Submitted to ICASSP*, 2006.*

STATISTICAL PHRASE-BASED SPEECH TRANSLATION

Lambert Mathias [1] *and William Byrne* [1,2]

Center for Language and Speech Processing, The Johns Hopkins University,
3400 N. Charles Street, Baltimore, MD 21218, U.S.A. [1]
Department of Engineering, Cambridge University
Trumpington Street, Cambridge, CB2 1PZ, U.K. [2]
lambert@jhu.edu , wjb31@cam.ac.uk

ABSTRACT

A generative statistical model of speech-to-text translation is developed as an extension of existing models of phrase-based text translation. Speech is translated by mapping ASR word lattices to lattices of phrase sequences which are then translated using operations developed for text translation. Performance is reported on Chinese to English translation of Mandarin Broadcast News.

1. INTRODUCTION

Statistical speech translation systems vary in the degree to which the Statistical Machine Translation (SMT) system and the automatic speech recognition (ASR) component are integrated within the overall translation process. In the 'pipeline' approach to speech translation, the transcription produced by ASR is translated as if it were any fluent, written sentence in the foreign language. This is a reasonable first approach to speech translation, and if ASR systems performed flawlessly, it would be perfectly adequate. However ASR systems are imperfect, and in imperfect statistical information processing systems it is generally desirable that initial processing procedures should pass on as much information as possible for use by subsequent stages.

With this motivation, speech translation architectures have been developed within which the ASR and SMT systems are *tightly coupled* (e.g. [1, 2, 3]). The objective is to allow the SMT system to search among many likely ASR hypotheses and hopefully produce a better translation than if it had been restricted to the single, best ASR hypothesis. In practice, the close coupling of ASR and MT can be realized by translating ASR N-Best lists [4, 5] or word lattices [6, 7]. N-Best translation is straightforward: a text-based SMT system can be used without modification to translate each entry, and the resulting translations can be sorted by some combination of ASR and SMT scores.

Although it is a complicated modeling and implementation problem, lattice-based translation offers potential advantages over translation of N-Best lists. Lattices provide larger search spaces, as well as detailed, sub-sentential information, such as word-level acoustic and language model scores, that can be passed directly to the SMT system. However it is not trivial to obtain gains in lattice-based translation relative to simply translating the ASR transcription. Initial attempts at incorporating word lattice information in translation did not yield consistent improvements in translation performance [6]. However approaches were subsequently developed by which lattices and confusion networks can be translated with improvements in translation quality [7, 8].

Motivated by this prior work, we present a novel approach to statistical phrase-based speech translation. This approach is based on a generative, source-channel model of translation, similar in spirit to the modeling approaches that underly HMM-based ASR systems - in fact, our model of speech-to-text translation contains the acoustic models of a large vocabulary ASR system as one of its components. We develop this model of speech-to-text translation as a direct extension of the phrase-based models used in our text translation systems, and we will show how lattice-based speech-to-text translation can be carried out easily, using the existing text-based translation systems essentially without modification.

We begin with a review of the underlying phrase-based translation model, and then extend it to speech translation by incorporating the acoustic models from a target language ASR system.

2. PHRASE-BASED GENERATIVE MODELS OF SPEECH TRANSLATION

The Translation Template Model (TTM) [9, 10] is a generative model of translation that consists of a series of transformative operations specified by conditional probability distributions. A (simplified) description of the generative process has the following steps.

Step 1 The source language sentence $s_1, \ldots, s_I$ is generated by the *Source Language Model*, $P(s_1^I)$.

Step 2 The source language sentence is segmented into a series of source language phrases, u_1^K. There are many possible sequences of phrases that can be derived from a single sentence, as defined by the *Source Phrase Segmentation* distribution, $P(u_1^K, K|s_1^I)$.

