Augmenting Translation Models with Simulated Acoustic Confusions for Improved Spoken Language Translation

Yulia Tsvetkov Florian Metze Chris Dyer
Language Technologies Institute
Carnegie Mellon University
Pittsburgh, PA 15213; U.S.A.
{ytsvetko, fmetze, cdyer}@cs.cmu.edu

Abstract

We propose a novel technique for adapting text-based statistical machine translation to deal with input from automatic speech recognition in spoken language translation tasks. We simulate likely misrecognition errors using only a source language pronunciation dictionary and language model (i.e., without an acoustic model), and use these to augment the phrase table of a standard MT system. The augmented system can thus recover from recognition errors during decoding using synthesized phrases. Using the outputs of five different English ASR systems as input, we find consistent and significant improvements in translation quality. Our proposed technique can also be used in conjunction with lattices as ASR output, leading to further improvements.

1 Introduction

Spoken language translation (SLT) systems generally consist of two components: (i) an automatic speech recognition (ASR) system that transcribes source language utterances and (ii) a machine translation (MT) system that translates the transcriptions into the target language. These two components are usually developed independently and then combined and integrated (Ney, 1999; Matusov et al., 2006; Casacuberta et al., 2008; Zhou, 2013; He and Deng, 2013).

While this architecture is attractive since it relies only on components that are independently useful, such systems face several challenges. First, spoken language tends to be quite different from the highly edited parallel texts that are available to train translation systems. For example, disfluencies, such as repeated words or phrases, restarts, and revisions of content, are frequent in spontaneous speech,[1] while these are usually absent in written texts. In addition, ASR outputs typically lack explicit segmentation into sentences, as well as reliable casing and punctuation information, which are crucial for MT and other text-based language processing applications (Ostendorf et al., 2008). Second, ASR systems are imperfect and make recognition errors. Even high quality systems make recognition errors, especially in acoustically similar words with similar language model scores, for example morphological substitutions like confusing bare stem and past tense forms, and in high-frequency short words (function words) which often lack both disambiguating context and are subject to reduced pronunciations (Goldwater et al., 2010).

One would expect that training an MT system on ASR outputs (rather than the usual written-style texts) would improve matters. Unfortunately, there are few corpora of speech paired with text translations into a second language that could be used for this purpose. This has been an incentive to various MT adaptation approaches and development of speech-input MT systems. MT adaptation has been done via input text pre-processing, by transformation of spoken language (ASR output) into written language (MT input) (Peitz et al., 2012; Xu et al., 2012); via decoding ASR n-best lists (Quan et al., 2005), or confusion networks (Bertoldi et al., 2007; Casacuberta et al., 2008), or lattices (Dyer et al., 2008; Onishi et al., 2010); via additional translation features capturing acoustic information (Zhang et al., 2004); and with methods that follow a paradigm of unified decoding (Zhou et al., 2007; Zhou, 2013). In line with the previous research, we too adapt a standard MT system to a speech-input MT, but by altering the translation model itself so it is better able to

[1] Disfluencies constitute about 6% of word tokens in spontaneous speech, not including silent pauses (Tree, 1995; Kasl and Mahl, 1965)

deal with ASR output (Callison-Burch et al., 2006; Tsvetkov et al., 2013a).

We address speech translation in a resource-deficient scenario, specifically, adapting MT systems to SLT when ASR is unavailable. We augment a discriminative set that translation models rescore with **synthetic translation options**. These automatically generated translation rules (henceforth synthetic phrases) are noisy variants of observed translation rules with simulated plausible speech recognition errors (§2). To simulate ASR errors we generate acoustically and distributionally similar phrases to a source (English) phrase with a phonologically-motivated algorithm (§4). Likely phonetic substitutions are learned with an unsupervised algorithm that produces clusters of similar phones (§3). We show that MT systems augmented with synthetic phrases increase the coverage of input sequences that can be translated, and yield significant improvement in the quality of translated speech (§6).

This work makes several contributions. Primary is our framework to adapt MT to SLT by populating translation models with synthetic phrases.[2] Second, we propose a novel method to generate acoustic confusions that are likely to be encountered in ASR transcription hypotheses. Third, we devise simple and effective phone clustering algorithm. All aforementioned algorithms work in a low-resource scenario, without recourse to audio data, speech transcripts, or ASR outputs: our method to predict likely recognition errors uses phonological rather than acoustic information and does not depend on a specific ASR system. Since our source language is English, we operate on a phone level and employ a pronunciation dictionary and a language model, but the algorithm can in principle be applied without pronunciation dictionary for languages with a phonemic orthography.

2 Methodology

We adopt a standard ASR-MT cascading approach and then augment translation models with synthetic phrases. Our proposed system architecture is depicted in Figure 1.

