPUS AND EXPERIMENTAL SETUP

We use a Spanish-English parallel corpus [17] that we recently created by translating the reference transcripts of the Spanish FISHER corpus (LDC2010T04) into English. The 819 conversations in the 160 hour (2M word) corpus have been partitioned into a 1.8M word *training* set (ca 750 conversations), and three held-out sets of 48K-50K words (20 conversations) each, named *dev*, *dev2* and *test*. One English translation is available for each utterance in the training set, and four translation for each utterance in the held-out sets.

The primary contributions of [17] are (i) describing the process of creating the parallel text corpus and (ii) demonstrating the utility of in-domain SMT training data over out-of-domain data (e.g., newswire and parliamentary discourse). The primary contribution of this paper is quantifying the impact of the informal register of conversational speech on ASR and SMT, and its implications for ASR-SMT integration.

For the work reported here, we have used the training partition of [17] to train the ASR and SMT systems, the *dev* partition to tune ASR meta-parameters (e.g., LM order and scale) and SMT model combination parameters via MERT [18], and the *dev2* partition to evaluate ASR and SMT performance. The *test* partition is set aside for evaluating future work.

3.1. ASR System Development

We use the Kaldi speech recognition tools [19] to build our Spanish ASR systems. The speech is represented by 13-dim PLP coefficients, plus their first and second derivatives. A standard GMM-HMM system is trained with triphone acoustic models, and used to initialize the training of another triphone system with LDA+MLLT features. This is followed by speaker adapted training (SAT) with fMLLR transforms. This *SAT system* is comparable to the one used in [17], and is our baseline. The dictionary is composed of the Spanish CALLHOME lexicon (LDC96L16), augmented automatically using pronunciation rules provided with that lexicon to cover all the words in the ASR training transcripts and the most frequent words in the Spanish Gigaword corpus (LDC2011T12), to yield a dictionary of 64K words. The Spanish language model used throughout this paper is a Kneser-Ney trigram estimated from the FISHER Spanish training transcripts.

To study the impact of ASR improvements on SMT, we bootstrap a speaker adapted subspace GMM system from the SAT models, yielding the second best ASR system used in this work, the *SGMM system*. The best system, which we refer to as the *bMMI system*, results from discriminative training of the SGMM system via the boosted MMI criterion.

When processing test data, full decoding and lattice generation is performed with the SAT and SGMM systems, but the bMMI system is used only to rescore the SGMM lattices.

Finally, recall that the SMT system is trained on parallel text whose Spanish side is the same as the reference transcripts used for acoustic model training. To train an SMT system on ASR output, we carry out 10-fold jack-knifing: we divide the FISHER training set into 10 roughly equal parts, and automatically transcribe each part using a complete ASR system trained as described above on the remaining 9 parts.

The 1-best output, which thus contains realistic ASR errors, comprises the Spanish side of the new parallel text.

3.2. SMT System Development

We use the Joshua toolkit [20] to build our Spanish-English SMT system. Joshua uses hierarchical phrase-based translation models, and supports lattice input.[3] The parallel text is the Spanish transcripts in the training set of [17] paired with their English translations. The English language model is derived by interpolating two 5-gram language models, one estimated from the English side of the parallel text, and another from the transcripts of the English FISHER corpus (LDC2004T19 and LDC2005T19). We follow a standard Joshua recipe (available at `joshua-decoder.org`) for SMT system training, which entails tokenization/normalization of the text, followed by word-alignment and phrase table extraction from the parallel text, language model estimation from the English texts, and MERT tuning [18] on the *dev* set.

Now, the SMT phrase tables are conventionally extracted from Spanish reference transcripts. But we ultimately wish to use them to translate ASR output. Clearly, fully matched SMT training would require a large parallel corpus of Spanish ASR outputs paired with English translations of the Spanish speech, which may not always be available. But it is reasonable to assume having a matched "tuning" set.

Therefore, as a first mitigating step, we *tune* the SMT system on the type of data it must translate; i.e., for translating 1-best ASR output on *dev2*, we use a system whose the MERT step is performed using the Spanish 1-best ASR output on *dev*. Similarly, for translating Spanish lattices from *dev2*, we perform MERT on *dev* using Spanish lattices.