Step 3 The sequences of source language phrases are translated into target language phrase sequences, x_1^K. The target language phrases are generated in source language phrase order, and each was generated by a single source phrase. Since source language phrases can generate multiple target phrases, the generation of target phrase sequences is specified by the *Phrase Translation* distribution, $P(x_1^K|u_1^K)$.

Step 4 New target language phrases are allowed to insert themselves into the target language sequences which are then (optionally) reordered; the tendency towards phrase insertion is controlled by a single parameter, the *Phrase Exclusion Probability*. This generates modified target language

Supported by NSF ITR IIS Award No. 0122466.

phrase sequences, v_1^R, under the *Phrase Movement and Insertion* distribution, $P(v_1^R|x_1^K, u_1^K)$.

Step 5 The target language phrase sequences are transformed to target language word sequences, $t_1, \ldots, t_J$, under the *Target Phrase Segmentation* distribution, $P(t_1^J|v_1^R)$. In practice, this is a degenerate transformation which maps every target phrase sequence to its unique word sequence.

Taken together, these distributions form a joint probability distribution over the source and target language sentences, and over the possible intermediate source and target phrase sequences. Moreover, the component distributions are formulated so that each can be implemented as a Weighted Finite State Machine (WFSM) [11, 12]. The component distributions form $P(t_1^J, v_1^R, x_1^K, u_1^K, s_1^I)$ as

$$\underset{\Omega}{P(t_1^J|v_1^R)}\ \underset{\Phi}{P(v_1^R|x_1^K, u_1^K)}\ \underset{R}{P(x_1^K|u_1^K)}\ \underset{W}{P(u_1^K|s_1^I)}\ \underset{G}{P(s_1^I)}$$

where the symbol beneath each distribution denotes its FSM.

To translate a given target language sentence t_1^J into the source language, we construct an acceptor T for the target sentence. In theory, we could then create a lattice of translations via the following sequence of FSM compositions

$$\mathcal{T} = G \circ W \circ R \circ \Phi \circ \Omega \circ T$$

and extract the translation $\widehat{s_1^I}$ as the path in the translation lattice $\mathcal{T}$ with the least cost (negative log likelihood), to approximate $\widehat{s_1^I} = \text{argmax}_{s_1^I}\ P(t_1^J|s_1^I)\, P(s_1^I)$ as

$$\widehat{s_1^I} = \underset{s_1^I}{\text{argmax}}\{\max_{v_1^R, x_1^K, u_1^K, K} P(t_1^J, v_1^R, x_1^K, u_1^K, s_1^I)\}\ .$$

In practice, we perform translation in distinct steps. We first generate the *target phrase lattice*, Q, which is a WFSM acceptor for all phrase sequences in the target sentence. Q is found by composition $\Omega \circ T$ followed by projection onto the input side of the resulting transducer. We next list all the unique target phrases in Q ; these are the phrases for which source language translations are needed, and candidate source language phrases are extracted for them from bilingual training text [13, 14]. This collection of source and target translation pairs is the *phrase pair inventory*.

At this point, we have the statistics to construct a compact Phrase Translation transducer R for the sentence to be translated, as well as the Source Phrase Segmentation transducer W and the reordering and insertion transducer, Φ [9]. The translation lattice is then generated as

$$\mathcal{T} = \underbrace{G}_{\substack{Source\\Language\\Model}} \circ \underbrace{W \circ R \circ \Phi}_{\substack{Source\,Word\,to\\Target\,Phrase\\Translation}} \circ \underbrace{Q}_{\substack{Target\\Phrase\\Acceptor}}$$

The point to stress here is that translation is actually carried out through a series of FSM compositions acting on a *phrase lattice*, i.e. an acceptor of target language phrases. In text translation, this accepts all the phrase sequences that can be derived from the single sentence to be translated. To translate speech, we can simply use an acceptor for all the target language phrase sequences in an ASR word lattice. We now extend the model formulation to support this.

