An Attentional Model for Speech Translation Without Transcription

Long Duong,[12] **Antonios Anastasopoulos,**[3] **David Chiang,**[3] **Steven Bird**[14] and **Trevor Cohn**[1]
[1]Department of Computing and Information Systems, University of Melbourne
[2]National ICT Australia, Victoria Research Laboratory
[3]Department of Computer Science and Engineering, University of Notre Dame
[4]International Computer Science Institute, University of California Berkeley

Abstract

For many low-resource languages, spoken language resources are more likely to be annotated with translations than transcriptions. This bilingual speech data can be used for word-spotting, spoken document retrieval, and even for documentation of endangered languages. We experiment with the neural, attentional model applied to this data. On phone-to-word alignment and translation reranking tasks, we achieve large improvements relative to several baselines. On the more challenging speech-to-word alignment task, our model nearly matches GIZA++'s performance on gold transcriptions, but without recourse to transcriptions or to a lexicon.

1 Introduction

For many low-resource languages, spoken language resources are more likely to come with translations than with transcriptions. Most of the world's languages are not written, so there is no orthography for transcription. Phonetic transcription is possible but too costly to produce at scale. Even when a minority language has an official orthography, people are often only literate in the language of formal education, such as the national language. Nevertheless, it is relatively easy to provide written or spoken *translations* for audio sources. Subtitled or dubbed movies are a widespread example.

One application of models of bilingual speech data is documentation of endangered languages. Since most speakers are bilingual in a higher-resource language, they can listen to a source language recording sentence by sentence and provide a spoken translation (Bird, 2010; Bird et al., 2014). By aligning this data at the word level, we hope to automatically identify regions of data where further evidence is needed, leading to a substantial, interpretable record of the language that can be studied even if the language falls out of use (Abney and Bird, 2010; Bird and Chiang, 2012).

We experiment with extensions of the neural, attentional model of Bahdanau et al. (2015), working at the phone level or directly on the speech signal. We assume that the target language is a high-resource language such as English that can be automatically transcribed; therefore, in our experiments, the target side is text rather than the output of an automatic speech recognition (ASR) system.

In the first set of experiments, as a stepping stone to direct modeling of speech, we represent the source as a sequence of phones. For phone-to-word alignment, we obtain improvements of 9–24% absolute F1 over several baselines (Och and Ney, 2000; Neubig et al., 2011; Stahlberg et al., 2012). For phone-to-word translation, we use our model to rerank n-best lists from Moses (Koehn et al., 2007) and observe improvements in BLEU of 0.9–1.7.

In the second set of experiments, we operate directly on the speech signal, represented as a sequence of Perceptual Linear Prediction (PLP) vectors (Hermansky, 1990). Without using transcriptions or a lexicon, the model is able to align the source-language speech to its English translations nearly as well as GIZA++ using gold transcriptions.

Our main contributions are: (i) proposing a new task, alignment of speech with text translations, including a dataset extending the Spanish

Proceedings of NAACL-HLT 2016, pages 949–959,
San Diego, California, June 12-17, 2016.

Fisher and CALLHOME datasets; (ii) extending the neural, attentional model to outperform existing models at both alignment and translation reranking when working on source-language phones; and (iii) demonstrating the feasibility of alignment directly on source-language speech.

2 Background

To our knowledge, there has been relatively little research on models that operate directly on parallel speech. Typically, speech is transcribed into a word sequence or lattice using ASR, or at least a phone sequence or lattice using a phone recognizer. This normally requires manually transcribed data and a pronunciation lexicon, which can be costly to create. Recent work has introduced models that do not require pronunciation lexicons, but train only on speech with text transcriptions (Lee et al., 2013; Maas et al., 2015; Graves et al., 2006). Here, we bypass phonetic transcriptions completely, and rely only on translations.

Such data can be found, for example, in subtitled or dubbed movies. Some specific examples of corpora of parallel speech are the European Parliament Plenary Sessions Corpus (Van den Heuvel et al., 2006), which includes parliamentary speeches in the 21 official EU languages, as well as their interpretation into all the other languages; and the TED Talks Corpus (Cettolo et al., 2012), which provides speech in one language (usually English) together with translations into other languages.

As mentioned in the introduction, a stepping-stone to model parallel speech is to assume a recognizer that can produce a phonetic transcription of the source language, then to model the transformation from transcription to translation. We compare against three previous models that can operate on sequences of phones. The first is simply to run GIZA++ (IBM Model 4) on a phonetic transcription (without word boundaries) of the source side. Stahlberg et al. (2012) present a modification of IBM Model 3, named Model 3P, designed specifically for phone-to-word alignment. Finally, pialign (Neubig et al., 2011), an unsupervised model for joint phrase alignment and extraction, has been shown to work well at the character level (Neubig et al., 2012) and extends naturally to work on phones.

