interspeech 2008 incorporating SST 08

Language Model Adaptation for a Speech to Sign Language Translation System using Web Frequencies and a MAP Framework

Luis Fernando D'Haro[1], Rubén San-Segundo[1], Ricardo de Córdoba[1]
Jan Bungeroth[2], Daniel Stein[2], Hermann Ney[2]

[1] Speech Technology Group.- Dpto. Ingeniería Electrónica – ETSI de Telecomunicación
Universidad Politécnica de Madrid, Spain
[2]Lehrstuhl für Informatik 6, Computer Science Department, RWTH Aachen Universtity, Germany
{lfdharo, lapiz, cordoba}@die.upm.es, {bungeroth, stein, ney}@informatik.rwth-aachen.de

Abstract

This paper presents a successful technique for creating a new language model (LM) that adapts the original target LM used by a machine translation (MT) system. This technique is especially useful for situations where there are very scarce resources for training the target side (Spanish Sign Language (LSE) in our case) in order to properly estimate the target LM, the Sign Language Model (SLM), used by the MT system. The technique uses information from the source language, Spanish in our task, and from the phrase-based translation matrix in order to create a new LM, estimated using web frequencies, which adapts the counts of the SLM through the Maximum A Posteriori method (MAP). The corpus consists of common used sentences spoken by an officer when assisting people in applying for, or renewing, the National Identification Document. The proposed technique allows relative reductions of 15.5% on perplexity and 2.7% on WER for translation, which are close to half the maximum performance obtainable when only the LM is optimized.
Index Terms: language model adaptation, machine translation, sign language, web counts.

1. Introduction

Nowadays, the significant improvements in automatic speech recognition and statistical machine translation (SMT) have made possible to face new challenges such as speech to sign language translation, which is especially useful to help deaf people to communicate with non-deaf people. Besides, many deaf people have problems when reading lips, and even written texts, as they are used to the sign language grammar [13]. Several examples of recent studies in this area follow:

- In [7], Morrissey and Way describe corpus-based methods for example-based translation from English to the sign language of the Netherlands.
- In [9] and [10], San Segundo et al. describe a speech to gesture architecture, and compare three different MT approaches: rule-based, statistical phrase-based and stochastic finite state transducers.
- In [11], Stein et al. describe a German-to-German sign language for weather reports, where specific pre and post processing methods for improving the translation results are also described.

It is well know that in order to train efficiently any SMT it is necessary to have a big parallel corpus. Unfortunately, most of the currently available Sign Language (SL) corpora are too small or too general for training purposes. For example, [11] and [3] both consider corpora of about 2000 sentences; while [9][12] and [7] rely on corpora of only a few hundred sentences. In addition, it is too hard to find such kind of corpus available from online content that is usual in spoken languages. Currently, the most important available corpus for LSE is provided by Biblioteca Virtual Miguel de Cervantes; it consists of several videos with poetry, literature for kids and small pieces of classical Spanish books. Unfortunately, this corpus does not provide any transcriptions, just video content (that is common in most SL corpora), and it is very different from our current task domain. In addition, there is not a standard representation, or grammar, for the LSE, which makes the problem of data scarcity even worse.

Taking into account the problem of having a small parallel corpus for estimating a good SLM (i.e. reducing the effect of data sparseness) to ensure correct grammatical sentences during the MT process, we focused on creating a new adapted SLM. Our proposal relies in the adaptation of the original n-gram counts on the target side, using "translated" n-gram counts retrieved from the web. In order to do this, we use the phrase-based translation matrix to, in first place, select well-trained parallel n-grams, in the source side, whose counts are retrieved from the web, and second, to convert the retrieved web counts into target n-gram counts. Finally, MAP adaptation is used for merging both counts, and a linear interpolation is performed between the original and the target LM to improve the system's translations

The paper is organized as follows: section 2 provides an overview of LM adaptation techniques; and the phrase-based approach in SMT; section 3 describes the runtime system and the parallel corpus used in this task. In section 4 the proposed adaptation technique is described in detail, and section 5 presents the perplexity and translation results obtained; finally, section 6 outlines some conclusions and future work.

2. Background

2.1. Language Model Adaptation

One of the main problems when training n-gram based LMs is the data sparseness. In [2] several methods to overcome this problem are described. In most cases, the technique consists of building two LMs, one trained from the in-domain corpus and another from a background corpus (out-of-domain, or less specific corpus which is expected to be bigger than the in-domain one), and then applying an adaptation formula that modifies the well estimated background model using information from the in-domain model. Among the best adaptation technique, which operates at frequency count level, we have Maximum A-Posteriori (MAP) [1]. The adaptation is made using Eq. 1.

$$\Pr(w_q \mid h_q) = \frac{\alpha C^I(h_q w_q) + \beta C^O(h_q w_q)}{\alpha C^I(h_q) + \beta C^O(h_q)} \quad (1)$$

Here, C^I and C^O are the frequency counts for the in-domain and out-of-domain corpora for history h_q and n-gram $h_q w_q$ respectively. α and β are weight factors, estimated empirically, which reduce the bias of the estimators.