Synthetic phrases are generated from entries in the original translation model–phrase translation

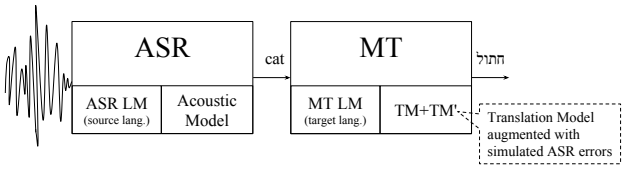

Figure 1: SLT architecture: ASR and MT are trained independently and then cascaded. We improve SLT by populating MT translation model with synthetic phrases. Each synthetic phrase is a variant of an original phrase pair with simulated ASR errors on the source side.

pairs acquired from parallel data. From a source side of an original phrase pair we generate list of its plausible misrecognition variants (pseudo-ASR outputs with recognition errors) and add them as a source side of a synthetic phrase. For k-best simulated ASR outputs we construct k synthetic phrases: a simulated ASR output in the source side is coupled with its translation–an original target phrase (identical for all k phrases). Synthetic phrases are annotated with five standard phrasal translation features (forward and reverse phrase and lexical translation probabilities and phrase penalty); these were found in the original phrase and remain unchanged. In addition, we add three new features to all phrase pairs, both synthetic and original. First, we add a boolean feature indicating the origin of a phrase: synthetic or original. Two other features correspond to an ASR language model score of the source side. One is LM score of the synthetic phrase, another is a score of a phrase from which the source side was generated. We then append synthetic phrases to a phrase table: k synthetic phrases for each original phrase pair, with eight features attached to each phrase. We show synthetic phrases example in Figure 2.

3 Acoustically confusable phones

The phonetic context of a given phone affects its acoustic realization, and a variability in a production of the same phone is possible depending on coarticulation with its neighboring phones.[3] In addition, there are phonotactic constraints that can restrict allowed sequences of phones. English has strong constraints on sequences of consonants; the sequence *[zdr]*, for example, cannot be a legal En-

[2]We augment phrase tables only with synthetic phrases that capture simulated ASR errors, the methodology that we advocate, however, is applicable to many problems in translation (Tsvetkov et al., 2013a; Ammar et al., 2013; Chahuneau et al., 2013).

[3]These are the reasons why in context-dependent acoustic modeling different HMM models are trained for different contexts.

Source phrase	Target phrase	Original phrase translation features	Synthetic indicator	Synthetic LM score	Original LM score
tells the story	*raconte l'histoire*	f_1, f_2, f_3, f_4, f_5	0	3.9×10^{-3}	3.9×10^{-3}
tell their story	raconte l'histoire	f_1, f_2, f_3, f_4, f_5	1	5.9×10^{-3}	3.9×10^{-3}
tells a story	raconte l'histoire	f_1, f_2, f_3, f_4, f_5	1	2.2×10^{-3}	3.9×10^{-3}
tell the story	raconte l'histoire	f_1, f_2, f_3, f_4, f_5	1	1.7×10^{-3}	3.9×10^{-3}
tell a story	raconte l'histoire	f_1, f_2, f_3, f_4, f_5	1	1.3×10^{-3}	3.9×10^{-3}
tell that story	raconte l'histoire	f_1, f_2, f_3, f_4, f_5	1	1.0×10^{-3}	3.9×10^{-3}
tell their stories	raconte l'histoire	f_1, f_2, f_3, f_4, f_5	1	0.9×10^{-3}	3.9×10^{-3}
tells the stories	raconte l'histoire	f_1, f_2, f_3, f_4, f_5	1	0.8×10^{-3}	3.9×10^{-3}
tells her story	raconte l'histoire	f_1, f_2, f_3, f_4, f_5	1	0.7×10^{-3}	3.9×10^{-3}
chelsea star	raconte l'histoire	f_1, f_2, f_3, f_4, f_5	1	0.5×10^{-3}	3.9×10^{-3}

Figure 2: Example of acoustically confusable synthetic phrases. Phrases were synthesized from the original phrase pair in Row 1 by generating acoustically similar phrases for the English phrase *tells the story*. All phrases have the same (target) French translation *me raconte l'histoire* and the same five basic phrase-based translation rule features. To these, three additional features are added: a synthetic phrase indicator, the source language LM score of the source phrase, and the source language LM score of a source phrase in the original phrase pair.

	T_{left}	T_{right}	TW_{left}	WIH_{right}	...
T				$P(T\|WIH)$	...
W	$P(W\|T)$				...
IH		$P(IH\|T)$	$P(IH\|TW)$		...
ER	$P(ER\|T)$				...
...	...	...	...	...	...

Figure 3: A fragment of the co-occurrence matrix for phone sequence [T W IH T ER]. Rows correspond to phones; columns correspond to left/right context phones of lengths one and two.

glish syllable onset (Jurafsky and Martin, 2000).

Motivated by the constraining effect of context on phonetic distribution, we cluster phones using a distance-based measure. To do so, we build a vector space model representation of each phone by creating a co-occurrence matrix from a corpus of phonetic forms where each row represents a phone and columns indicate the contextual phones. We take into account left/right context windows of lengths one and two. A cell $r_{p,c}$ in the vector space dictionary matrix represents phone p and context c using the empirical relative frequency $f(p \mid c)$, as estimated from a pronunciation dictionary. Figure 3 shows a fragment of the co-occurrence matrix constructed from a dictionary containing just the pronunciation of *Twitter* – [T W IH T ER].

