THE MSR SYSTEM FOR IWSLT 2011 EVALUATION

Xiaodong He, Amittai Axelrod[1]*, Li Deng, Alex Acero, Mei-Yuh Hwang,*
Alisa Nguyen[2]*, Andrew Wang*[3]*, Xiahui Huang*[4]

Microsoft Research
One Microsoft Way, Redmond, WA 98052

{xiaohe, deng, alexac, mehwang}@microsoft.com, [1]amittai@u.washington.edu,
[2]alisanguyen@college.harvard.edu, [3]andrewkw@berkeley.edu, [4]xiahuihuang@gmail.com

Abstract

This paper describes the Microsoft Research (MSR) system for the evaluation campaign of the 2011 international workshop on spoken language translation. The evaluation task is to translate TED talks (www.ted.com). This task presents two unique challenges: First, the underlying topic switches sharply from talk to talk. Therefore, the translation system needs to adapt to the current topic quickly and dynamically. Second, only a very small amount of relevant parallel data (transcripts of TED talks) is available. Therefore, it is necessary to perform accurate translation model estimation with limited data. In the preparation for the evaluation, we developed two new methods to attack these problems. Specifically, we developed an unsupervised topic modeling based adaption method for machine translation models. We also developed a discriminative training method to estimate parameters in the generative components of the translation models with limited data. Experimental results show that both methods improve the translation quality. Among all the submissions, ours achieves the best BLEU score in the machine translation Chinese-to-English track (MT_CE) of the IWSLT 2011 evaluation that we participated.

1. Introduction

The IWSLT benchmark is an annual evaluation of spoken language translation (SLT) held by the International Workshop on Spoken Language Processing (IWSLT) [5]. The task of IWSLT2011 has been the translation of TED talks (www.ted.com). TED talks are given by leaders in various fields and cover an open set of topics in Technology, Entertainment, Design, and other domains. Compared with conventional machine translation tasks, this task presents two unique challenges: First, the underlying topic switches sharply from talk to talk, and each talk contains only tens to hundreds of utterances. Therefore, the system needs to adapt to the current topic dynamically and automatically. Second, unlike text based machine translation where a large parallel training corpus is often available, there is only a small amount of talk-style parallel data consisting of human translations of TED talks. Therefore, methods of estimating accurate translation models from limited parallel data are needed.

In this paper, we present the Microsoft Research (MSR) system on the IWSLT2011 TED talk translation task. In order to address the first problem, we use a topic model-based method for fast unsupervised topic adaptation. Machine translation systems are more effective when used to translate input that closely matches the training and tuning data. Here the wide-ranging subject of the talks contraindicates the use of a single domain-specific system for the task. A topic model [2] is a generative model for explaining broad topical variety in a corpus. The importance of this model is that it is unsupervised, and that after training it can be used to perform statistical inference on the new input. This allows previously-unheard utterances to be related to the topics learned during training. In the past, topic models have been used to select additional monolingual data to create a topic-specific language model [19], and these models have been applied to the task of statistical machine translation (SMT) [17][18]. Combining topic models with prior work on selecting relevant out-of-domain sub-corpora [1][7], we propose a method for selecting additional parallel corpora using an unsupervised topic model. In IWSLT2011, we have submitted the topic-adaptive phrase-based translation system as our contrastive system 2.

In order to address the second challenge, we develop a discriminative training method to estimate the translation channel models more accurately. The machine translation problem is commonly modeled by a log-linear model with multiple features that capture different dependencies between the source language and the target language [15]. Although the log-linear model is discriminative in nature, many of the feature functions, such as the phrase-level translation probability features and the lexicon-level translation probability features (e.g., lexical weighting), are derived from generative models. Further, these features are usually trained by conventional maximum likelihood (ML) estimation [11]. In the case of sparse training data, the ML estimation could lead to sub-optimal distribution [10]. In order to address this problem, we introduce a discriminative training method for these generative translation models based on a technique called growth transformation (GT). In IWSLT2011, we have submitted a phrase-based system with discriminative translation models as our contrastive system 1.