Like other techniques, MAP also needs a big background corpus to provide robustness to the in-domain LMs with n-grams from other domains. An alternative to generate the background corpus is to collect web frequency counts using information retrieval (IR) techniques. [5] and [14] report different experiments that confirm that LMs estimated using web frequency counts can be used for adaptation purposes providing comparable or better results. This paper follows a similar approach, but differs from these papers in the mechanism for selecting the n-grams to query the web, in the process of converting the n-gram counts from the source to the target side in a SMT system, and in the adaptation framework used (i.e. MAP).

2.2. Statistical Machine Translation

In automatic language translation, given a string in a source language, Spanish for this task, $f_1^J = f_1 \ldots f_J$,, into a target language, Spanish Sign Language here, $e_1^I = e_1 \ldots e_I$. Among all possible target strings, the system will choose the highest probability string, given by Bayes decision rule (Eq. 2).

$$\begin{aligned} \hat{e}_1^I &= \arg\max_{e_1^I} \{\Pr(e_1^I \mid f_1^J)\} \\ &= \arg\max_{e_1^I} \{\Pr(e_1^I)\Pr(f_1^J \mid e_1^I)\} \end{aligned} \quad (2)$$

Here, $\Pr(e_1^I)$ is the probability given by the target LM, whereas $\Pr(f_1^J|e_1^I)$ is the translation model. The arg max operation denotes the search problem, i.e. the generation of the output sentence in the target language. The language and the translation models provide complementary information that can be estimated individually.

Currently, one of the most widely SMT approaches is the phrase-based translation method [6]. It is done in three steps:

- Word alignment: The goal is to calculate the best alignments between words and signs. It was performed using GIZA++ [8] and optimized on the dev set.
- Phrase extraction: All phrase pairs that are consistent with the word alignment are collected. The maximum phrase size was fixed to seven. However, when creating the phrase table used for the LM adaptation, this parameter was set to three (in order to simplify the selection of the n-grams to be used to query the web).
- Phrase scoring: In this step, the translation probabilities are computed for all phrase pairs. Both translation probabilities are calculated: forward and backward.

Finding the best translation is equal to finding the best path, for which we employed a monotone search. For translation, we used the Pharaoh[1] toolkit that is a beam search decoder for phrase-based SMT models, and the LMs were trained using the SRILM toolkit[2].

[1] http://www.isi.edu/publications/licensed-sw/pharaoh/

[2] http://www.speech.sri.com/projects/srilm/

3. System and Corpus Description

3.1. Run-time system

It consists of three main modules [9]. In the first one, the sentence uttered by the non-deaf user is recognized. Then, the second module, the SMT system, translates the recognized utterance into a sequence of semantic symbols, called glosses, representing the grammar structure and sequence that follow the Spanish Sign Language. Finally, the third module is the avatar, we use VGuido[3], that uses a predefined dictionary to convert the sequence of glosses into an animated sequence of movements to play the sign. The glosses in the dictionary are defined and stored using HamNoSys and SiGML notation.

Table 1. Corpus statistics summary.

	Train		Dev and Test	
	Spanish	**LSE**	**Spanish**	**LSE**
Sentences Pairs	266		150	
Number of words	3153	2952	1776	1688
Vocabulary	534	292	427	250
OOV	0	0	90	30

3.2. Corpus description

For this task, the translation system is focused on a limited domain, consisting in common spoken sentences used when providing information for applying or renewing the National Identification Document (NID). In this context, a speech to sign language MT system is very useful because most of the officers do not know the Spanish Sign Language (LSE).

Table 1 shows the main statistics of this corpus. It sums up to 416 sentences that contain 624 different words, and were translated by hand, by an expert, into LSE, generating 322 different glosses. For example, the sentence "you will have to pay 20 euros as document fee" is translated into the following sequence of glosses "FUTURE YOU TWENTY EURO DOC_FEE PAY COMPULSORY", or "the NID must be renewed every five years" is translated into "EVERY FIVE PLURAL YEAR RENEW NID YOU COMPULSORY". Observe the order of the glosses and the semantic-like representation. The sentences were randomly divided into three sets, with 266 phrases for training. With the remaining sentences, we created three-fold cross validation sets leaving 50 sentences for development and 100 for test each time. For both text-to-sign and speech-to-sign translation purposes the same test and dev sets have been used.