We also conduct a contrastive experiment in which the tokenization, word alignment and phrase table extraction steps in the SMT training pipeline are also performed using the Spanish 1-best outputs (instead of the reference transcripts) paired with the reference English translations. The intuition is that if the ASR output contains systematic errors relative to the reference transcripts, then the SMT system could learn to overcome them.

3.3. ASR and SMT Evaluation Metrics

Standard metrics and scoring tools are used throughout this paper. ASR output is evaluated against the reference transcripts using the NIST sclite tool, albeit without the benefit of a GLM file tailored for *dev* and *dev2*. SMT output is evaluated using BLEU-n4r4 and the NIST TER scripts, with lower-case, punctuated reference English translations.

[3] Kaldi word lattices are deterministic by design but permit epsilon arcs, and store separate acoustic- and language-model scores on each arc. Joshua requires epsilon-free lattices and treats arc weights like local probabilities. Therefore, weight-pushing and epsilon-removal is carried out on the Kaldi lattices using the Google OpenFST tools before passing them on to Joshua.

ASR System	Dataset	1-best WER	Lattice WER
SAT	*dev*	41.2%	19.2%
SAT	*dev2*	39.8%	18.6%
SGMM	*dev*	38.1%	12.8%
SGMM	*dev2*	37.0%	12.4%
bMMI	*dev*	35.9%	13.5%
bMMI	*dev2*	34.5%	12.9%

Table 1. WER improvements going from SAT to SGMM to bMMI models. This also improves BLEU and TER, as seen by comparing the corresponding rows of Tables 3, 4 and 5.

Count	Correct Word	ASR Output
54	sí	si
53	si	sí
47	mm	mhm
41	qué	que
41	y	sí
36	[noise]	[laughter]
32	las	la
32	que	qué
29	mhm	mm
29	mja	mhm

Table 2. The 10 most frequent substitution errors on *dev* suggest that an SMT system could learn to translate incorrect Spanish words to the correct English word.

4. CONVERSATIONAL SPEECH TRANSLATION

4.1. ASR System Evaluation

We begin by evaluating the performance of the Spanish ASR system on *dev* and *dev2*, as summarized in Table 1.

Note from the table that the WER on *dev2* improves from 39.8% to 34.5% as the acoustic models are improved, and one hopes that this improvement will lead to improved SMT output. This possibility is investigated in Section 4.2.3.

Note also that the most accurate path in the lattice has a "Lattice WER" that is a factor of 2-3 lower, and one hopes that the SMT system will benefit from the presence of such alternatives when translating entire lattices. This possibility is investigated in Section 4.2.1.

Finally, the ten most frequent substitution errors on the *dev* data are listed in Table 2. One may ascribe some, such as *sí* $\leftrightarrow$ *si* (which both translate to *yes*) and *qué* $\leftrightarrow$ *que* (which both translate to *what*), to inconsistent transcription conventions. But others represent systematic ASR errors. If such errors were present in the Spanish (parallel) text used for SMT training, the SMT system could potentially learn to overcome them by correctly translating the incorrect ASR output. This possibility is investigated in Section 4.2.2.

Train on	Tune on	Translate	BLEU	TER
Transcript	Transcript	Transcript	64.3%	28.5%
Transcript	Oracle	Oracle	40.8%	49.4%
Transcript	Lattice	Lattice	32.0%	59.7%
Transcript	1-best	1-best	31.8%	60.4%
1-best	1-best	1-best	33.7%	58.0%
1-best	Lattice	Lattice	34.0%	56.8%

Table 3. SMT performance on *dev2* as a function of different SMT training and tuning choices, when translating the ASR output of the SAT system with 39.8% WER.

Train on	Tune on	Translate	BLEU	TER
Transcript	Transcript	Transcript	64.3%	28.5%
Transcript	Oracle	Oracle	44.0%	46.3%
Transcript	Lattice	Lattice	34.4%	57.1%
Transcript	1-best	1-best	34.7%	57.0%
1-best	1-best	1-best	36.3%	55.5%
1-best	Lattice	Lattice	36.7%	54.4%

Table 4. SMT performance on *dev2* as a function of different SMT training and tuning choices, when translating the ASR output of the SGMM system with 37.0% WER.