Our primary submission is a combination of four systems, including the topic-adaptive system and the discriminative translation model system described above, plus a regular phrase-based machine translation system [11] and a Hiero system [3]. System combination is performed based on the incremental indirect hidden Markov model proposed in [20][21].

2. Data

For training, we use exclusively the monolingual and parallel texts supplied by the evaluation campaign. No additional

[1,2,3,4] The work was performed when Amittai Axelrod, Alisa Nguyen, Andrew Wang, and Xiahui Huang were interns at Microsoft Research.

datasets, web data, or other resources were used.

2.1. TED relevant training data

The TED parallel corpus consists of about 110K sentences of English transcription and their Chinese translation of archived TED talks (www.ted.com) as provided by the IWSLT evaluation campaign.

2.2. Supplementary training data

In addition, the IWSLT evaluation campaign also provides out-of-domain data for potential usage. These include about 7.7M parallel sentences of UN proceedings and 115M of monolingual English sentences, mainly from the EuroMatrixPlus project, Europarl corpus, and LDC Gigawords corpus [5].

2.3. Development data

The evaluation campaign provides two sets of development data, namely, dev2010 and tst2010. A summary of these two development data sets are presented in table 1.

Table 1. Development sets.

Data set	# sentences	OOV
Dev2010	934	1.31%
Tst2010	1664	0.67%

3. System Details

3.1 MSR Phrase-based translation system

The MSR phrase-based translation system is implemented as described in [11][23], e.g.,

$$\hat{E} = \underset{E}{\operatorname{argmax}}\, P(E|F) \tag{1}$$

where

$$P(E|F) = \frac{1}{Z} exp\left\{\sum_i \lambda_i log\varphi_i(E,F)\right\} \tag{2}$$

where $Z = \sum_E exp\{\sum_i \lambda_i log\varphi_i(E,F)\}$ being the normalization denominator to ensure that the probabilities sum to one. In the log-linear model, $\{\varphi_i(E,F)\}$ are the feature functions constructed from *E* and *F*.

In our system, features include hypothesis length, number of phrases, lexicalized reordering model scores, language model scores, and translation model scores. Details of these models are described in the following sections.

3.1.1 Translation phrase tables

In our system, we first perform word alignment on the TED parallel corpus using the word-dependent HMM-based alignment method proposed in [6]. Then, a phrase table is constructed from the word aligned TED corpus as described in [11]. In the phrase table, each phrase pair has four translation model scores. They are:

- Forward phrase translation feature: $\varphi_{F2Eph}(E,F,X) = P_{TMph}(E|F) = \prod_k p(\tilde{e}_k|\tilde{f}_k)$, where $\tilde{e}_k$ and $\tilde{f}_k$ are the *k*-th phrase in *E* and *F*, respectively, and $p(\tilde{e}_k|\tilde{f}_k)$ is the probability of translating $\tilde{f}_k$ to $\tilde{e}_k$. This is usually modeled by a multinomial model.
- The backward phrase translation feature is defined similarly.
- Forward word translation feature: $\varphi_{F2Ewd}(E,F,X) = P_{TMwd}(E|F) = \prod_k \prod_m \sum_n p(e_{k,m}|f_{k,n})$, where $e_{k,m}$ is the *m*-th word of the k-th target phrase $\tilde{e}_k$, $f_{k,n}$ is the *n*-th word in the *k*-th source phrase $\tilde{f}_k$, and $p(e_{k,m}|f_{k,n})$ is the probability of translating word $f_{k,n}$to word $e_{k,m}$. (This is also referred to as the lexical weighting feature.) Note, although this feature is derived from the probability distribution $\{ p(e_{k,m}|f_{k,n}) \}$ which is modeled by a multinomial model.
- The backward word translation feature is defined similarly.

In order to mitigate the data sparse issue, we also selected 500K of TED-like parallel sentences from the supplied UN parallel corpus based on the bi-lingual cross-entropy data selection method as described in [1]. Then, an additional phrase table was constructed based these 500K sentences of UN data. Both TED and UN phrase tables are integrated into the log-linear model at decoding.