4. Proposed Adaptation Technique

As mentioned before, the target LM is useful for ensuring that the translated sentences are well formed and fluent. In our task, it is a Sign Language Model (SLM). Unfortunately, the training corpus is very small and the LM probabilities cannot be reliably estimated. In this case, LM adaptation techniques with a large corpus are required. However, it was impossible to find an available background corpus to adapt with. However, since the source language is Spanish, we found that it was possible to take advantage of the phrase-based translation table created during the training of the MT model. The proposed technique is done in three steps:

[3] http://www.sign-lang.uni-hamburg.de/eSIGN/

Backward: To start with, the system uses the phrase pairs table created independently during the training of the translation probability $Pr(f_1^J|e_1^I)$ (Eq. 2). This table consists of a list of n-gram pairs that are consistent translations between the source and target language, with their probabilities $p(\bar{f}_i|\bar{e}_i)$ and $p(\bar{e}_i|\bar{f}_i)$, and lexical weights [6]. Using this table, the system creates a list of source-side n-grams, used in the next step, that satisfy $p(\bar{f}_i|\bar{e}_i) \geq \theta$. We decided to impose this threshold θ in order to reduce the number of n-gram pairs to be queried in the web, so that they are more reliable. In our experiments, θ was set to $1/n_i$, where n_i is the number of reverse translations for $\bar{f}_i$. However, it could be fixed as a function of the corpus size and the translation model quality. The final list consisted of 1270 n-grams (410 unigrams, 497 bigrams, and 362 trigrams).

Information Retrieval (IR): Using the n-gram list, the system queries the internet to obtain web frequency counts using the Google-API[4]. Then, a new source LM is created interpolating the original one (in-domain) and the MAP-adapted LM (Eq. 1).

Forward: Finally, the translation table is applied again, but on the opposite direction, to obtain the n-gram frequency counts on the target side. The conversion is done taking each n-gram pair in the list, $\bar{f}_i$, multiplying the retrieved web count, $N^{web}(\bar{f}_i)$, by the phrase translation probability, $p(\bar{e}_i|\bar{f}_i)$, and summing up all the contributions that satisfy $p(\bar{e}_i|\bar{f}_i) \geq \delta$, to obtain the counts for the target n-gram, $N(\bar{e}_i)$ (see Eq. 3). δ is set to $1/n_i$, the same as θ but on the target side. Then, Eq. 1 is applied to merge the counts from the original sign corpus with the converted counts. Finally, a new SLM is created form the linear interpolation of the original SLM and the merged counts.

$$C^O(\bar{e}_i) = \frac{\sum_{\forall \bar{e}_i: p(\bar{e}_i|\bar{f}_i) \geq \delta} N^{Web}(\bar{f}_i) * p(\bar{e}_i|\bar{f}_i)}{\sum_{\forall \bar{e}_i: p(\bar{e}_i|\bar{f}_i) \geq \delta} p(\bar{e}_i|\bar{f}_i)} \quad (3)$$

The weight factors from Eq. 1 were optimized on the dev sets (cross-fold) running a downhill simplex algorithm, resulting in the following average values for the source side: $\beta_s = 0.000417$, $\alpha_s = 36.7$ and $\lambda_s = 0.51$; and on the target side: $\beta_t = 0.0001$, $\alpha_t = 48$ and $\lambda_t = 0.52$. λ is the interpolation weight between the original and the adapted LMs.

We will use Table 2 to show an example of applying Eq. 3 for a trigram value. In this case, for the trigram of glosses "Tú NECESITAR ENTREGAR" (YOU DELIVER COMPULSORY) there are three suitable translations on the source side. Given the condition, $p(\bar{f}_i|\bar{e}_i) \geq \theta = 0.333$, the system only selects the n-grams pairs: one and three during the backward step (for the bigram case, θ = ¼ =0.25, it would select n-grams pairs b and c). In the forward step, using Eq. 3, and considering that the original count for the trigram gloss is 12, and for the bigram gloss it is 76, the out-of-domain trigram count will be C^O(trigram)=(135000*0.5 + 80420*0.4)/(0.5 + 0.4)= 110742, and the bigram count will be C^O(bigram)=(148000*0.3637+179000*0.41029)/(0.3637+0.4 1029) = 164433. Then, using Eq. 1 the final adapted count are C^{MAP}(trigram) = 48*12 + 0.0001*110742 = 587, and C^{MAP}(bigram) = 48*76 + 0.0001*164433= 3664. Using these adapted counts, the unsmoothed adapted trigram probability will be P(ENTREGAR | Tú NECESITAR) = 587/3664 = 0.160.