Under this representation, the similarity of phones can be easily quantified by measuring their distance in the vector space, the cosine of the angle between them:

$$Sim(p_1, p_2) = \frac{p_1 \cdot p_2}{||p_1|| \cdot ||p_2||}$$

Armed with this similarity function, we apply the K-means algorithm[4] to partition the phones into disjoint sets.

4 Plausible misrecognition variants

For an input English sequence we generate top-k pseudo-ASR outputs, that are added as a source side of a synthetic phrase. Every ASR output that we simulate is a plausible misrecognition that has two distinguishing characteristics: it is **acoustically** and **linguistically** confusable with the input sequence. Former corresponds to phonetic similarity and latter to distributional similarity of these two phrases in corpus.

Given a reference string–a word or sequence of words $\mathbf{w}$ in the source language, we generate k-best hypotheses $\mathbf{v}$. This can be modeled as a weighted finite state transducer:

$$\{\mathbf{v}\} = G \circ D^{-1} \circ T \circ D \circ \{\mathbf{w}\} \qquad (1)$$

where

- D maps from words to pronunciations
- T is a phone confusion transducer
- D^{-1} maps from pronunciations to words
- G is an ASR language model

D maps words to their phonetic representation[5], or multiple representations for words with several

[4] Value of K=12 was determined empirically.

[5] Using the CMU pronounciation dictionary http://www.speech.cs.cmu.edu/cgi-bin/cmudict

pronunciation variants. To create a phone confusion transducer T maps source to target phone sequences by performing a number of edit operations. Allowed edits are:

- Deletion of a consonant (mapping to ϵ).
- Doubling of a vowel.
- Insertion of one or two phones in the end of a sequence from the list of possible suffixes: S (*-s*), IX NG (*-ing*), D (*-ed*).
- Substitution of a phone by an acoustically similar phone. The clusters of the similar phones are $\{$Z, S$\}$, $\{$XL, L, R$\}$, $\{$AA, AO, EY, UH$\}$, $\{$AXR, AX$\}$, $\{$XN, XM$\}$, $\{$P, B, F$\}$, $\{$DH, CH, ZH, T, SH$\}$, $\{$OY, AE$\}$, $\{$IY, AY, OW$\}$, $\{$EH, AH, IH, AW, ER, UW$\}$. The phone clustering algorithm that produced these is detailed in the previous section.

After a series of edit operations, D^{-1} transducer maps new phonetic sequences from pronunciations to n-grams of words. The k-best variants resulting from the weighted composition are the k-best plausible misrecognitions.

One important property of this method is that it maps words in decoding vocabulary (41,487 types are possible inputs to transducer D) into CMU dictionary which is substantially larger (141,304 types are possible outputs of transducer D^{-1}). This allows to generate out-of-vocabulary (OOV) words and phrases, which are not only recognition errors, but also plausible variants of different source phrases that can be translated to one target phrase, e.g., verb past tense forms or function words.

Consider a bigram *tells the* from our synthetic phrase example in Figure 2. We first obtain its phonetic representation [T EH L Z] [DH IY], and then a sequence of possible edit operations is Substitute(T, CH), Substitute(Z, S), Delete(DH) and translation of phonetic sequence [CH EH L S IY] back to words brings us to *chelsea*. See Figure 4 for visualization.

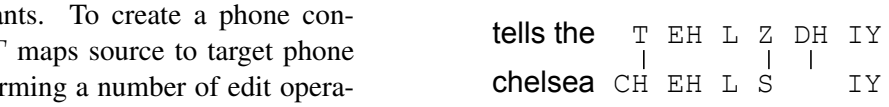

Figure 4: Pseudo-ASR output generation example for a bigram *tells the*. Phonetic edits are Substitute(T, CH), Substitute(Z, S), Delete(DH).

5 Experimental setups

To establish the effectiveness and robustness of our approach, we conducted two sets of experiments—expASR and expMultilingual—with transcribed and translated TED talks (Cettolo et al., 2012b).[6] English is the source language in all the experiments.

In expASR we used tst2011–the official test set of the SLT track of the IWSLT 2011 evaluation campaign on the English-French language pair (Federico et al., 2011).[7] This test set comprises reference transcriptions of 8 talks (approximately 1.1h of speech, segmented to 818 utterances), 1-best hypotheses from five different ASR systems, a ROVER combination of four systems (Fiscus, 1997), and three sets of lattices produced by the participants of the IWSLT 2011 ASR track.