3.1.2 Language Models

Two language models (LM) are used in our system. The first is a 3-gram LM trained on the English side of the TED parallel corpus. In addition, we also trained a LM based on the 115M of monolingual English sentences. Since there are much more data in this monolingual English dataset, a 5-gram LM can be trained to capture longer contextual information without severe data sparsity issue. Both LMs use Kneser-Ney smoothing.

3.1.3 Tuning of Lambdas

The linear weights of these features, e.g., $\boldsymbol{\lambda} = \{\lambda_i\}$, are tuned by minimum error rate training (MERT) [14]:

$$\hat{\boldsymbol{\lambda}} = \underset{\boldsymbol{\lambda}}{\operatorname{argmax}}\, BLEU\left(\hat{E}(\boldsymbol{\lambda}, X), E^*\right) \tag{3}$$

where E^* is the translation reference(s), and $\hat{E}(\boldsymbol{\lambda}, X)$ is the translation output. In our system, dev2010 is used for MERT training.

3.2 Topic-Adaptive Phrase-based translation system

Latent Dirichlet Allocation (LDA) [2] is a probabilistic topic model for decomposing the content of a (heterogeneous) corpus according to some number of topics *K*. In particular, for a fixed number of topics, each part of the corpus is assumed to reflect some combination of all of those topics. Probabilistic inference can then be used to extract an underlying topical structure from the corpus. One advantage to topic models is that they can be trained in an unsupervised manner, using freely-available toolkits such as MALLET [13].

Let *P*(z) be the distribution, over all *Z* topics, in a particular utterance *W* which consists of words *w*. In LDA, *P*(z) is taken to have a Dirichlet distribution. Now let *P*(*w*|*z*) is the probability distribution of words given the particular topic *z*. The generative story of probabilistic topic models supposes that each word *w* in an utterance is produced by first sampling a topic *z* from *P*(*z*), and then selecting a word *w* according to *P*(*w*|*z*). The probability of a word within an utterance is thus:

$$P(w) = \sum_{k=1}^{K} P(w|z=k)P(z=k) \quad (4)$$

Once the topic model has been trained, it can be used to infer the topic mixture of new utterances. These topic scores can be used to cluster the new input relative to the existing *K* topics. Prior work has shown that the data in each topical cluster in a corpus can be used to train targeted language models which outperform the general corpus-wide model on topic-specific input [19]. This approach has applied to Statistical Machine Translation (SMT) as well, whereby language models are adapted to the parallel corpus topics and added to the system to improve translation performance [17] [18].

In this work we instead consider the case where there is both an in-domain and an out-of-domain bilingual parallel corpus. Rather than adapting a topical language model to use in combination with a background model, we wish to identify parts of the external parallel corpus that are similar to the individual topics in the in-domain corpus. The 2011 IWSLT task included the use of 7.7 million sentences of parallel UN data, which can be considered out-of-domain relative to the TED talks in the training corpus. Our experiments show that the UN corpus, when used in its entirety as a second translation model, does not positively impact translation. However, prior work by [1] shows that relevant subsets of an unrelated corpus can be more beneficial for training a second translation model than using the entire additional corpus. This motivates the use of a topic model trained on the input (Chinese) side of the TED talks to select the most relevant subset of the UN corpus for each particular topic, based on thresholding the scores of the single-most-likely topic. In this way, the UN parallel corpus is trimmed to four pieces totaling the 1.4M most topically-relevant sentences. Each of these topic-specific subsets is used to train a topic-specific translation model. The TED training corpus for IWSLT is not large enough to split into topics that are big enough to use to train a reasonable translation model, so all the TED data is used together as a general TED-domain model and adaptation is performed by using a different subset of the UN data to train the topic-adapted model.

The tuning and evaluation data was split into topics via the same model that had been trained on the TED data, and assigning it to the single most likely topic. Even concatenating the 2010 dev and test sets, we were limited to 4 topics to keep each topical tuning set be large enough to prevent overfitting. Each topical subset of the input data was decoded using the corresponding topical model. During MERT learning and run-time testing, two translation models, one general and one topic-specific, were used in combination with two language models trained on the in-domain data and some additional monolingual data. These four models were tuned for each topic in a log-linear combination.