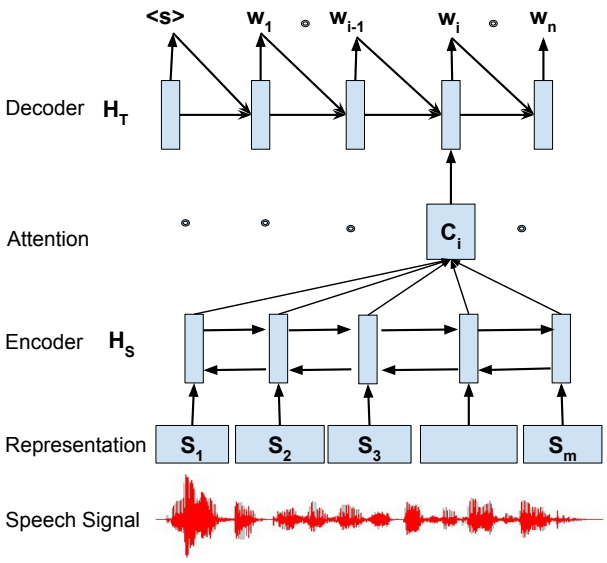

Figure 1: The attentional model as applied to our tasks. We consider two types of input: discrete phone input, or continuous audio, represented as PLP vectors at 10ms intervals

3 Model

We base our approach on the attentional translation model of Cohn et al. (2016), an extension of Bahdanau et al. (2015) which incorporates more fine grained components of the attention mechanism to mimic the structural biases in standard word based translation models. The attentional model encodes a source as a sequence of vectors, then decodes it to generate the output. At each step, it "attends" to different parts of the encoded sequence. This model has been used for translation, image caption generation, and speech recognition (Luong et al., 2015; Xu et al., 2015; Chorowski et al., 2014; Chorowski et al., 2015). Here, we briefly describe the basic attentional model, following Bahdanau et al. (2015), review the extensions for encoding structural biases (Cohn et al., 2016), and then present our novel means for adapting the approach handle parallel speech.

3.1 Base attentional model

The model is shown in Figure 1. The speech signal is represented as a sequence of vectors $S_1, S_2, \ldots, S_m$. For the first set of experiments, each S_i is a 128-dimensional vector-space embedding of a phone. For the second set of experiments, each S_i is the

39-dimensional PLP vector of a single frame of the speech signal. Our model has two main parts: an encoder and a decoder. For the encoder, we used a bidirectional recurrent neural network (RNN) with Long Short-Term Memory (LSTM) units (Hochreiter and Schmidhuber, 1997); we also tried Gated Recurrent Units (Pezeshki, 2015), with similar results. The source speech signal is encoded as sequence of vectors $H_S = (H_S^1, H_S^2, \ldots, H_S^m)$ where each vector H_S^j $(1 \leq j \leq m)$ is the concatenation of the hidden states of the forward and backward LSTMs at time j.

The attention mechanism is added to the model through an alignment matrix $\alpha \in \mathbb{R}^{n \times m}$, where n is the number of target words. We add <s> and </s> to mark the start and end of the target sentence. The row $\alpha_i \in \mathbb{R}^m$ shows where the model should attend to when generating target word w_i. Note that $\sum_{j=1}^m \alpha_{ij} = 1$. The "glimpse" vector c_i of the source when generating w_i is $c_i = \sum_j \alpha_{ij} H_S^j$.

The decoder is another RNN with LSTM units. At each time step, the decoder LSTM receives c_i in addition to the previously-output word. Thus, the hidden state[1] at time i of the decoder is defined as $H_T^i = \text{LSTM}(H_T^{i-1}, c_i, w_{i-1})$, which is used to predict word w_i:

$$p(w_i \mid w_1 \cdots w_{i-1}, H_S) = \text{softmax}(g(H_T^i)), \quad (1)$$

where g is an affine transformation. We use 128 dimensions for the hidden states and memory cells in both the source and target LSTMs.

We train this model using stochastic gradient descent (SGD) on the negative log-likelihood for 100 epochs. The gradients are rescaled if their L2 norm is greater than 5. We tried Adagrad (Duchi et al., 2011), AdaDelta (Zeiler, 2012), and SGD with momentum (Attoh-Okine, 1999), but found that simple SGD performs best. We implemented dropout (Srivastava et al., 2014) and the local attentional model (Luong et al., 2015), but did not observe any significant improvements.

3.2 Structural bias components

As we are primarily interested in learning accurate alignments (roughly, attention), we include the modelling extensions of Cohn et al. (2016) for incorporating structural biases from word-based translation models into the neural attentional model. As shown later, we observe that including these components result in a substantial improvement in measured alignment quality. We now give a brief overview of these components.