2.1. Speech Translation from ASR Phrase Lattices

We assume we have an ASR system with target language acoustic models $P(A|t_1^J)$ and a target language model. To describe how source language text might generate a a target language utterance A, we define $P(A, t_1^J, v_1^R, x_1^K, u_1^K, s_1^I)$ as

$$\underset{\mathcal{L}}{P(A|t_1^J)}\ \underset{\Omega}{P(t_1^J|v_1^R)}\ \underset{\Phi}{P(v_1^R|x_1^K, u_1^K)}\ \underset{R}{P(x_1^K|u_1^K)}\ \underset{W}{P(u_1^K|s_1^I)}\ \underset{G}{P(s_1^I)}$$

where $\mathcal{L}$ is an weighted acceptor containing the word sequences and acoustic scores from an lattice generated by the ASR system over the utterance A. In translation from speech, the ideal translation $\widehat{s_1^I} = \text{argmax}_{s_1^I}\ P(A|s_1^I)P(s_1^I)$ is approximated as

$$\widehat{s_1^I} = \underset{s_1^I}{\text{argmax}}\{\max_{t_1^J, v_1^R, x_1^K, u_1^K, K} P(A, t_1^J, v_1^R, x_1^K, u_1^K, s_1^I)\}\ .$$

As an aside, this differs from translation of the ASR transcript, which would be $\widehat{t_1^J} = \text{argmax}_{t_1^J} P(A, t_1^J)$

$$\widehat{s_1^I} = \underset{s_1^I}{\text{argmax}}\{\max_{v_1^R, x_1^K, u_1^K, K} P(\widehat{t_1^J}, v_1^R, x_1^K, u_1^K, s_1^I)\}\ .$$

We briefly digress to contrast our model to the previously mentioned lattice-based speech translation approaches [7, 8]. While different, those are based on joint generation and translation, i.e. a parameterized distribution $P_\lambda(t_1^J, s_1^I|A)$ describes the simultaneous generation of a source sentence and its translation. This differs from generative approaches, which rely on distinct parameterized distributions, e.g. $P_{\lambda_1}(A|t_1^J)P_{\lambda_2}(t_1^J|s_1^I)P_{\lambda_3}(s_1^I)$. The two approaches have their advantages and disadvantages, but it is worth noting that they arise from fundamentally different formulations, and involve quite different estimation and decoding procedures.

Resuming the discussion of implementing the speech translation process via WFSMs, we could replace the (unweighted) acceptor T constructed for a single target sentence to be translated by the (weighted) acceptor for the word strings in the ASR lattice

$$\mathcal{T} = G \circ W \circ R \circ \Phi \circ \Omega \circ \mathcal{L}\ .$$

But what is done follows the text translation approach: the Target Phrase Segmentation transducer is applied to the word lattice acceptor, as $\Omega \circ \mathcal{L}$, to generate a lattice of phrases, Q. The translation lattice is then found as $\mathcal{T} = G \circ W \circ R \circ \Phi \circ Q$, and the translation $\widehat{s_1^I}$ is found as the minimum cost path through $\mathcal{T}$.

There are of course modeling and implementation issues that arise in translating ASR lattices relative to translating individual text strings. For example, it is considerably easier to enumerate all the phrases in a sentence than in a word lattice. This and other issues are non-trivial modeling problems, and our current approaches to them are discussed next. However, despite these modeling challenges, we emphasize that this approach to speech translation neatly avoids the difficult problem of developing statistical translation systems that can process ASR word lattices. That problem is replaced instead by a modeling problem, namely how to extract phrase sequences from word lattices.

2.2. Transforming ASR Word Lattices into Phrase Lattices

We describe our initial approach to transforming ASR word lattices into phrase lattices suitable for translation by the TTM. Formally, we would like to extract the target phrase sequences under

the posterior distribution provided by the ASR system:

$$Q = P(v_1^R|A) = \frac{\sum_{t_1^J} P(v_1^R|t_1^J)\, P(A|t_1^J)\, P(t_1^J)}{P(A)}$$

based on the acoustic scores $P(A|t_1^J)$ and the target language model scores $P(t_1^J)$. The latter does not appear in the formulation of overall translation; however we include it simply because it improves translation performance, and note that proper inclusion of the target language model will require extension of the TTM itself.