3.3 Discriminative translation model based phrasal system

Although the log-linear model of (2) is discriminative in nature, many of the feature functions, such as the translation models based features, are derived from generative models. Conventionally, these features are usually trained by maximum likelihood (ML) estimation [11]. However, when data are sparse, the ML training could lead to sub-optimal estimation of probability distributions [10].

Recently, effort has been made to further extend the max-BLEU training method. In [12], model parameters are optimized with a perceptron using the best possible translation hypothesis as the approximated reference. On the other hand, in [4], the linear model is extended to include tens of thousands of fine-grained features, where most of them are binary indicators. In order to effectively training the weights of this many features, an MIRA-based optimization method is used.

In this work, we introduce a discriminative training method for the estimation of translation models based on a technique called growth transformation (GT) [9]. Unlike [12], we use the expected BLEU score as the objective function and the true reference is used without approximation. Compared to [4], our focus is on discriminative training of the phrase and lexicon translation probability distributions. With our method, we can train tens of millions of parameters effectively.

Let Λ denote the full parameter set of the translation models. The objective function of our method is expected BLEU:

$$O(\Lambda) = \sum_{E} p(E|F,\Lambda) C_{DT}(E) \quad (5)$$

where $C_{DT}(E)$ is the evaluation metric, which for translation is BLUE score. In this work, we adopt:

$$C_{DT}(E) = \sum\nolimits_{r} BLEU(E_r, E_r^*) \quad (6)$$

Optimization of the objective function is discussed in [8] and comprehensive study will be detailed in a future paper. In the following, we just present the preliminary estimation formula for the phrase and lexicon translation models directly. Using the backward phrase translation model as an example, the GT formula is:

$$p(\tilde{f}|\tilde{e},\Lambda) = \frac{\sum_k \sum_{\substack{E,F:\\ e_k=\tilde{e}\\ f_k=\tilde{f}}} p(F|E,\Lambda')\Delta_E + D_{\tilde{e}} \cdot p(\tilde{f}|\tilde{e},\Lambda')}{\sum_k \sum_{\substack{E:\\ e_k=\tilde{e}}} p(F|E,\Lambda')\Delta_E + D_{\tilde{e}}} \quad (7)$$

where $\Delta_E = [C(E) - O(\Lambda')]$ and $D_{\tilde{e}}$ is a constant independent of Λ. In our implementation, the following formula is used to compute $D_{\tilde{e}}$:

$$D_{\tilde{e}} = \tau + \rho \cdot \sum_{k} \sum_{\substack{E,F:\\ e_k=\tilde{e}}} p(F|E,\Lambda') \max\{O(\Lambda') - C(E), 0\} \quad (8)$$

We set τ to be a small positive value and $\rho \geq 1$, so that the denominator of (7) is guaranteed to be positive. The forward phrase translation model has a similar GT estimation formula and will be omitted here. For the backward lexical weighting feature, the GT formula for the lexicon translation model $p(g|h,\Lambda)$ is:

$$p(g|h,\Lambda) = \frac{\sum_{\substack{k,m:\\ f_{k,m}=g}} \sum_{E,F} p(E,F|X,\Lambda')\Delta_E \gamma_h(k,m) + D_h \cdot p(g|h,\Lambda)'}{\sum_{k,m} \sum_{E,F} p(E,F|X,\Lambda')\Delta_E \gamma_h(k,m) + D_h} \quad (9)$$

where

$$\gamma_h(k,m) = \frac{\sum_{n:e_{k,n}=h} p(f_{k,m}|e_{k,n},\Lambda')}{\sum_n p(f_{k,m}|e_{k,n},\Lambda')} \quad (10)$$

and D_h is set in a similar way as (8). Again, the forward word translation model has a similar GT estimation formula.

3.4 Hiero system

We also implemented the hierarchical phrase-based system as described in [3]. It uses a statistical phrase-based translation model that uses hierarchical phrases. The model is a synchronous context-free grammar and it is learned from parallel data without any syntactic information.

