The NTT Statistical Machine Translation System for IWSLT2005

Hajime Tsukada, Taro Watanabe, Jun Suzuki, Hideto Kazawa, and Hideki Isozaki

NTT Communication Science Labs.

Purpose

- A large number of reportedly effective features is evaluated by our system.
- Additional monolingual and bilingual resources are also evaluated.
  - Monolingual resources for generated language modeling
  - Bilingual resources for translation modeling

SMT based on Log-linear Models [Och, 2002][Och, 2003]

$$\hat{e}_1^I = \underset{e_1^I}{\operatorname{argmax}} \frac{1}{Z(f_1^J)} \exp\left(\sum_j \lambda_j f_j(e_1^I, f_1^J)\right)$$

- $f_j(e_1^I, f_1^J)$: feature functions
- λ_j is calculated based on the minimum error rate criterion in our system.

Easy to combine various features for translation modeling, language modeling, and lexical reorder modeling

Language Model Features

- Features:
  - 6-gram
  - Class-based 9-gram
  - Prefix-4 9-gram
  - Suffix-4 9-gram

- Training Conditions:
  - Mixed casing
  - Prefix-4 (suffix-4) takes only 4-letter prefixes (suffixes) [Och, 2005].

Examples of prefix-4

I 'd like to reserve -> I 'd like to rese+

I 'd like to make a reservation -> I 'd like to make a rese+

Phrase-based Features

- Phrase translation probabilities, $\phi(\tilde{e}|\tilde{f})$ and $\phi(\tilde{f}|\tilde{e})$:

$$\phi(\tilde{e}|\tilde{f}) = \frac{\text{count}(\tilde{f}, \tilde{e})}{\sum_{\tilde{f}} \text{count}(\tilde{f}, \tilde{e})}$$

- χ^2(f, e)
- Dice(f, e)

Phrase Based Features (cont'd)

- Phrase extraction probability of source/target:

$$\frac{\#\text{ of extracted source/target phrases}}{\#\text{ of source/target phrases appearing in the corpus}}$$

- Phrase pair extraction probability:

$$\frac{\#\text{ of sentences phrase pairs extracted}}{\#\text{ of sentences phrase pairs appearing in the corpus}}$$

Phrase Based Features (cont'd)

- Adjusted Dice coefficient:

$$\text{Dice}(\tilde{f}, \tilde{e}) log(\text{count}(\tilde{f}, \tilde{e}) + 1)$$

Word-level Features

- Lexical weights, $p_w(\tilde{f}|\tilde{e})$ and $p_w(\tilde{e}|\tilde{f})$, where

$$p_w(\tilde{f}|\tilde{e}) = max_a \prod_{j=1}^{J} \frac{1}{|\{i|(i,j) \in a)\}|} \cdot \sum_{\forall(i,j)\in a} w(f_j|e_i)$$

$$w(f|e) = \frac{\text{count}(f,e)}{\sum_{f'} \text{count}(f',e)}$$

Word-level Features (cont'd)

- IBM model 1 scores, $p_{M1}(\tilde{f}|\tilde{e})$ and $p_{M1}(\tilde{e}|\tilde{f})$, where

$$p_{M1}(\tilde{f}|\tilde{e}) = \frac{1}{(\tilde{I}+1)^{\tilde{J}}} \prod_{j}^{\tilde{J}} \sum_{i}^{\tilde{I}} w(\tilde{f}_j|\tilde{e}_i)$$

Word-level Features (cont'd)

- Viterbi IBM model 1 scores, $p_{M1'}(\tilde{f}|\tilde{e})$ and $p_{M1'}(\tilde{e}|\tilde{f})$, where

$$p_{M1'}(\tilde{f}|\tilde{e}) = \frac{1}{(\tilde{I}+1)^{\tilde{J}}} \prod_{j}^{\tilde{J}} \max_{i} w(\tilde{f}_j|\tilde{e}_i)$$

Word-level Features (cont'd)

- Noisy OR gates, $p_{NOR}(\tilde{f}|\tilde{e})$ and $p_{NOR}(\tilde{e}|\tilde{f})$, where

$$p_{NOR}(\tilde{f}|\tilde{e}) = \prod_j (1 - \prod_i (1 - w(\tilde{f}_j|\tilde{e}_i)))$$

Word-level Features (cont'd)

- Deletion penalty, $p_{del}(\tilde{e}, \tilde{f})$, where

$$p_{del}(\tilde{e}, \tilde{f}) = \sum_j del(\tilde{e}_1^{\tilde{I}}, \tilde{f}_j)$$

$$del(\tilde{e}_1^{\tilde{I}}, \tilde{f}_j) = \begin{cases} 1 & i \text{ does not exist s.t.} \\ & w(\tilde{e}_i|\tilde{f}_j) > threshold \\ 0 & \text{otherwise.} \end{cases}$$

Lexical Reordering Features

- Distortion model $d(a_i - b_{i-1}) = \exp^{-|a_i - b_{i-1} - 1|}$, where
  - a_i denotes the starting position of the foreign phrase translated into the i-th English phrase,
  - b_{i-1} denotes the end position of the foreign phrase translated into the (i-1)-th English phrase.