In addition to words, ASR lattices can contain silence markers, fillers, and sentence breaks. Since these do not occur within sentences in our bilingual text collections, it is difficult to extract phrases that cover them. We map these symbols to NULL. Consequently, some of the phrases extracted will span what the ASR system hypothesized as sentence breaks. This is less than ideal, and we note this as an opportunity to incorporate metadata extraction techniques to guide phrase extraction through improved detection of phrase and sentence boundaries [15].

After this initial processing, the list of all phrases is extracted from the word lattices, as the list of all phrases is extracted from the target sentence in text translation. To extract the phrases from the ASR word lattice, we use the GRM Library *grmcount* tool which counts subsequences in a WFSM [16].

3. SPEECH TO TEXT TRANSLATION PERFORMANCE

We investigate the performance of our systems on the TC-STAR Chinese to English (C-E) Broadcast News translation task [1].

3.1. Mandarin Broadcast News Translation Task Description

The speech-to-text translation corpus is based on six Mandarin news broadcasts which were manually segmented and transcribed into Chinese sentences for use as reference transcriptions for ASR system evaluation. Two English translations of each Chinese sentence transcription were commissioned for translation system scoring. Three documents form the Development Set, and the other three the Evaluation Set, as specified by the 2005 TC-STAR evaluation; they contain 525 and 494 sentences, respectively. The overall statistics are given in Table 1.

	C	E-1	E-2
Dev	3,156 / 12,648	3,232 / 12,865	3,107 / 12,177
Eval	2,993 / 13,023	2,809 / 13,199	2,771 / 13,101

Table 1. Dev and Eval Set Vocabularies (types/tokens).

In addition to the six audio documents, their Chinese text transcriptions, and their corresponding English translations, we also have Mandarin ASR lattices in HTK format [17]. These were generated by the LIMSI Mandarin Broadcast News System [18] incorporating cross-word triphone acoustic models and a 4-gram language model.

The LIMSI Mandarin Broadcast News ASR system was applied to the complete audio document. The system was allowed generate its own acoustic segmentation independently of the manual acoustic segmentation performed during the initial transcription. Corresponding to this automatic segmentation, there are 231 ASR lattices for the Dev set and 181 ASR lattices for the Eval set.

[1] http://www.tc-star.org

Since the audio segmentation was performed automatically by the ASR system, the number of lattices is not the same as the number of manually segmented sentences in the reference transcriptions. Prior to scoring, the ASR hypotheses and the reference transcriptions are each concatenated in temporal order to form a single, long document. The ASR Character Error Rate (CER) over the Dev set was found to be 8.7%.

3.2. Mandarin Broadcast News Phrase-Based SMT System

Translation experiments are based on the TTM phrase-based SMT system [9, 10], and the experiments reported here are performed on the basic system submitted by JHU/CU to the 2005 TC-STAR and NIST Chinese-English MT evaluations.

The underlying training bitext consists of C-E parallel corpora provided by LDC (http://www.ldc.org), mainly consisting of FBIS, Xinhua, Hong Kong News, Sinorama news sources, the Chinese Treebank, and the Hong Kong Hansards and UN proceedings; the bitext contains 175M Chinese words and 200M English words. The Chinese text was word segmented using the LDC segmenter followed by rule-based number grouping. The English text was processed using a slightly modified version of the tokenizer distributed in the NIST MT-eval toolkit [19].