In this system, only one phrase table is used, which is estimated from the TED parallel corpus. Then, we merged the English side of the TED parallel corpus and the 115M WMT11 sentences to form one big corpus, and trained a 5-gram LM from it.

3.5 System Combination

In testing, each of these four system produced 10-best output. Then, we combined these output based on the Incremental indirect hidden Markov model proposed in [20][21]. The system combination parameters are tuned on a big tuning set, i.e., the concatenation of dev2010 and tst2010.

3.6 Case restoration

In our system, a language model based truecaser is used. The LM is trained on the original (cased) English transcript of the TED corpus. Further, the cases of the original English words embedded in the input Chinese sentences, mostly people names or acronyms, are kept with no change.

4. Submissions

MSR has participated in both the machine translation Chinese-to-English track (MT_CE) and the machine translation sysem combination Chinese-to-English track (MT_SC_CE).

4.1 Submissions to the MT_CE track

For the MT_CE track, we submitted one primary submission and two contrastive submissions. The primary submission is a combination of the four single systems described above. The contrastive-1 system is a single phrase-based system with discriminative translation models, which is also the best one in the four single systems we built. The contrastive-2 system is a single phrase-based system with adaptive translation models. Their performances on the IWSLT2011 test set are tabulated in Table 2.

Table 2. Performance of MSR MT_CE submissions

submission	BLEU case+punc	BLEU no_case+no_punc
primary	0.1689	0.1545
contrastive-1	0.1592	0.1463
contrastive-2*	0.1345	0.1171

* Due to the lack of resource, contrastive-2 system uses only 1% of the supplied monolingual English data for the second LM.

4.2 Submissions to the MT_SC_CE track

There are a total of five primary submissions from different sites in the MT_CE track. The translations of these five entries are used for system combination in the MT_SC_CE track. In addition, the participants are suggested to submit a preliminary run on the dev2010 and tst2010 data sets in August so that these preliminary submissions can be used to tune the system combination parameters. However, only four sites submitted the output in the preliminary run. Moreover, it was found that there is severe mismatch between performances of individual systems in the preliminary run and the formal evaluation. For example, after comparing the relative rank of the performance of the four systems in the preliminary run and the formal run (the latter is from a notice provided to the participants of the MT_SC_CE track by the organizer), system-1 seems improved significantly after the preliminary run. These issues make the tuning of the combination parameters difficult.

In the MSR submission, we submitted one primary submission and two contrastive submissions. In all three submissions, only the translations from the four sites who have submitted preliminary runs are used for combination.

In our primary submission, we jointly optimize the word alignment, ordering, and lexical selection decisions according to a set of feature functions combined in a single log-linear model as described in [22]. Regarding tuning of combination parameters, due to the severe mismatch of performances of individual systems in the preliminary run and the formal evaluation, the system weights estimated from the preliminary run is not reliable. Therefore, in our primary run, we heuristically set the system weights (according to the rank of systems in the formal run from a notice by the organizer), i.e., 0.25 : 0.20 : 0.35 : 0.20. All other parameters such as LM weight, word-voting weight etc. are still tuned on the data of the preliminary run. In contrast, contrastive-1 uses system weights trained on the preliminary run. On the other hand, contrastive-2 also uses system weights trained on the preliminary run and use the incremental HMM based combination method[21].

The performances of the four single systems and combined systems are given in table 3. As shown in the table, no significant gain is obtained by system combination, and the performance even degrades for the two contrastive systems. This may indicate that, due to the mismatch of performances of individual systems at the preliminary run (i.e., used for tuning of system combination parameters) and the formal run, the system combination parameters are severely twisted and are no longer suitable for combing the four systems at the formal run.

Table 3. Performance of MSR MT_CE submissions

system	BLEU case+punc	BLEU no_case+no_punc
System-1	0.1513	0.1361
System-2	0.1212	0.1130
System-3	0.1689	0.1545
System-4	0.1315	0.1178
MSR-Comb-p	0.1702	0.1565
MSR-Comb-c1	0.1662	0.1524
MSR-Comb-c2	0.1637	0.1505

5. Summary and Discussion

The 2011 IWSLT evaluation results validate the effectiveness of two new methods that we developed recently. In particular, the major gain has been achieved using the

discriminative learning method based on a comprehensive theoretical framework and optimization technique [8][9]. While the evaluation we participated is text translation only, its effectiveness provides an indirect evidence that its extension to speech translation will be promising, which is a more natural task targeted by our theoretical framework presented in [8]. For the method of topic adaptation, with more data available, we expect the adaptation technique will show greater strength than presented in this paper.