Previous attention. In the basic attentional model, the alignment is calculated based on the source encoding H_S and the previous hidden state H_T^{i-1} of the target, $\alpha_i = \text{Attend}(H_T^{i-1}, H_S)$, where Attend is a function that outputs m attention coefficients. This attention mechanism is overly simplistic, in that it is incapable of capturing patterns in the attention over different positions i. Recognising and exploiting these kinds of patterns has proven critical in traditional word based models of translation (Brown et al., 1993; Vogel et al., 1996; Dyer et al., 2013). For this reason Cohn et al. (2016) include explicit features encoding structural biases from word based models, namely absolute and relative position, Markov conditioning and fertility:

1. previous alignment, α_{i-1}
2. sum of previous alignments, $\sum_{j=1}^{i-1} \alpha_j$
3. source index vector, $(1, 2, 3, \ldots, m)$; and
4. target index vector $(i, i, i, \ldots, i)$.

These features are concatenated to form a feature matrix $\beta \in \mathbb{R}^{4 \times m}$, which are added to the alignment calculation, i.e., $\alpha_i = \text{Attend}(H_T^{i-1}, H_S, \beta)$.

Coverage penalty. The sum over previous alignments feature, described above provides a basic fertility mechanism, however as it operates locally it is only partially effective. To address this, Cohn et al. (2016) propose a global regularisation method for implementing fertility.

Recall that the alignment matrix $\alpha \in \mathbb{R}^{n \times m}$, each α_i is normalized, such that $\sum_j \alpha_{ij} = 1$. However, nothing in the model requires that every source element gets used. This is remedied by encouraging the columns of the alignment matrix to also sum to one, that is, $\sum_i \alpha_{ij} = 1$. To do so, we add a regularization penalty, $\lambda \sum_{j=1}^m \left\| \sum_{i=1}^n \alpha_{ij} - 1 \right\|_2^2$ to the objective function where λ controls the regularization strength. We tune λ on the development set and found that $\lambda = 0.05$ gives the best performance.

[1] The LSTM also carries a memory cell, along with the hidden state; we exclude this from the presentation for clarity of notation.

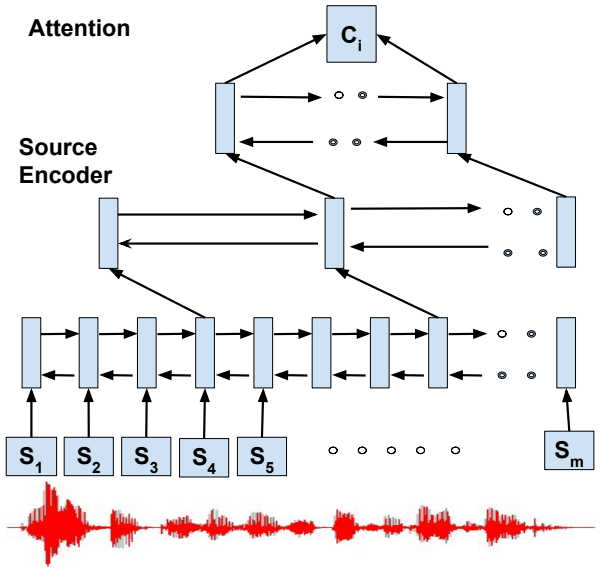

Figure 2: Stacking three layers of LSTM to the source side as in the second set of experiments

4 Extensions for Speech

We can easily apply the attentional model to parallel data, where the source side is represented as a sequence of phones. In cases where no annotated data or lexicon are available, we expect it is difficult to obtain phonetic transcriptions. Instead, we would like to work directly with the speech signal. However, dealing with the speech signal is significantly different than the phone representation, and so we need to modify the base attentional model.

4.1 Stacked and pyramidal RNNs

Both the encoder and decoder can be made more powerful by stacking several layers of LSTMs (Sutskever et al., 2014). For the first set of experiments below, we stack 4 layers of LSTMs on the target side; further layers did not improve performance on the development set.

For the second set of experiments, we work directly with the speech signal as a sequence of PLP vectors, one per frame. Since the frames begin at 10 millisecond intervals, the sequence can be very long. This makes the model slow to train; in our experiments, it seems not to converge at all. Following Chan et al. (2016), we use RNNs stacked into a pyramidal structure to reduce the size of the source speech representation. As illustrated in Figure 2, we stack 3 layers of bidirectional LSTMs. The first layer is the same as the encoder H_S described in Figure 1. The second layer uses every fourth output of the first layer as its input. The third layer selects every other output of the second layer as its input. The attention mechanism is applied only to the top layer. This reduces the size of the alignment matrix by a factor of eight, giving rise to vectors at the top layer representing 80ms intervals, which roughly correspond in duration to input phones.

4.2 Alignment smoothing

In most bitexts, source and target sentences have roughly the same length. However, for our task of aligning text and speech where the speech is represented as a sequence of phones or PLP vectors, the source can easily be several times larger than the target. Therefore we expect that a target word will commonly align to a run of several source elements. We want to encourage this behavior by smoothing the alignment matrix.