The documents were aligned at the sentence and sub-sentence level [20] to produce 7M bilingual sentence or subsentence chunk pairs. The chunk-aligned bitext was then aligned at the word level under the Word-to-Phrase alignment model [14]. Phrase-pairs were extracted by commonly used heuristics [13]; phrase pairs were extracted only as needed to cover the Chinese phrases to be translated. This process was complicated by inconsistent Chinese tokenization and word segmentation schemes between the ASR system and the SMT bitext; this is discussed in the next section.

The English language model training data consists of 380M words of text from the LDC English Gigaword (AFP and Xinhua), the English side of FBIS, and the online archives of People's Daily. On the C-E task we estimated an interpolated 3-gram target LM with uniform weights over the three LM English sources. For LM training, the corpus was lower-cased and punctuation removed.

Prior to translation, the Chinese ASR lattices were converted into weighted finite state acceptors in AT&T FSM format [11, 12]; time information was removed, and the lattices were reduced in size by applying ϵ-removal, determinization, and minimization [11, 12]. Lattices were in joint-likelihood form with acoustic and language model scores were combined using a Word Insertion Penalty and a Grammar Scale Factor optimized for ASR Word Error Rate by LIMSI. The ASR word lattices are pruned as necessary, and after composition with the target phrase segmentation transducer, phrases are extracted up to 5 target word in length. Parameters were optimized over the dev set.

Translation performance was measured under the BLEU metric [21] with respect to the two sets of English transcriptions. Casing was preserved in the reference translation, and the SMT output was re-cased using the SRILM *disambig* tool [22] with a modified Kneser-Ney 3-gram LM trained over the English LM text.

Baseline translation performance is reported by applying various translation system configurations to the Chinese reference transcriptions. In these baseline experiments, BLEU scores are reported at both the sentence level (sBLEU) and document level (dBLEU). However, since there is no easily found correspondence between the ASR acoustic segmentation and the manual segmentation from which the sentence level translations are derived, only

	Mandarin Source	DEV	EVAL
Monotone Phrase Order	Ref. Transcription	12.8 / 16.1	14.1 / 18.8
	ASR 1-Best	14.8	13.6
	ASR lattice	15.0	13.8
MJ-1 VT Phrase Reordering	Ref. Transcription	12.9 / 16.1	14.1 / 19.3
	ASR 1-Best	15. 0	13.8
	ASR lattice	15.1	14.0

Table 2. Mandarin Broadcast News Translation Performance.

dBLEU scores are reported for the speech translation experiments.

Translation performance is reported in Table 2 over the manual reference transcriptions (sBLEU/dBLEU scores are provided), the ASR 1-Best hypotheses, and the ASR lattices. Two configurations of the TTM are investigated. In the first, target phrases appear in monotone phrase order, i.e. Chinese phrases appear in English phrase order. In the second, MJ-1 VT phrase reordering allows target phrases swap places with their immediate neighbors as determined by reordering probabilities estimated over bitext (with a backoff swap probability of 0.02) [10]. Although improvements are not large, we find improvements in all scenarios by translating ASR lattices instead of ASR 1-Best hypothesis.

In comparing the speech and text translation systems, we note that from the Dev Set reference transcriptions we extract 44,744 Chinese phrases; 11,617 of these appear in the bitext, accompanied by 59,589 English phrases (after pruning). By comparison, we extract 58,395 Chinese phrases from the ASR lattices - 1.3 times as many phrases as appear in the reference transcriptions. However, we are able to find only 12,983 of these Chinese phrases in the bitext, and these are accompanied by 60,574 English translations. In summary, we find that in translating the lattice we have increased the number of Chinese phrases and their English alternatives only slightly. There are two factors at work. The first is that our phrase extraction procedure was developed for phrases extracted from text; different modeling procedure will be needed to translate phrases hypothesized by ASR systems, which tend to be disfluent and are relatively unlikely to appear in training bitext text. The second, dominant, problem is the mismatch in tokenization and word segmentation between the ASR system and the Chinese side of the bitext. We anticipate performance improvements when we integrate ASR and SMT systems constructed with consistent text formatting.