In this set of experiments we compare baseline systems performance to a performance of systems augmented with synthetic phrases on (1) reference transcriptions, (2) 1-best hypotheses from all released ASR systems, and (3) a set of ASR lattices produced by FBK (Ruiz et al., 2011).[8] Experiments with individual systems are aimed to validate that MT augmented with synthetic phrases can better translate ASR outputs with recognition errors and sequences that were not observed in the MT training data. Consistency in performance across different ASRs is expected if our approach to generate plausible misrecognition variants is universal, rather than biased to a specific system. Comparison of 1-best system with synthetic phrases to lattice decoding setup without synthetic phrases should demonstrate whether n-best plausible misrecognition variants that we generate assemble multiple paths through a lattice.

The purpose of expMultilingual is to show that translation improvement is consistent across different target languages. This multilingual experiment is interesting because typologically different languages pose different challenges to translation (degree and locality of reordering, morphological richness, etc.). By showing that we improve results across languages (even with

[6] http://www.ted.com/

[7] http://iwslt2011.org/doku.php?id=06_evaluation#slt_track_english_to_french

[8] Pruning threshold for lattices is 0.08.

the same underlying ASR system), we show that our technique is robust to the different demands that languages place on the translation model. We could not find any publicly available multilingual data sets of the translated speech,[9] therefore we constructed a new test set.

We use our in-house speech recognizer and evaluate on locally crawled and pre-processed TED audio and text data. We build SLT systems for five target languages: French, German, Russian, Hebrew, and Hindi. Consequently, our test systems are diverse typologically and trained on corpora of different sizes. We sample a test set of seven talks, representing approximately two hours of English speech, for which we have translations to all five languages;[10] talks are listed in Table 1.

Due to segmentation differences in the released TED (text) corpora and then several automatic preprocessing stages, numbers of sentences for the same talks are not identical across languages. Therefore, we select English-French system as an oracle (this is the largest dataset), and first align it with the ASR output. Then, we filter out test sets for non-French MT systems, to retain only sentence pairs that are included in the English-French test set. Thus, our test sets for non-French MT systems are smaller, and source-side sentences in the English-French MT is a superset of source-side sentences in all five languages. Training, tuning, and test corpora sizes are listed in Table 2. Same training and development sets were used in both `expASR` and `expMultilingual` experiments.

	Training	Dev	Test
EN–FR	140,816	2,521	843
EN–DE	130,010	2,373	501
EN–RU	117,638	2,380	735
EN–HE	135,366	2,501	540
EN–HI	126,117	2,000	300

Table 2: Number of sentences in training, dev and `expMultilingual` test corpora.

5.1 ASR

In the `expMultilingual` set of experiments, we employ the JANUS Recognition Toolkit that features the IBIS single pass decoder (Soltau et al., 2001). The acoustic model is maximum likelihood system, no speaker adaptation or discriminative training applied. The acoustic model training data is 186h of Broadcast News-style data. 5-gram language model with modified Kneser-Ney smoothing is trained with the SRILM toolkit (Stolcke, 2002) on the EPPS, TED, News-Commentary, and the Gigaword corpora. The Broadcast News test set contains 4h of audio; we obtain 25.6% word error rate (WER) on this test set.

We segment the TED test audio by the timestamps of transcripts appearance on the screen. Then, we manually detect and discard noisy hypotheses around segmentation boundaries, and manually align the remaining hypotheses with the references which are the source side of the English-French MT test set. The resulting test set of 843 hypotheses, sentence aligned with transcripts, yields 30.7% WER. Higher error rates (relatively to the Broadcast News baseline) can be explained by the idiosyncratic nature of the TED genre, and the fact that our ASR system was not trained on the TED data.

For the `expASR` set of experiments the ASR outputs and lattices in standard lattice format (SLF) were produces by the participants of IWSLT 2011 evaluation campaign.

5.2 MT

We train and test MT using the TED corpora in all five languages. For French, German and Russian we use sentence-aligned training and development sets (without our test talks) released for the IWSLT 2012 evaluation campaign (Cettolo et al., 2012a); we split Hebrew and Hindi to training and development respectively.[11] We split Hebrew and Hindi to sentences with simple heuristics, and then sentence-align with the Microsoft Bilingual Sentence Aligner (Moore, 2002). Punctuation marks were removed, corpora were lowercased, and tokenized using the `cdec` scripts (Dyer et al., 2010).

In all MT experiments, both for sentence and lattice translation, we employ the Moses toolkit (Koehn et al., 2007), implementing the phrase-based statistical MT model (Koehn et al., 2003) and optimize parameters with MERT (Och, 2003). Target language 3-gram Kneser-Ney smoothed

[9] After we conducted our experiments, a new multilingual parallel corpus of translated speech was released for SLT track of IWSLT 2013 Evaluation Campaign, however, this data set does not include Russian, Hebrew and Hindi, which are a subject of this research.

[10] Since TED translation is a voluntary effort, not all talks are available in all languages.

[11] Since TED Hindi corpus is very small (only about 6K sentences) we augment it with additional parallel data (Bojar et al., 2010); however, this improved Hindi system quality only marginally, probably owing to domain mismatch.