Table 2. Example of the phrase translation and counts retrieved from Google.

Source ($\bar{f}_i$)	**Target ($\bar{e}_i$)**	$p(\bar{e}_i\|\bar{f}_i)$	$p(\bar{f}_i\|\bar{e}_i)$
1.) necesitas entregar (you need to provide) Web Count:135000	**TRIGRAM:** Tú NECESITAR ENTREGAR	0.5	**1.0**
2.) que traer (to bring) Web Count: 206000		0.1	0.2
3.) tienes que entregar (you have to provide) Web Count: 80420	YOU DELIVER COMPULSORY	0.4	**0.5**
a.) que (that) Web Count: 26400000	**BIGRAM:** Tú NECESITAR YOU NEED	0.18	0.071
b.) tienes que (have to) Web Count: 148000		0.3637	**0.739**
c.) debes (must) Web Count: 190000		0.0460	0.143
d.) necesitas (you need) Web Count: 179000		0.4103	**0.952**

5. Experiments

5.1. Speech Recognition Results

The speech recognition system used in this section is a state of the art recognizer developed in our group [4]. The recognizer uses context-dependent continuous Hidden Markov Models (HMMs). These HMM models were trained with more than 40 hours of speech and 4000 different speakers from SpeechDat. In addition, CMN and CVN techniques were used to compensate differences in the acoustic channel. As front-end, it uses 13 PLP coefficients, plus delta and delta-delta coefficients summing up 39 parameters for each 10 ms frame. The original source LM used for speech recognition was a bigram language model due to the data sparseness. For the experiments, 15 speakers were recorded (8 males and 7 females). Each test sentence was uttered by 5 speakers, obtaining a total of 750 utterances. Table 3 shows the recognition results and the significant influence of a poorly trained LM (due to the small amount of data). With a robust LM, in a similar task [4], the recognition system has a 4.2% WER.

Table 3. Speech Recognition Results

WER	Ins (%)	Del (%)	Sub (%)
26.39	3.53	6.92	15.95

5.2. Language Model Experiments

Table 4 shows perplexities results provided by the baseline LMs and the adapted ones on train, dev and test sets. The results for the test and dev sets correspond to the averaged perplexities for the three-fold cross validation. The baseline LMs are backoff trigram with Good-Turing discount. The perplexities on both sides correspond to the adapted LMs.

[4] http://code.google.com/apis/ajaxsearch/

Values in parenthesis are relative improvements over the baseline perplexities.

Although the adaptation reduces perplexities in both sides, during the forward step, the translation table introduces some mismatch that reduces the improvement on the target side from 18.9% to 15.5% in the test set.

Table 4.Perplexity results

	Train		Dev		Test	
	Source	**Target**	**Source**	**Target**	**Source**	**Target**
Baseline	5.65	5.02	15.34	10.8	15.37	10.7
Adapted	3.01 (46.7%)	3.16 (37.1%)	11.92 (22.4%)	8.75 (18.7%)	12.45 (18.9%)	9.04 (15.5%)

5.3. Machine Translation Experiments

Table 5 shows the averaged MT results for the text-to-sign and speech-to-sign experiments on the test set for three different conditions we have considered. In Exp1, the system uses the baseline SLM. In Exp2, the system uses the adapted SLM. Finally, in Exp3 the SLM is trained considering all sentences (train, development and test sets). Since this model has all the available information, it corresponds to the top performance that it is possible to obtain only due to the LM component. The table shows the four common evaluation measures for assessing the quality of the obtained translations, i.e. WER (Word Error Rate), PER (Position Independent WER), BLEU (BiLingual Evaluation Understudy), and NIST. The former two are error measures (the higher the value, the worse the quality) whereas the latter two are accuracy measures (the higher, the better). We have also considered BLEU and NIST since we want to obtain similar translations to the ones created by the experts (see section 3.2).

Table 5. Machine translation results (Exp 1-3)

		WER	PER	BLEU	NIST
Text-to-Sign	Exp 1	34.74	29.59	0.50	6.30
	Exp 2	**33.79** (2.73%)	**29.1** (1.68%)	**0.51** (2.61%)	**6.36** (1.06%)
	Exp 3	32.62 (6.1%)	28.06 (5.48%)	0.55 (9.91%)	6.57 (4.23%)
Speech-to-Sign	Exp 1	42.87	38.94	0.43	5.65
	Exp 2	**42.53** (0.78%)	**38.57** (0.95%)	**0.44** (3.75%)	**5.70** (0.89%)
	Exp 3	41.43 (3.36%)	37.8 (2.9%)	0.47 (9.96%)	5.86 (3.62%)

For the text-to-sign MT system, the results show that the proposed technique is able to reach approximately half (2.73%) of the maximum improvement (6.1%) in WER that it is possible to obtain due only to the LM component.