6. Acknowledgements

We would like to thank the organization committee of the IWSLT2011 evaluation campaign that makes the evaluation presented in this paper possible.

7. References

[1] A. Axelrod, X. He, J. Gao. "Domain adaptation via pseudo in-domain data selection". Proceedings of Empirical Methods in Natural Language Processing, 2011.

[2] D. Blei, A. Ng, M.I. Jordan, "Latent Dirichlet allocation," Journal of Machine Learning Research, 2003.

[3] D. Chiang. "A hierarchical phrase-based model for statistical machine translation." In Proc. of ACL, 2005.

[4] D. Chiang, K. Knight and W. Wang, "11,001 new features for statistical machine translation," in *Proc. NAACL-HLT*, 2009.

[5] M. Federico, L. Bentivogli, M. Paul, and S. Stueker, "Overview of the IWSLT 2011 Evaluation Campaign", In Proc. IWSLT, 2011

[6] X. He, "Using word-dependent transition models in HMM-based word alignment for statistical machine translation," *Proc. ACL-WMT*, 2007.

[7] X. He and L. Deng, "Robust speech translation by domain adaptation." in *Proc. Interspeech*, 2011

[8] X. He and L. Deng. "Speech Recognition, Machine Translation, and Speech Translation – A Unified Discriminative Learning Paradigm" IEEE Sig, Proc. Mag., Sept 2011.

[9] X. He, L. Deng, W. Chou, "Discriminative learning in sequential pattern recognition." *IEEE Sig. Proc. Mag*., Sept., 2008.

[10] B.-H. Juang, W. Chou, and C.-H. Lee, "Minimum classification error rate methods for speech recognition," *IEEE Trans. Speech Audio Processing*, May 1997.

[11] P. Koehn, F. Och, and D. Marcu. "Statistical phrase-based translation," *Proc. HLT-NAACL*, 2003

[12] P. Liang, A. Bouchard-Cote, D. Klein and B. Taskar, "An end-to-end discriminative approach to machine translation," in *Proc. COLING-ACL*, 2006

[13] A. McCallum, "MALLET: A machine learning for language toolkit", http://mallet.cs.umass.edu, 2002.

[14] F. Och, "Minimum error rate training in statistical machine translation." *Proc. ACL*, 2003.

[15] F. Och and H. Ney. "Discriminative training and maximum entropy models for statistical machine translation." In Proc. ACL 2002.

[16] K. Papineni, S. Roukos, T. Ward, and W. Zhu, "BLEU: A method for automatic evaluation of machine translation." *Proc. ACL*, 2002.

[17] N. Ruiz, M. Federico. "Topic adaptation for lecture translation through bilingual latent semantic models". In *Proc of WMT* 2011.

[18] Y.C. Tam, I. Lane, T. Schultz, "Bilingual-LSA based LM adaptation for spoken language translation". In *Proc of ACL*, 2007.

[19] Y.C. Tam and T. Schultz, "Unsupervised language model adaptation using latent semantic marginals", In *Proc of Interspeech*, 2006.

[20] X. He, M. Yang, J. Gao, P. Nguyen, R. Moore. "Indirect-HMM-based Hypothesis Alignment for Combining Outputs from Machine Translation Systems." In Proc. EMNLP, 2008

[21] C-H. Li, X. He, Y. Liu and N. Xi. "Incremental HMM Alignment for MT System Combination." In Proc. ACL, 2009

[22] X. He and K. Toutanova. "Joint Optimization for Machine Translation System Combination." In Proc. EMNLP, 2009.

[23] R. Moore and C. Quirk. "Faster Beam-Search Decoding for Phrasal Statistical Machine Translation." In Proc. of MT Summit XI, 2007.