TED id	TED talk
1	Al Gore, 15 Ways to Avert a Climate Crisis, 2006
39	Aubrey de Grey: A roadmap to end aging, 2005
142	Alan Russell: The potential of regenerative medicine, 2006
228	Alan Kay shares a powerful idea about ideas, 2007
248	Alisa Miller: The news about the news, 2008
451	Bill Gates: Mosquitos, malaria and education, 2009
535	Al Gore warns on latest climate trends, 2009

Table 1: Test set of TED talks.

language models are trained on the training part of each corpus. Results are reported using case-insensitive BLEU with a single reference and no punctuation (Papineni et al., 2002). To verify that our improvements are consistent and are not just an effect of optimizer instability (Clark et al., 2011), we train three systems for each MT setup. Statistical significance is measured with the MultEval toolkit.[12] Reported BLEU scores are averaged over three systems.

In MT adaptation experiments we augment baseline phrase tables with synthetic phrases. For each entry in the original phrase table we add (at most) five[13] best acoustic confusions, detailed in Section 4. Table 3 contains sizes of phrase tables, original and augmented with synthetic phrases.

	Original	Synthetic
EN–FR	4,118,702	24,140,004
EN–DE	2,531,556	14,807,308
EN–RU	1,835,553	10,743,818
EN–HE	2,169,397	12,692,641
EN–HI	478,281	2,674,025

Table 3: Sizes of phrase tables from the baseline systems, and phrase tables with synthetic phrases.

6 Experiments

6.1 `expASR`

We first measure the phrasal coverage of recognition errors that our technique is able to predict. We compute a number of 1- and 2-gram phrases in ASR hypotheses from the `tst2011` that are not in the references: these are ASR errors. Then, we compare their OOV rate in the English-French phrase tables, original vs. synthetic. The purpose of synthetic phrases is to capture misrecognized sequences, ergo, reduction in OOV rate of ASR errors in synthetic phrase tables corresponds to the portion of errors that our method was able to predict. Table 4 shows that the OOV rate of n-grams in phrase tables augmented with synthetic phrases drops dramatically, up to 54%. Consistent reduction of recognized errors across outputs from five different ASR systems confirms that our error-prediction approach is ASR-independent.

`tst2011`	#1-grams	#2-grams
system0	29 (50.9%)	230 (20.3%)
system1	27 (41.5%)	234 (21.3%)
system2	36 (36.0%)	230 (20.1%)
system3	34 (44.1%)	275 (20.1%)
system4	46 (52.9%)	182 (16.8%)
ROVER	30 (54.5%)	183 (18.7%)

Table 4: Phrasal coverage of recognition errors that our technique is able to predict. These are raw counts of 1-gram and 2-gram types that are OOVs in the baseline system and are recovered by our method when we augment the system with plausible misrecognitions. Percentages in parentheses show OOV rate reduction due to recovered n-grams.

Next, we explore the effect of synthetic phrases on translation performance, across different (1-best) ASR outputs. For references, ASR hypotheses, and ROVERed hypotheses we compare translations produced by MT systems trained with and without synthetic phrases. We detail our findings in Table 5.

Improvements in translation are significant for all systems with synthetic phrases. This experiment corroborates the underlying assumption that simulated ASR errors are paired with correct target phrases. Moreover, this experiment supports the claim that incorporating noisier translations in the translation model successfully adapts MT to SLT scenario and has indeed a positive effect on speech translation. Interestingly, improvement of reference translations is also observed. We speculate that this stems from better lexical selection due to a smoothing effect that our technique may

[12] https://github.com/jhclark/multeval

[13] This threshold is of course rather arbitrary. In future experiments we are planning to conduct an in-depth investigation of the threshold value, based on ASR LM score and phonetic distance from the original phrase.

	WER	BLEU Baseline	BLEU Synthetic	p
references	-	30.8	31.2	0.05
system0	22.0	24.3	25.0	<0.01
system1	23.3	23.8	24.3	<0.01
system2	21.1	23.9	24.4	0.02
system3	32.4	20.8	21.3	<0.01
system4	19.5	24.5	25.0	0.01
ROVER	17.4	25.0	25.6	0.01

Table 5: Comparison of the baseline translation systems with the systems augmented with synthetic phrases. We measure EN–FR MT performance on the `tst2011` test set: reference transcripts and ASR outputs on from five systems and their ROVER combination. Improvements in translation of all ASR outputs are statistically significant. This confirms the claim that incorporating simulated ASR errors via synthetic phrases effectively adapts MT to SLT scenario.

have.

Finally, we contrast the proposed approach of translation models adaptation to a conventional method of lattice translation. We decode FBK lattices produced for IWSLT 2011 Evaluation Campaign, and compare results to FBK 1-best translation results, which correspond to system1 in Table 5. Table 6 summarizes our main finding: 1-best system with synthetic phrases significantly outperforms lattice decoding setup with baseline translation table.[14] The additional small improvement in lattice decoding with synthetic phrases suggests that lattice decoding and phrase table adaptation are two complementary strategies and their combination is beneficial.