The easiest way to do this is by post-processing the alignment matrix. We train the model as usual, and then modify the learned alignment matrix α by averaging each cell over a window, $\alpha'_{ij} := \frac{1}{3}(\alpha_{i,j-1} + \alpha_{ij} + \alpha_{i,j+1})$. The modified alignment matrix, α', is only used for generating hard alignments in our alignment evaluation experiments. We can smooth further by changing the computation of α_{ij} during training. We flatten the softmax by adding a temperature factor, $T \geq 1$:

$$\alpha_{ij} = \frac{\exp(e_{ij}/T)}{\sum_k \exp(e_{ik}/T)}$$

Note that when $T = 1$ we recover the standard softmax function; we set $T = 10$ in both experiments.

5 Experimental Setup

We work on the Spanish CALLHOME Corpus (LDC96S35), which consists of telephone conversations between Spanish native speakers based in the US and their relatives abroad. While Spanish is not a low-resource language, we pretend that it is by not using any Spanish ASR or resources like transcribed speech or pronunciation lexicons (except in the construction of the "silver" standard for evaluation, described below). We also use the English translations produced by Post et al. (2013).

We treat the Spanish speech as a sequence of 39-dimensional PLP vectors (order 12 with energy and first and second order delta) encoding the power spectrum of the speech signal. We do not have gold standard alignments between the Spanish speech and English words for evaluation, so we produced "silver" standard alignments. We used a forced aligner (Gorman et al., 2011) to align the speech to its transcription, and GIZA++ with the `gdfa` symmetrization heuristic (Och and Ney, 2000) to align the Spanish transcription to the English translation. We then combined the two alignments to produce "silver" standard alignments between the Spanish speech and the English words.

Cleaning and splitting the data based on dialogue turns, resulted in a set of 17,532 Spanish utterances from which we selected 250 for development and 500 testing. For each utterance we have the corresponding English translation, and for each word in the translation we have the corresponding span of Spanish speech.

The forced aligner produces the phonetic sequences that correspond to each utterance, which we use later in our first set of experiments as an intermediate representation for the Spanish speech.

In order to evaluate an automatic alignment between the Spanish speech and English translation against the "silver" standard alignment, we compute alignment precision, recall, and F1-score as usual, but on links between Spanish PLP vectors and English words.

6 Phone-to-Word Experiments

In our first set of experiments, we represent the source Spanish speech as a sequence of phones. This sets an upper bound for our later experiments working directly on speech.

6.1 Alignment

We compare our model against three baselines: GIZA++, Model 3P, and pialign. For pialign, in order to better accommodate the different phrase lengths of the two alignment sides, we modified the model to allow different parameters for the Poisson distributions for the average phrase length, as well as different null alignment probabilities for each side.[2] We used the settings `-maxsentlen 200 -maxphraselen 20 -avgphraselenF 10 -nullprobF 0.001`, improving performance by 1.6% compared with the default setting. For Model 3P, we used the settings `-maxFertility 15 -maxWordLength 20`, unrestricted `max[Src/Trg]SenLen` and 10 `Model3Iterations`. We chose the iteration with the highest score to report as the baseline.

Model	F-score	Δ
GIZA++	29.7	−13.0
Model 3P	31.2	−11.5
Pialign (default)	42.4	−0.3
Pialign (modified)	44.0	+1.3
Base model	42.7	+0
+ alignment features	46.2	+3.5
+ coverage penalty	48.6	+5.9
+ stacking	46.3	+3.6
+ alignment smoothing	47.3	+4.6
+ alignment/softmax smoothing	48.2	+5.5
All modifications	53.6	+10.9

Table 1: On the alignment task, the base model performs much better than GIZA++ and Model 3P, and at roughly the same level as pialign; modifications to the model produce further large improvements. The Δ column shows the score difference compared with the base model.

The attentional model produces a soft alignment matrix, whose entries α_{ij} indicate $p(s_j \mid w_i)$ of aligning source phone s_j to target word w_i. For evaluation, we need to convert this to a hard alignment that we can compare against the "silver" standard. Since each word is likely to align with several phones, we choose a simple decoding algorithm: for each phone s_j, pick the word w_i that maximizes $p(w_i \mid s_j)$, where this probability is calculated from alignment matrix α using Bayes' Rule.