From these experiments, it is possible to guess that the quality of the translation model limits significantly the improvement reached by better SLMs. This intuition was confirmed when we tested an optimal MT system, i.e. trained using all the available sentences. In this case, the WER was 13.06%. It is interesting to observe that for the speech-to-sign language experiments the improvements are lower. The probable explanation is that speech recognition introduces errors that affect some n-gram pairs, and so reduce the improvements of the target language model.

6. Conclusions

This paper has presented a successful technique to adapt LM for MT systems, which provides a relative improvement of 18.9% and 15.5% in perplexity over the base system for the source and target language respectively. The difference between both improvements are mainly due to the mismatch introduced by the translation table used to convert the frequencies retrieved from the web into frequencies on the target side. Besides, the MT experiments for text-to-sign provided a 2.73% relative reduction on WER that is near to half the performance that it is possible to achieve when only the LM is optimized. The MT experiments for speech-to-sign did not produce considerable improvements, which is probably due to the effect of recognition errors in the web counts for the n-grams with errors. As future tasks, we plan to work in improving the robustness against recognition errors, in applying more complex adaptation techniques, e.g. entropy models, topic adaptation, etc., and specific pre-processing techniques to improve the translation model.

7. Acknowledgements

This work has been supported by ANETO (CCG07-UPM/TIC-1823), ROBONAUTA (DPI2007-66846-c02-02) and EDECAN (TIN2005-08660-C04).

8. References

[1] Bacchiani, M., Riley, M., Roark, B., and Sproat, R. "MAP adaptation of stochastic grammars", Computer Speech & Language, 20(1):41–68, January 2006.

[2] Bellegarda, J. R. "Statistical language model adaptation: review and perspectives", Speech Communication, (42):93–108, 2004.

[3] Chiu, Y.-H., Wu, C.-H., Su, H.-Y., and Cheng, C.-J. "Joint Optimization of Word Alignment and Epenthesis Generation for Chinese to Taiwanese Sign Synthesis", IEEE Trans. Pattern Analysis and Machine Intelligence, 29(1):28–39, 2007.

[4] Córdoba, R., Macías-Guarasa, J., Sama, V., Barra, R., Pardo, J.M. "New Advances in Cross-Task and Speaker Adaptation for Air Traffic Control Tasks". Revista de Procesamiento del Lenguaje Natural Nº 35, pp. 21-27 ISSN 1135-5948, 2005.

[5] Keller, F., and Lapata, M. "Using the Web to Obtain Frequencies for Unseen Bigrams", Computational Linguistics, 29(3):459–484, 2003.

[6] Koehn, P., Och, F. J., and Marcu, D. "Statistical Phrase- Based Translation", HLT/NAACL 2003, pp. 48-54, Canada.

[7] Morrissey, S., and Way, A. "An Example-Based Approach to Translating Sign Language", Proc. Workshop Example-Based Machine Translation (MT X05), pp. 109-116, Thailand, 2005.

[8] Och, F. J., and Ney, H. "Improved Statistical Alignment Models", ACL 2000, pp. 440–447, Hong Kong, China.

[9] San-Segundo, R., Barra, R., D'Haro L. F., et al.. "A Spanish Speech to Sign Language Translation System for assisting deaf-mute people", Interspeech. 2006, pp. 1399–1402. USA.

[10] San-Segundo, R., Pérez, A., Ortiz, D., D'Haro, L. F., Torres, M. I., Casacuberta, F. "Alternatives on Speech to Sign Language Translation". Interspeech 2007, pp 2529-2532. Belgium.

[11] Stein, D., Bungeroth, J., and Ney H. "Morpho-Syntax Based Statistical Methods for Sign Language Translation", EAMT 2006. pp. 169–177, Oslo, Norway.

[12] Stein, D., Dreuw, P., Ney, H., Morrisey, S., and Way, A. "Hand in hand: Automatic Sign Language to English Translation". TMI 2007, pp.214-220.

[13] Zhao, L., Kipper, K., Schuler, W., Vogler, C., Badler, N. and Palmer, M. "A machine translation system from English to American Sign Language". AMTA, 2000. pp. 54-67.

[14] Zhu, X., and Rosenfeld, R. "Improving trigram language modeling with the world wide web", ICASSP, pp. 533–536, 2001.