6.2 `expMultilingual`

In the multilingual experiment we train ten MT setups: five baseline setups and five systems with synthetic phrases, three systems per setup. For each system we compare translations of the reference transcripts and ASR hypotheses on the multilingual test set described in Section 6. We evaluate translations produced by MT systems trained with and without synthetic phrases. Table 7 summarizes experimental results, along with the test set WER for each language.

[14] Automatic evaluation results (in terms of BLEU) published during the IWSLT 2011 Evaluation Campaign (Federico et al., 2011) (p. 21) are 26.1 for FBK systems. Unsurprisingly, performance of our systems is lower, as we focus only on translation table and do not optimize factors, such as LMs and others.

	BLEU Baseline	BLEU Synthetic
FBK 1-best	23.8	24.3
FBK lattices	24.0	24.4

Table 6: Comparison of the baseline EN–FR translation systems with the systems augmented with synthetic phrases, in 1-best and lattice decoding setups. 1-best synthetic system significantly outperforms baseline lattice decoding setup. Additional improvement in lattice decoding with synthetic phrases suggests that lattice decoding and phrase table adaptation are two complementary strategies.

	WER	Baseline		Synthetic	
		Ref	ASR	Ref	ASR
EN–FR	30.7	23.3	17.8	23.9	18.1
EN–DE	33.6	14.0	11.1	14.2	11.4
EN–RU	30.7	12.3	10.7	12.2	10.6
EN–HE	29.7	9.2	7.0	9.5	7.2
EN–HI	32.1	5.5	4.5	5.6	4.8

Table 7: Comparison of the baseline translation systems with the systems augmented with synthetic phrases. We measure MT performance on the reference transcripts and ASR outputs. Consistent improvements are observed in four out of five languages.

Modest but consistent improvements are observed in four out of five setups with synthetic phrases. Only French setup yielded statistically significant improvement ($p < .01$). However, if we concatenate the outputs of all languages, the improvement in translation of references with BLEU score averaged over all systems becomes statistically significant (p = .03), improving from 16.8 for the baseline system to 17.3 for the adapted MT outputs. While more careful evaluation is required in order to estimate the effect of acoustic confusions, the accumulated result show that synthetic phrases facilitate MT adaptation to SLT across languages.

7 Analysis

We conducted careful manual analysis of actual usages of synthetic phrases in translation. The purpose of this qualitative analysis is to verify that predicted ASR errors are paired with phrases that contribute to better translation to a target language. Table 8 shows some examples. In the first sentence from the `tst2011` test set (output from system 4) the word *area* was erroneously recognized as *airy*,

English ref	so what they do is they move into an **area**
ASR output	so what they do is they move into an **airy**
Baseline MT	donc ce qu'ils font c'est qu'ils se déplacer dans un **airy**
Synthetic MT	donc ce qu'ils font c'est qu'ils se déplacer dans une **zone**
French ref	donc ce qu'ils font c'est qu'ils emménagent dans une **zone**
English ref	so **i started thinking** and listing what all it was that i thought would make a perfect biennial
ASR output	so on **i started a thinking** and listing was all it was that i thought would make a pretty by neil
Baseline MT	donc **j'ai commencé à une pensée** et listing était tout c'était que je pensais ferait un assez par neil
Synthetic MT	donc **j'ai commencé à penser** et une liste était tout c'était que je pensais ferait un assez par neil
French ref	alors **j'ai commencé à penser** et à lister tout ce qui selon moi ferait une biennale parfaite

Table 8: Examples of translations improved with synthetic phrases.

which is an OOV word for the baseline system. Our confusion generation algorithm also produced the word *airy* as a plausible misrecognition variant for the word *area* and attached it to a correct target phrase *zone*, and this synthetic phrase was selected during decoding, yielding to a correct translation for the ASR error. Second example shows a similar behavior for an indefinite article *a*. Third example is taken from the English-Russian system in the multilingual test set. *Gauge* was produced as a plausible misrecognition variant to *age*, and therefore correctly translated (albeit incorrectly inflected) as возраста(age+sg+m+acc). Synthetic phrases were also used in translations containing misrecognized function words, segmentation-related examples, and longer n-grams.

8 Related work

Predicting ASR errors to improve speech recognition quality has been explored in several previous studies. Jyothi and Fosler-Lussier (2009) develop weighted finite-state transducer framework for error prediction. They build a confusion matrix FST between phones to model acoustic errors made by the recognizer. Costs in the confusion matrix combine acoustic variations in the HMM representations of the phones (information from the acoustic model) and word-based phone confusions (information from the pronunciation model). In their follow-up work, Jyothi and Fosler-Lussier (2010) employ this error-predicting framework to train the parameters of a global linear discriminative language model that improves ASR.

Sagae et al. (2012) examined three protocols for 'hallucinating' ASR n-best lists. First approach generates confusions on the phone level, with a phone-based finite-state transducer that employs real n-best lists produced by the ASR system. Second is generating confusions at the word level with a MT-based approach. Third is a phrasal cohorts approach, in which acoustically confusable phrases are extracted from ASR n-best lists, based on *pivots*–identical left and right contexts of a phrase. All three methods were evaluated on the task of ASR improvement through decoding with discriminative language models. Discriminative language models trained on simulated n-best lists produced with phrasal cohorts method yielded the largest WER reduction on the telephone speech recognition task.