Table 1 shows the results of the alignment experiment. The base attentional model achieved an F-score of 42.7%, which is much better than GIZA++ and Model 3P (by 13% and 11.5% absolute, respectively) and at roughly the same level as pialign. Adding our various modifications one at a time

[2] Our modifications have been submitted to the pialign project.

aligner	decoder	reranker none	reranker AM
AM (all mods)		14.6	
GIZA++	Moses	18.2	19.9
pialign	Moses	18.9	19.8
pialign (mod)	Moses	20.2	21.1
Word-based Reference		34.1	

Table 2: BLEU score on the translation task. Using the attentional model (AM) alone (first row) significantly underperformed Moses. However, using the AM as a reranker yielded improvements across several settings. The word-based reference translation provides the upper bound for our phoneme-based systems.

yields improvements ranging from 3.5% to 5.9%. Combining all of them yields a net improvement of 10.9% over the base model, which is 9.4% better than the modified pialign, 22.4% better than Model 3P, and 23.9% better than GIZA++.

6.2 Translation

In this section, we evaluate our model on the translation task. We compare the model against the Moses phrase-based translation system (Koehn et al., 2007), applied to phoneme sequences. We also provide baseline results for Moses applied to word sequences, to serve as an upper bound. Since Moses requires word alignments as input, we used various alignment models: GIZA++, pialign, and pialign with our modifications. Table 2 shows that translation performance roughly correlates with alignment quality.

For the attentional model, we used all of the modifications described above except alignment smoothing. We also used more dimensions (256) for hidden states and memory cells in both encoder and decoder. The decoding algorithm starts with the symbol <s> and uses beam search to generate the next word. The generation process stops when we reach the symbol </s>. We use a beam size of 5, as larger beam sizes make the decoder slower without substantial performance benefits.

As shown in Table 2, the attentional model achieved a BLEU score of 14.6 on the test data, whereas the Moses baselines achieve much better BLEU scores, from 18.2 to 20.2. We think this is because the attentional model is powerful, but we don't have enough data to train it fully given that the output space is the size of the vocabulary. Moreover, this attentional model has been configured to optimize the alignment quality rather than translation quality.

We then tried using the attentional model to rerank 100-best lists output by Moses. The model gives a score for generating the next word $p(w_i|w_1 \cdots w_{i-1}, H_S)$ as in equation (1). We simply compute the score of a hypothesis by averaging the negative log probabilities of the output words,

$$\text{score}(w_1 \cdots w_n) = -\frac{1}{n}\sum_{i=1}^{n} \log(p(w_i|w_1 \cdots w_{i-1}, H_S)),$$

and then choosing the best scoring hypothesis. Table 2 shows the result using the attentional model as the reranker on top of Moses, giving improvements of 0.9 to 1.7 BLEU over their corresponding baselines. These consistent improvements suggest that the probability estimation part of the attentional model is good, but perhaps the search is not adequate. Further research is needed to improve the attentional model's translation quality. Another possibility, which we leave for future work, is to include the attentional model score as a feature in Moses.

Table 3 shows some example translations comparing different models. In all examples, it appears that using pialign produced better translations than GIZA++. Using the attentional model as a reranker for pialign further corrects some errors. Using the attentional model alone seems to perform the worst, which is evident in the third example where the attentional model simply repeats a text fragment (although all models do poorly here). Despite the often incoherent output, the attentional model still captures the main keywords used in the translation.

We test this hypothesis by applying the attentional model for a cross-lingual keyword spotting task where the input is the English keyword and the outputs are all Spanish sentences (represented as phones) containing a likely translation of the keyword. From the training data we select the top 200 terms as the keyword based on tf.idf. The relevance judgment is based on exact word matching. The attentional model achieved 35.8% precision, 43.3%

recall and 36.0% F-score on average on 200 queries. Table 4 shows the English translations of retrieved Spanish sentences. In the first example, the attentional model identifies *mañana* as the translation of *tomorrow*. In the second example, it does reasonably well by retrieving 2 correct sentences out of 3, correctly identifying *dejamos* and *salgo* as the translation of *leave*.

7 Speech-to-Word Experiments

In this section, we represent the source Spanish speech as a sequence of 39 dimensional PLP vectors. The frame length is 25ms, and overlapping frames are computed every 10ms. As mentioned in Section 4.1, we used a pyramidal RNN to reduce the speech representation size. Other than that, the model used here is identical to the first set of experiments.

Using this model directly for translation from speech does not yield useful output, as is to be expected from the small training data, noisy speech data, and an out-of-domain language model. However, we are able to produce useful results for the ASR and alignment tasks, as presented below.