4. CONCLUSION

We have presented a modeling framework for statistical speech-to-text translation as an extension of the phrase-based TTM text translation model. This formulation leads to a tight coupling of the ASR and SMT subsystems, both as statistical models and as implemented by the WFSM phrase-based translation system. We have identified and described weaknesses in this initial formulation and its implementation, and we intend to improve upon these in subsequent work. Mandarin-to-English Broadcast News translation experiments demonstrate that the approach is feasible, and we anticipate further improvements in translation performance by integrated development of the component ASR and SMT systems.

5. REFERENCES

[1] E. Vidal, "Finite-state speech-to-speech translation," in *Proc. ICASSP*, 1997.

[2] H. Ney, "Speech translation: Coupling of recognition and translation," in *Proc. ICASSP*, 1999.

[3] Casacuberta et al., "Architectures for speech-to-speech translation using finite-state models," in *Proc. Workshop on Speech-to-Speech Translation*, 2002.

[4] R. Zhang et al, "A unified approach in speech-to-speech translation: integrating features of speech recognition and machine translation," in *Proc. COLING*, 2004.

[5] V.H. Quan et al., "Integrated N-best re-ranking for spoken language translation," in *In EuroSpeech*, 2005.

[6] S.Saleem, S. C. Jou, S. Vogel, and T. Schultz, "Using word lattice information for a tighter coupling in speech translation systems," in *Proc. ICSLP*, 2004.

[7] E. Matusov, S.Kanthak, and H. Ney, "On the integration of speech recognition and statistical machine translation," in *Proc. InterSpeech*, 2005.

[8] N. Bertoldi and M. Federico, "A new decoder for spoken language translation based on confusion networks," in *IEEE ASRU Workshop*, 2005.

[9] S. Kumar, Y. Deng, and W. Byrne, "A weighted finite state transducer translation template model for statistical machine translation," *J. Natural Language Engineering*, vol. 11, no. 3, 2005.

[10] S. Kumar and W. Byrne, "Local phrase reordering models for statistical machine translation," in *Proc. of HLT-EMNLP*, 2005.

[11] M. Mohri, F. Pereira, and M. Riley, "Weighted automata in text and speech processing," in *European Conference on Artificial Intelligence*, 1996.

[12] M. Mohri, F. Pereira, and M. Riley, *ATT General-purpose finite-state machine software tools*, 2001, http://www.research.att.com/sw/tools/fsm/.

[13] F. Och, *Statistical Machine Translation: From Single Word Models to Alignment Templates*, Ph.D. thesis, RWTH Aachen, Germany, 2002.

[14] Y. Deng and W. Byrne, "HMM word and phrase alignment for statistical machine translation," in *Proc. of HLT-EMNLP*, 2005.

[15] Y. Liu *et al.*, "Structural metadata research in the EARS program," in *Proc. ICASSP*, 2005.

[16] C. Allauzen, M. Mohri, and B. Roark, "Generalized algorithms for constructing statistical language models," in *41st Meeting of the ACL*, July 2003.

[17] S. Young *et al.*, *The HTK Book, Version 3.1*, Dec. 2001.

[18] L. Chen, L. Lamel, and J.-L. Gauvain, "Transcribing Mandarin Broadcast News," in *IEEE ASRU Workshop*, 2003.

[19] NIST, *The NIST Machine Translation Evaluations*, 2004, http://www.nist.gov/speech/tests/mt/.

[20] Y. Deng, S. Kumar, and W. Byrne, "Bitext chunk alignment for statistical machine translation," *J. Natural Language Engineering*, Submitted.

[21] K. Papineni, S. Roukos, T. Ward, and W. Zhu, "BLEU: a method for automatic evaluation of machine translation," Tech. Rep. RC22176(W0109-022), IBM Research, 2001.

[22] A. Stolcke, "SRILM – an extensible language modeling toolkit," in *Proc. ICSLP*, 2002.