Train on	Tune on	Translate	BLEU	TER
Transcript	Transcript	Transcript	64.3%	28.5%
Transcript	Oracle	Oracle	44.1%	46.2%
Transcript	Lattice	Lattice	36.6%	55.6%
Transcript	1-best	1-best	34.9%	57.3%
1-best	1-best	1-best	37.2%	54.4%
1-best	Lattice	Lattice	38.0%	53.8%

Table 5. SMT performance on *dev2* as a function of different SMT training and tuning choices, when translating the ASR output of the bMMI system with 34.5% WER.

4.2. SMT System Evaluation

We begin by translating the Spanish reference transcripts from *dev2* using an SMT system trained and tuned on Spanish reference transcripts. We then translate the ASR 1-best output from the SAT system using the same SMT system, albeit tuned on 1-best output of the SAT system on *dev*. The resulting SMT performance is reported in Table 3.

It is clear that Spanish CTS, per se, is not too difficult to translate: the BLEU score is 64.3%. However, it drops dramatically to 31.8% when translating the 1-best output of the SAT system, whose WER is 39.8% (cf. Table 1).

4.2.1. Translating ASR Lattices Instead of 1-Best Outputs

Recall from Table 1 that the most accurate path in lattices generated by the SAT system has a much lower WER (18.6%). Therefore, a natural step is to pass on the entire ASR lattice as the SMT input. The result of this exercise, also shown in Table 3, is a modest SMT improvement (32.0% BLEU).

Table 3 also shows that if an *oracle* were to select and pass on the most accurate transcript available in the lattice, then the improvement would be much more dramatic (40.8% BLEU). This suggests further research on tighter ASR-SMT integration and SMT-guided rescoring of ASR lattices.

4.2.2. SMT Training on ASR Output Instead of Reference

Inspired by the observations in Table 2, we carry out the 10-fold decoding of the speech in the training set, and use the 1-best output as the Spanish side of the parallel text for SMT training. The SMT models are tuned on *dev*, as before, and used to translate the ASR output on *dev2*. The resulting performance is shown in the lower block of Table 3.

Clearly, for both 1-best translation and lattice translation, the SMT system trained on ASR output is better: BLEU improves 31.8% to 33.7% (resp. 32.0% to 34.0%), and TER reduces from 60.4% to 58.0% (resp. 59.7% to 56.8%).

While Table 2 inspired this investigation, the ASR output differs from reference transcripts in other systematic ways, e.g., sentence segmentation and punctuation. These improvements should therefore not be attributed solely to overcoming substitution errors. Further investigations are underway.

4.2.3. Improving ASR to Improve SMT

All the results of Table 3 are based on using an ASR system with SAT acoustic models. We next study how these results change as the ASR system is incrementally improved, first by replacing the SAT models with SGMMs, and then with bMMI-trained SGMMs. Past experience with the GALE SLT tasks may lead one to suspect that small improvements in ASR of the kind shown in Table 1 will *not* lead to substantial SMT improvements. But the results of Tables 4 and 5 show a pleasantly surprising improvement on the results of Table 3.

As the WER is reduced from 39.8% to 37.0% to 34.5%, BLEU for 1-best translation increases from 33.7% to 36.3% to 37.2%, while BLEU for lattice translation goes up from 34.0% to 36.7% to 38.0%. TER reduces commensurately.

5. CONCLUDING REMARKS

The cumulative impact of the three error mitigating steps of Section 4.2 is a BLEU improvement from 31.8% to 38.0%. The largest part is from improving core ASR, the next from modified SMT training, and the rest from translating lattices.

Yet, the BLEU score for translating the oracle-best ASR output is even higher: 44.1%. This gap suggests that research is still needed to better integrate the ASR and SMT systems.

Finally, the large gap between the BLEU score for translating the oracle-best ASR output versus the reference transcript (44.1% v/s 64.3%) suggests that ASR performance is still a dominant hurdle in the path towards high quality translation of conversational telephone speech.

6. REFERENCES

[1] W. Wahlster, *Verbmobil: Foundations of speech-to-speech translation*, Springer, Sept. 2000.

[2] J. Olive, C. Christianson, and J. McCary, Eds., *Handbook of natural language processing and machine translation: DARPA Global Autonomous Language Exploitation*, Springer, Mar. 2011.

[3] E. Vidal, "Finite-state speech-to-speech translation," in *Proceedings of the IEEE International Conference in Acoustics, Speech and Signal Processing*, Apr. 1997, vol. 1, pp. 111–114.