Lexical Reordering Features (cont'd)

- Right and left monotone model $P_R(\tilde{e}, \tilde{f})$ and $P_L(\tilde{e}, \tilde{f})$, where

$$P_R(\tilde{f}, \tilde{e}) = \frac{\text{count}_R}{\text{count}(\tilde{f}, \tilde{e})}$$

and count_R denotes the number of right connected phrases that are monotone.

Other features

- Number of words that constitute a translation
- Number of phrases that constitute a translation

Decoder

- Beam search + A* search
- Constraints for reordering:
  - Window size constraint, restricting number of words to be skipped in the source
  - ITG-constraint

Experimental Purpose

- To validate the use of the reportedly effective features
  - All features introduced previously are used.
- Evaluation of additional language resources
  - Comparable experiments with both s*upplied* and *unrestricted* data tracks are conducted.
  - Target language is English:
    - Japanese-to-English
    - Chinese-to-English
    - Korean-to-English
    - Arabic-to-English

Experimental Conditions

- Mixed casing and prefix-4 form for word alignment
- Mixed casing for language models
- Language models are trained by SRI toolkit

Monolingual Corpora for Unrestricted Data Track

English data sets	Corpus size (words)
IWSLT (supplied)	190,177
ATR	1,100,194
WEB	8,482,782
Gigaword	1,799,531,558

ATR: ATR spoken language database

WEB: WEB pages on traveling

Bilingual Corpora for Unrestricted Data Track

Data sets	Language pairs	Corpus size (English words)
IWSLT (supplied)	JE/CE/AE/KE	190,177
ATR	JE	1,334,852
LDC	CE	76,939,292
-	AE	-
-	KE	-

ATR: ATR spoken language database

LDC: LDC2004T08 and LDC2005T10

Other Setups

- Use NIST score for estimating feature function scaling factors
- ITG-constraints for J-to-E and K-to-E
- Window size constraints up to 7 for A-to-E and C-to-E
- On-the-fly estimation of language models
  1. Vocabulary set is limited to that observed in the supplied corpus and ATR database when counting n-grams.
  2. N-gram models for decoding are derived from the vocabulary set generated by using the extracted phrase pairs and the test set.

Evaluation of Additional Monolingual Corpora

-- Output Language Perplexity of N-grams for Decoding --

Test sets	6-gram				class-9-gram	
	ATR	IWSLT	WEB	GIGA	IWSLT	IWSLT+ATR
devset1	37.7	41.0	81.2	93.8	40.7	41.1
devset2	41.1	44.3	88.4	92.0	45.0	44.8

Test sets	prefix4-9-gram		suffix4-9-gram	
	IWSLT	IWSLT+ATR	IWSLT	IWSLT+ATR
devset1	41.5	34.0	40.0	32.6
devset2	44.5	36.4	42.9	35.2

- The perplexities of n-grams trained by additional resources are small enough.

Evaluation of Additional Bilingual Corpora

-- Input Language Perplexity of Supplied-data Trigram --

Test sets	Japanese		Chinese	
	IWSLT	ATR	IWSLT	LDC
devset1	16.9	29.5	56.6	462
devset2	17.6	32.9	56.1	449
testset	24.5	28.6	50.7	432

$ATR \approx IWSLT$

$LDC \neq IWSLT$

Results

Language pair	Translation input	Training data	BLEU score	NIST score
AE	transcription	supplied	0.4350	9.1821
		^{m}unrestricted	0.4764	9.3674
CE	transcription	supplied	0.3275	8.0768
		^{m}unrestricted	0.4112	8.8418
		^{mb}unrestricted	0.3943	8.6804
	ASR 1-best	supplied	0.2739	6.5185
		^{mb}unrestricted	0.2965	6.9416
JE	transcription	supplied	0.3669	7.9669
		^{m}unrestricted	0.3679	8.1207
		^{mb}unrestricted	0.3932	8.6442
	ASR 1-best	supplied	0.3881	8.3855
		^{mb}unrestricted	0.3762	8.3502
KE	transcription	supplied	0.3218	7.8489
		^{m}unrestricted	0.3497	8.0160

- Supplied < Unrestricted
- Additional monolingual resources are helpful.

Conclusions

- Competitive accuracy is obtained.
- The log-linear model effectively utilized n-grams trained by out-of-domain corpora, and improved the translation accuracy of the supplied data.
- Future works:
  - Feature extraction
  - Why is our system extremely inferior in terms of BLEU scores?