	PER (%)
Our model	24.3
Our model + monotonic	22.3
Chorowski et al. (2014)	18.6
Graves et al. (2013)	17.7

Table 5: Phone-error-rate (PER) for various models evaluated on TIMIT

7.1 ASR Evaluation

To illustrate the utility of our approach to modelling speech input, first, we evaluate on the more common ASR task of phone recognition. This can be considered as a sub-problem of translation, and moreover, this allows us to benchmark our approach against the state-of-the-art in phone recognition. We experimented on the TIMIT dataset. Following convention, we removed all the `SA` sentences, evaluated on the 24 speaker core test set and used the 50 auxiliary speaker development set for early stopping. The model was trained to recognize 48 phonemes and was mapped to 39 phonemes for testing. We extracted 39 dimensional PLP features from the TIMIT dataset and trained the same model without any modification. Table 5 shows the performance of our model. It performs reasonably well compared with the state-of-the-art (Graves et al., 2013), considering that we didn't tune any hyper-parameters or feature representations for the task. Moreover, our model is not designed for the monotonic constraints inherent to the ASR problem, which process the input without reordering. By simply adding a masking function (equation 2 from Chorowski et al. (2014)) to encourage the monotonic constraint in the alignment function, we observe a 2% PER improvement. This is close to the performance reported by Chorowski et al. (2014) (Table 5), despite the fact that they employed user-adapted speech features.

7.2 Alignment Evaluation

We use alignment as a second evaluation, training and testing on parallel data comprising paired Spanish speech input with its English translations (as described in §5), and using the speech-based modelling techniques (see §4.) We compare to a naive baseline where we assume that each English letter (not including spaces) corresponds to an equal number of Spanish frames. The results of our attentional model and the baseline are summarized in Table 6. The attentional model is substantially lower than the scores in Table 1, because the PLP vector representation is much less informative than the gold phonetic transcription. Here, we have to identify phones and their boundaries in addition to phone-word alignment. However, the naive baseline does surprisingly well, presumably because our (unrealistic) choice of Spanish-English does not have very much reordering.

Figure 3 presents some examples of Spanish speech and English text, showing a heat map of the alignment matrix α (before smoothing). Due to the pyramidal structure of the encoder, each column roughly corresponds to 80ms. In the example on the left, the model is confident at aligning *a little* with columns 1–5, which corresponds roughly to their correct Spanish translation *algo*. We misalign the word *of* with columns 8–10, when the correct alignment should be columns 5–6, corresponding

Phones	sil e m e d i h o k e t e i B a a y a m a r sp a y e r o a n t e a y e r sp sil
Transcription	eh , me dijo que te iba a llamar , ayer , o anteayer
AM	eh , he told me that she was going to call , yesterday before yesterday
Giza	oh , he told me that you called yesterday or before yesterday .
Mod. Pialign	eh , she told me that I was going to call yesterday or before yesterday .
Mod. Pialign + AM	eh , he told me that I was going to call , yesterday or before yesterday .
Reference	eh , he told me that he was going to call you , yesterday , or the day before yesterday .
Phones	sil i t u k o m o a s e s t a D o h w a n i t o e s t a s t r a B a h a n d o k e sp e s t a s a s y e n d o sil
Transcription	y tú , cómo has estado , juanito , estás trabajando , qué estás haciendo.
AM	and how have you been working , are working ?
GIZA	and how are you Juanito , are you job , what are you doing ?
Mod. pialign	and how have you been Juanito are you working , what are you doing ?
Mod. pialign + AM	and how have you been Juan , are you working , what are you doing ?
Reference	and how have you been , Juanita , are you working , what are you doing .
Phones	sil t e n g o k e a s e r l e e l a s e o a s i k o m o a u n h a r D i n i n f a n t i l sp sil
Transcription	tengo que hacerle el aseo así como a un jardín infantil –
AM	I have to have to him like to like that to ⟨unkA⟩
GIZA	I have to do the , the how a vegetable information in the .
Mod. pialign	I have to do the that like to a and it was , didn't you don't have the .
Mod. pialign + AM	I have to make the or like to a and it was , didn't you don't have the –
Reference	I have to clean it like a kindergarten

Table 3: Translation examples for various models: the attentional model (AM), the standard Moses with GIZA++ aligner (giza), with modified Pialign aligner (Mod. pialign) and using the attentional model as reranker on top of pialign.

Keyword : **tomorrow**
El va **mañana** para Caracas. A qué va a Caracas él.
Y **mañana** , y **mañana** o pasado te voy a poner un paquete.
Oh , no , Julio no sé a dónde está y va **mañana** a Caracas , está con Richard.
Oye , qué bueno , entonces nos vamos tempranito en la **mañana**
No , aquí la gente se acuesta a las dos de la **mañana**.
Keyword : **leave**
Todo , organizar completo todo , desde los alquileres , la comida , mozo , cantina , todo lo pongo yo aquí
Y entonces dónde lo **dejamos** pagando estacionamiento y pagando seguro
Sí , el veintiuno. yo **salgo** de para aquí el dieciséis para florida , y el veintiuno llego a Caracas.

Table 4: Examples of cross-lingual keyword spotting using the attentional model. The bolded terms in the retrieved text are based on manual inspection.