[4] H. Ney, "Speech translation: coupling of recognition and translation," in *Proceedings of the IEEE International Conference in Acoustics, Speech and Signal Processing*, Mar. 1999, vol. 1, pp. 517–520.

[5] S. Saleem, S.-C. Jou, S. Vogel, and T. Schultz, "Using word lattice information for a tighter coupling in speech translation systems," in *Proceedings of the International Conference on Spoken Language Processing*, 2004, pp. 41–44.

[6] E. Matusov, S. Kanthak, and H. Ney, "On the integration of speech recognition and statistical machine translation.," in *Proceedings of the 9th European Conference on Speech Communication and Technology*, 2005, pp. 3177–3180.

[7] S. Bangalore and G. Riccardi, "Finite-state models for lexical reordering in spoken language translation.," in *Proceedings of the Sixth International Conference on Spoken Language Processing*, 2000, pp. 422–425.

[8] N. Bertoldi and M. Federico, "A new decoder for spoken language translation based on confusion networks," in *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding*, 2005, pp. 86–91.

[9] V. H. Quan, M. Federico, and M. Cettolo, "Integrated n-best re-ranking for spoken language translation," in *Proceedings of the 9th European Conference on Speech Communication and Technology*, 2005, pp. 3181–3184.

[10] L. Mathias and W. Byrne, "Statistical phrase-based speech translation," in *Proceedings of the IEEE International Conference in Acoustics, Speech and Signal Processing*, 2006, vol. I, pp. 561–564.

[11] N. Bertoldi, R. Zens, and M. Federico, "Speech translation by confusion network decoding," in *Proceedings of the IEEE International Conference in Acoustics, Speech and Signal Processing*, 2007, vol. IV, pp. 1297–1300.

[12] R. Zhang, G. Kikui, H. Yamamoto, T. Watanabe, F. Soong, and W. K. Lo, "A unified approach in speech-to-speech translation: integrating features of speech recognition and machine translation," in *Proceedings of the 20th international conference on Computational Linguistics*, 2004.

[13] H. Soltau, G. Saon, B. Kingsbury, J. Kuo, L. Mangu, D. Povey, and G. Zweig, "The IBM 2006 Gale arabic ASR system," in *Proceedings of the IEEE International Conference in Acoustics, Speech and Signal Processing*, 2007, vol. IV, pp. 349–352.

[14] X. Cui, L. Gu, B. Xiang, W. Zhang, and Y. Gao, "Developing high performance ASR in the IBM multilingual speech-to-speech translation system," in *Proceedings of the IEEE International Conference in Acoustics, Speech and Signal Processing*, 2008, pp. 5121–5124.

[15] L. Lamel, S. Courcinous, J. Despres, J.-L. Gauvain, Y. Josse, K. Kilgour, F. Kraft, V. B. Le, H. Ney, M. Nußbaum-Thom, I. Oparin, T. Schlippe, R. Schluter, T. Schultz, T. F. da Silva, S. Stuker, M. Sundermeyer, B. Vieru, N. T. Vu, A. Waibel, and C. Woehrling, "Speech recognition for machine translation in Quaero," in *Proceedings of the International Workshop on Spoken Language Translation*, 2011.

[16] B. Zhou, "Statistical machine translation for speech: A perspective on structures, learning, and decoding," *Proceedings of the IEEE*, vol. 101, no. 5, pp. 1180–1202, May 2013.

[17] M. Post, G. Kumar, A. Lopez, D. Karakos, C. Callison-Burch, and S. Khudanpur, "General lattice decoding for improved speech-to-text translation with the Fisher and Callhome Spanish-English speech translation corpus," *Proceedings of the International Workshop on Spoken Language Translation*, 2013.

[18] F. J. Och, "Minimum error rate training in statistical machine translation," in *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, July 2003.

[19] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The Kaldi speech recognition toolkit," in *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*, Dec. 2011.

[20] Z. Li, C. Callison-Burch, C. Dyer, J. Ganitkevitch, S. Khudanpur, L. Schwartz, W. Thornton, J. Weese, and O. Zaidan, "Joshua: An open source toolkit for parsing-based machine translation," in *Proceedings of the Fourth Workshop on Statistical Machine Translation*, Mar. 2009, pp. 135–139.