Our approach to generating plausible ASR misrecognitions is similar to previously explored FST-based methods. The fundamental difference, however, is in speech-free phonetic confusion transducer that does not employ any data extracted from acoustic models or ASR outputs. Simulated ASR errors are typically used to improve ASR applications. To the best of our knowledge no prior work has been done on integrating ASR errors directly in the translation models.

9 Conclusion

The idea behind the novel ASR error-prediction algorithm that we devise is to identify phonological neighbors with similar distributional properties, i.e. similar sounding words for which language model probabilities are insufficient for their disambiguation. These sequences have been identified as significant contributors to ASR errors (Goldwater et al., 2010). Additional and even more important factors that cause recognition errors are disfluencies in speech (Tsvetkov et al., 2013b). In the task of adapting MT to SLT these and other irregularities can effectively be incorporated in a useful general framework: *synthetic phrases* that augment phrase tables. Our experiments show that simulated acoustic confusions capture real ASR errors and that proposed framework effectively exploits them to improve translation.

Acknowledgments

We are grateful to João Miranda and Alan Black for providing us the TED audio with transcriptions, and to Zaid Sheikh for his help with ASR decoding. This work was supported in part by the U. S. Army Research Laboratory and the U. S. Army Research Office under contract/grant number W911NF-10-1-0533.

References

Waleed Ammar, Victor Chahuneau, Michael Denkowski, Greg Hanneman, Wang Ling, Austin Matthews, Kenton Murray, Nicola Segall, Yulia Tsvetkov, Alon Lavie, and Chris Dyer. 2013. The CMU machine translation systems at WMT 2013: Syntax, synthetic translation options, and pseudo-references.

Nicola Bertoldi, Richard Zens, and Marcello Federico. 2007. Speech translation by confusion network decoding. In *Proc. ICASSP*, pages 1297–1300. IEEE.

Ondrej Bojar, Pavel Stranak, and Daniel Zeman. 2010. Data issues in English-to-Hindi machine translation. In *Proceedings of LREC*.

Chris Callison-Burch, Philipp Koehn, and Miles Osborne. 2006. Improved statistical machine translation using paraphrases. In *Proceedings of HLT/NAACL*, pages 17–24. Association for Computational Linguistics.

Francisco Casacuberta, Marcello Federico, Hermann Ney, and Enrique Vidal. 2008. Recent efforts in spoken language translation. *Signal Processing Magazine, IEEE*, 25(3):80–88.

Mauro Cettolo, Marcello Federico, Luisa Bentivogli, Michael Paul, and Sebastian Stüker. 2012a. Overview of the IWSLT 2012 evaluation campaign.

Mauro Cettolo, Christian Girardi, and Marcello Federico. 2012b. WIT^3: Web inventory of transcribed and translated talks. In *Proceedings of EAMT*, pages 261–268, Trento, Italy.

Victor Chahuneau, Eva Schlinger, Noah A. Smith, and Chris Dyer. 2013. Translating into morphologically rich languages with synthetic phrases. In *Proceedings of EMNLP*.

Jonathan H. Clark, Chris Dyer, Alon Lavie, and Noah A. Smith. 2011. Better hypothesis testing for statistical machine translation: Controlling for optimizer instability. In *Proceedings of ACL*, pages 176–181. Association for Computational Linguistics.

Chris Dyer, Smaranda Muresan, and Philip Resnik. 2008. Generalizing word lattice translation. In *Proceedings of ACL-08: HLT*, pages 1012–1020. Association for Computational Linguistics.

Chris Dyer, Adam Lopez, Juri Ganitkevitch, Johnathan Weese, Ferhan Ture, Phil Blunsom, Hendra Setiawan, Vladimir Eidelman, and Philip Resnik. 2010. cdec: A decoder, alignment, and learning framework for finite-state and context-free translation models. In *Proceedings of ACL*.

Marcello Federico, Luisa Bentivogli, Michael Paul, and Sebastian Stüker. 2011. Overview of the IWSLT 2011 evaluation campaign. In *Proc. IWSLT*, pages 8–9.

Jonathan G. Fiscus. 1997. A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (ROVER). In *Proc. ASRU*, pages 347–352. IEEE.

Sharon Goldwater, Dan Jurafsky, and Christopher D Manning. 2010. Which words are hard to recognize? prosodic, lexical, and disfluency factors that increase speech recognition error rates. *Speech Communication*, 52(3):181–200.

Xiaodong He and Li Deng. 2013. Speech-centric information processing: An optimization-oriented approach. *IEEE*, 101(5):1116–1135.

Dan Jurafsky and James H Martin. 2000. *Speech & Language Processing*. Pearson Education India.

Preethi Jyothi and Eric Fosler-Lussier. 2009. A comparison of audio-free speech recognition error prediction methods. In *Proc. INTERSPEECH*, pages 1211–1214.