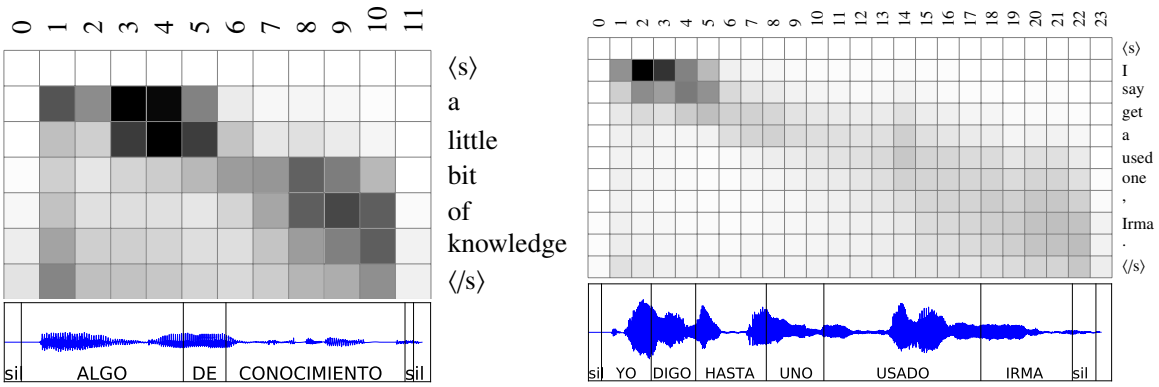

Figure 3: PLP-word alignment examples. The heat maps shows the alignment matrix which is time-aligned with the speech signals and their transcriptions.

ASR	aligner	F1
none	Naive baseline	31.7
none	AM (all mods)	26.4
cz	AM (all mods)	28.0
hu	AM (all mods)	27.9
ru	AM (all mods)	27.4
es	GIZA++	29.7

Table 6: Alignment of Spanish speech to English translations. In the first two rows, no gold or automatic transcriptions of any sort are used. In the next three rows, non-Spanish phone recognizers (cz, hu, ru) are used on the Spanish speech and the attentional model is run on the noisy transcription; this does better than no transcriptions. The last row is an unfair comparison because it uses gold Spanish (es) phonetic transcriptions; nevertheless, our model performs nearly as well.

to Spanish translation *de*. The word *knowledge* is aligned quite well with columns 7–10, corresponding to Spanish *conocimiento*. The example on the right is for a longer sentence. The model is less confident about this example, mostly because there are words that appear infrequently, such as the personal name *Irma*. However, we are still observing diagonal-like alignments that are roughly correct. In both examples, the model correctly leaves silence (`sil`) unaligned.

As a middle ground between assuming gold phonetic transcriptions (cf. Section 6) and no transcriptions at all, we use noisy transcriptions by running speech recognizers for other languages on the Spanish speech: Russian (ru), Hungarian (hu) and Czech (cz) (Vasquez et al., 2012). These distantly related languages were chosen to be a better approximation to the low-resource scenario. All three models perform better than operating directly on the speech signal (Table 6), and notably, the Russian result is nearly as good as GIZA++'s performance on gold phonetic transcriptions.

8 Conclusion

This paper reports our work to train models directly on parallel speech, i.e. source-language speech with English text translations that, in the low-resource setting, would have originated from spoken translations. To our knowledge, it is the first exploration of this type. We augmented the Spanish Fisher and CALLHOME datasets and extended the alignment F1 evaluation metric for this setting. We extended the attentional model of Bahdanau et al. to work on parallel speech and observed improvements relative to all baselines on phone-to-word alignment. On speech-to-word alignment, our model, without using any knowledge of Spanish, performs almost as well as GIZA++ using gold Spanish transcriptions.

Language pairs with word-order divergences and other divergences will of course be more challenging than Spanish-English. This work provides a proof-of-concept that we hope will spur future work towards solving this important problem in a true low-resource language.

Acknowledgments

This work was partly conducted during Duong's internship at ICSI, UC Berkeley and partially supported by the University of Melbourne and National ICT Australia (NICTA). We are grateful for support from NSF Award 1464553 and the DARPA LORELEI Program. Cohn is the recipient of an Australian Research Council Future Fellowship FT130101105.

References

Steven Abney and Steven Bird. 2010. The Human Language Project: Building a universal corpus of the world's languages. In *Proceedings of ACL*, pages 88–97.

Nii O. Attoh-Okine. 1999. Analysis of learning rate and momentum term in backpropagation neural network algorithm trained to predict pavement performance. *Advances in Engineering Software*, 30(4):291–302.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *Proceedings of ICLR*.

Steven Bird and David Chiang. 2012. Machine translation for language preservation. In *Proceedings of COLING*, pages 125–134, Mumbai, India.

Steven Bird, Lauren Gawne, Katie Gelbart, and Isaac McAlister. 2014. Collecting bilingual audio in remote indigenous communities. In *Proceedings of COLING*, pages 1015–1024.

Steven Bird. 2010. A scalable method for preserving oral literature from small languages. In *The Role of Digital Libraries in a Time of Global Change: 12th Inter-*

national Conference on Asia-Pacific Digital Libraries, pages 5–14, Berlin, Heidelberg. Springer-Verlag.

Peter E. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2).

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

William Chan, Navdeep Jaitly, Quoc V. Le, and Oriol Vinyals. 2016. Listen, attend and spell: A neural network for large vocabulary conversational speech recognition. In *Proceedings of ICASSP*.

Jan Chorowski, Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. End-to-end continuous speech recognition using attention-based recurrent NN: first results. In *Proceedings of NIPS Workshop on Deep Learning and Representation Learning*.

Jan K. Chorowski, Dzmitry Bahdanau, Dmitriy Serdyuk, Kyunghyun Cho, and Yoshua Bengio. 2015. Attention-based models for speech recognition. In *Proceedings of NIPS*, pages 577–585.

Trevor Cohn, Cong Duy Vu Hoang, Ekaterina Vymolova, Kaisheng Yao, Chris Dyer, and Gholamreza Haffari. 2016. Incorporating structural alignment biases into an attentional neural translation model. In *Proceedings of NAACL HLT*.

John Duchi, Elad Hazan, and Yoram Singer. 2011. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12:2121–2159.

Chris Dyer, Victor Chahuneau, and Noah A. Smith. 2013. A simple, fast, and effective reparameterization of IBM Model 2. In *Proceedings of NAACL HLT*, pages 644–648.

Kyle Gorman, Jonathan Howell, and Michael Wagner. 2011. Prosodylab-Aligner: A tool for forced alignment of laboratory speech. *Canadian Acoustics*, 39(3):192–193.

Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. 2006. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of ICML*, pages 369–376. ACM.

Alex Graves, Abdel-rahman Mohamed, and Geoffrey E. Hinton. 2013. Speech recognition with deep recurrent neural networks. In *Proceedings of ICASSP*, pages 6645–6649.

Hynek Hermansky. 1990. Perceptual linear predictive (PLP) analysis for speech. *Acoustical Society of America*, pages 1738–1752.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9(8):1735–1780.

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

Chia-ying Lee, Yu Zhang, and James Glass. 2013. Joint learning of phonetic units and word pronunciations for ASR. In *Proceedings of EMNLP*, pages 182–192.

Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective approaches to attention-based neural machine translation. In *Proceedings of EMNLP*, pages 1412–1421.

Andrew L. Maas, Ziang Xie, Dan Jurafsky, and Andrew Y. Ng. 2015. Lexicon-free conversational speech recognition with neural networks. In *Proceedings of NAACL HLT*, pages 345–354.

Graham Neubig, Taro Watanabe, Eiichiro Sumita, Shinsuke Mori, and Tatsuya Kawahara. 2011. An unsupervised model for joint phrase alignment and extraction. In *Proceedings of NAACL HLT*, pages 632–641.

Graham Neubig, Taro Watanabe, Shinsuke Mori, and Tatsuya Kawahara. 2012. Machine translation without words through substring alignment. In *Proceedings of ACL*, pages 165–174.

Franz Josef Och and Hermann Ney. 2000. Improved statistical alignment models. In *Proceedings of ACL*, pages 440–447.

Mohammad Pezeshki. 2015. Sequence modeling using gated recurrent neural networks. *arXiv preprint arXiv:1501.00299*.

Matt Post, Gaurav Kumar, Adam Lopez, Damianos Karakos, Chris Callison-Burch, and Sanjeev Khudanpur. 2013. Improved speech-to-text translation with the Fisher and Callhome Spanish–English speech translation corpus. In *Proceedings of IWSLT*.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15:1929–1958.

Felix Stahlberg, Tim Schlippe, Sue Vogel, and Tanja Schultz. 2012. Word segmentation through cross-lingual word-to-phoneme alignment. In *Proceedings of the IEEE Spoken Language Technology Workshop (SLT)*, pages 85–90.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *Proceedings of NIPS*, pages 3104–3112.

Henk Van den Heuvel, Khalid Choukri, Chr Gollan, Asuncion Moreno, and Djamel Mostefa. 2006. Tc-

star: New language resources for ASR and SLT purposes. In *Proceedings of LREC*, pages 2570–2573.

Daniel Vasquez, Rainer Gruhn, and Wolfgang Minker. 2012. *Hierarchical Neural Network Structures for Phoneme Recognition*. Springer.

Stephan Vogel, Hermann Ney, and Christoph Tillmann. 1996. HMM-based word alignment in statistical translation. In *Proceedings of COLING*, pages 836–841.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron C. Courville, Ruslan Salakhutdinov, Richard S. Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *Proceedings of ICML*, pages 2048–2057.

Matthew D. Zeiler. 2012. ADADELTA: an adaptive learning rate method. *arXiv preprint arXiv:1212.5701*.