Preethi Jyothi and Eric Fosler-Lussier. 2010. Discriminative language modeling using simulated asr errors. In *Proc. INTERSPEECH*, pages 1049–1052.

Stanislav V Kasl and George F Mahl. 1965. The relationship of disturbances and hesitations in spontaneous speech to anxiety. In *Journal of Personality and Social Psychology*, volume 1(5), pages 425–433.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of HLT/NAACL*, pages 48–54. Association for Computational Linguistics.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of ACL*, pages 177–180. Association for Computational Linguistics.

Evgeny Matusov, Stephan Kanthak, and Hermann Ney. 2006. Integrating speech recognition and machine translation: Where do we stand? In *Proc. ICASSP*, pages V–1217–V–1220. IEEE.

Robert C. Moore. 2002. Fast and accurate sentence alignment of bilingual corpora. In *Proceedings of AMTA*, pages 135–144, London, UK. Springer-Verlag.

Hermann Ney. 1999. Speech translation: Coupling of recognition and translation. In *Proc. ICASSP*, volume 1, pages 517–520. IEEE.

Franz Josef Och. 2003. Minimum error rate training in statistical machine translation. In *Proceedings of ACL*, pages 160–167, Morristown, NJ, USA. Association for Computational Linguistics.

Takashi Onishi, Masao Utiyama, and Eiichiro Sumita. 2010. Paraphrase lattice for statistical machine translation. In *Proceedings of ACL*.

Mari Ostendorf, Benoît Favre, Ralph Grishman, Dilek Hakkani-Tur, Mary Harper, Dustin Hillard, Julia Hirschberg, Heng Ji, Jeremy G Kahn, Yang Liu, Sameer Maskey, Evgeny Matusov, Hermann Ney, Andrew Rosenberg, Elizabeth Shriberg, Wen Wang, and Chuck Wooters. 2008. Speech segmentation and spoken document processing. *Signal Processing Magazine, IEEE*, 25(3):59–69.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of ACL*, pages 311–318. Association for Computational Linguistics.

Stephan Peitz, Simon Wiesler, Markus Nußbaum-Thom, and Hermann Ney. 2012. Spoken language translation using automatically transcribed text in training. In *Proc. IWSLT*.

Vu H Quan, Marcello Federico, and Mauro Cettolo. 2005. Integrated n-best re-ranking for spoken language translation. In *Proc. INTERSPEECH*, pages 3181–3184. IEEE.

Nick Ruiz, Arianna Bisazza, Fabio Brugnara, Daniele Falavigna, Diego Giuliani, Suhel Jaber, Roberto Gretter, and Marcello Federico. 2011. FBK@ IWSLT 2011. In *Proc. IWSLT*.

Kenji Sagae, M. Lehr, E. Prud'hommeaux, P. Xu, N. Glenn, D. Karakos, S. Khudanpur, B. Roark, M. Saraçlar, I. Shafran, D. Bikel, C. Callison-Burch, Y. Cao, K. Hall, E. Hasler, P. Koehn, A. Lopez, M. Post, and D. Riley. 2012. Hallucinated n-best lists for discriminative language modeling. In *Proc. ICASSP*. IEEE.

H. Soltau, F. Metze, C. Fügen, and A. Waibel. 2001. A one-pass decoder based on polymorphic linguistic context assignment. In *Proc. ASRU*.

Andreas Stolcke. 2002. SRILM—an extensible language modeling toolkit. In *Proc. ICSLP*, pages 901–904.

Jean E Fox Tree. 1995. The effects of false starts and repetitions on the processing of subsequent words in spontaneous speech. *Journal of memory and language*, 34(6):709–738.

Yulia Tsvetkov, Chris Dyer, Lori Levin, and Archna Bhatia. 2013a. Generating English determiners in phrase-based translation with synthetic translation options. In *Proceedings of WMT*. Association for Computational Linguistics.

Yulia Tsvetkov, Zaid Sheikh, and Florian Metze. 2013b. Identification and modeling of word fragments in spontaneous speech. In *Proc. ICASSP*. IEEE.

Ping Xu, Pascale Fung, and Ricky Chan. 2012. Phrase-level transduction model with reordering for spoken to written language transformation. In *Proc. ICASSP*, pages 4965–4968. IEEE.

Ruiqiang Zhang, Genichiro Kikui, Hirofumi Yamamoto, Taro Watanabe, Frank Soong, and Wai Kit Lo. 2004. A unified approach in speech-to-speech translation: integrating features of speech recognition and machine translation. In *Proceedings of COLING*, page 1168. Association for Computational Linguistics.

Bowen Zhou, Laurent Besacier, and Yuqing Gao. 2007. On efficient coupling of ASR and SMT for speech translation. In *Proc. ICASSP*, volume 4, pages IV–101. IEEE.

Bowen Zhou. 2013. Statistical machine translation for speech: A perspective on structures, learning, and decoding. *IEEE*, 101(5):1180–1202.