

Model Adaptation in Statistical Machine Translation for Synchronous Context-Free Grammars

Manaal Faruqui Baskaran Sankaran Anoop Sarkar



Statistical Machine Translation (SMT)

- The process of translating one human language to another human language by the computer using probabilistic models.

Let x be a source language sentence

y be a target language sentence

Our work is to maximize the probability of 'x' translating into 'y'

$$\hat{y} = \operatorname{argmax}_y \log P(y | x)$$

Context Free Grammar

- A grammar in theoretical computer science is a re-writing system
 - It consists of some terminal (think of **values**) and non-terminals (think of **variables**)
 - It has some production rules: LHS \longrightarrow RHS
 - It means we can replace/re-write the LHS with RHS
- A CFG is a grammar in which every production rule is of the form :-

Non-terminal \longrightarrow A combination of **Terminals** and **Non-terminals**

For example:-

$$\begin{aligned} X &\longrightarrow Y a Z \\ Y &\longrightarrow b \\ Z &\longrightarrow c \end{aligned}$$

where,

$$\begin{aligned} X, Y, Z &\in \{ \text{Non-terminals} \} \\ a, b, c &\in \{ \text{Terminals} \} \end{aligned}$$

- A synchronous CFG combines two CFG together .

Using SCFG to do Machine Translation

- Suppose we have the following rules ,

$$\begin{array}{ll} S \longrightarrow (X_1 , X_1) & \log P_1 = -8.0 \\ X \longrightarrow (\text{Je } X_1 , \text{I } X_1) & \log P_2 = -3.0 \\ X \longrightarrow (X_1 \text{ étudiante} , X_1 \text{ student}) & \log P_3 = -5.2 \\ X \longrightarrow (\text{suis} , \text{am a}) & \log P_4 = -2.6 \end{array}$$

We can derive a translation using these rules,

$$\begin{array}{ll} S \longrightarrow (X_1 , X_1) & \log P = -8.0 \\ \longrightarrow (X_1 \text{ étudiante} , X_1 \text{ student}) & \log P = -13.2 \\ \longrightarrow (\text{Je } X_1 \text{ étudiante} , \text{I } X_1 \text{ étudiante}) & \log P = -16.2 \\ \longrightarrow (\text{Je suis étudiante} , \text{I am a student}) & \log P = -18.8 \end{array}$$

Where do the rules come from?

- In training phase of our SCFG translation model, we extract the rules from a big **parallel corpus**.
- In a parallel corpus we have all the sentences in the source language corpus translated into the target language corpus. The source language corpus is translated by **humans** into the target language.



- The parallel corpus used in our experiments was Europarl-v3 French-English corpus which is freely available.

It had 1.27 M sentences of both French and English language.

- We derive translation rules from this corpus using phrase alignments for each sentence.

Model Adaptation

- Consider the situation where we have lots of parallel text to train our SMT system ; but all the text comes from one source, e.g. Parliamentary proceedings (the Europarl Corpus) .
- We actually would like to translate newswire text or blogs -- we need to adapt to the new domain
- We use a log-linear model for domain adaptation .

Approach

- Recall that we find the best translation as :-

$$\hat{y} = \operatorname{argmax}_y \log P(y | x)$$

- We compute the best translation using a log-linear model :-

$$\log P(y | x) = (\lambda_1 * feature_1 + \lambda_2 * feature_2 + \dots - \lambda_n * feature_n) - \log Z$$

Where,

$feature_i \in$ Various components of our translation model

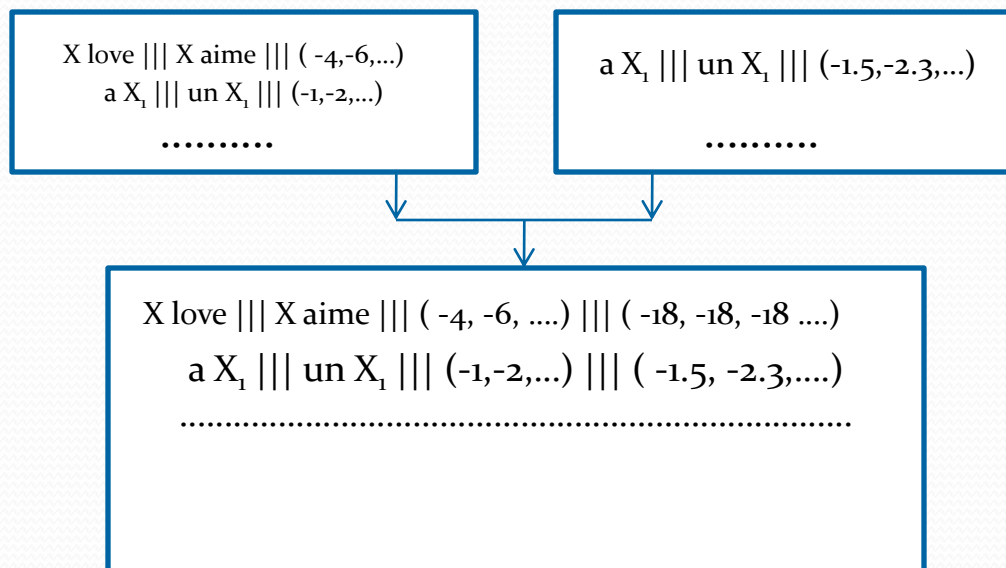
$\lambda_i \in$ Respective feature weights

- We add new components for our new domain and they get new feature weights λ

The new λ 's are set to values that allow the specific in-domain information to be combined with the general out of domain information

Merging Rule Tables

- We merge the rule tables in such a way that the new table contains all the rules present in both tables .



- So every rule in the merged table has the following form :-
source ||| target ||| probset₁ ||| probset₂
- This ensures that every candidate translation proposed by our system obtains a score from the log-linear model for domain adaptation

Filtering Rules

Rule Tables	Number of Rules
Rules from in-domain corpus	8.54 M
Rules from the merged data	321.76 M
Rules from the merged Rule-Tables	320.48M

Due to the large size of the rule-tables , I also worked on filtering rules to improve translation speed.

Rule filtering according to the sentence structure in the given data-set

- The filtering was done according to sentence structure in the given data-set.

X est belle . \longrightarrow X is beautiful .

-If the word “belle” is not present in the set of given sentences then the above rules can't be used .

je X étudiante . \longrightarrow I X student .

-Even if “je” and “étudiante” occur in the data-set, but if in no sentence “je” occurs before “étudiante” then the above rule can't be used.

Individual Rule-Filtering

- We did not allow a target phrase to have more than “n” number of translations .
- For our experiments, we fixed $n = 10$.
- The translations having the highest target side counts were chosen as the top 10 translations .
- This filtering reduced the existing rule table to $1/3$ of its original cardinality .

After Rule-Filtering

Rule-Tables	No. of Rules before filtering	No. of Rules after filtering
Rules from the in-domain corpus	8.54 M	0.45 M
Rules from the merged data	321.76 M	2.60 M
Rules from the merged Rule-Tables	320.48M	2.67 M

Evaluation Metrics

- Each evaluation metric compares the system output to a set of reference translations. There is **no** such thing as **one correct translation** -- there are **many**
- **Bilingual Evaluation Understudy (BLEU)** : BLEU is the geometric mean of the number of phrase matches of different lengths combined with a penalty for being too short compared to the input .
- **Multi-Reference Word Error Rate (mWER)** : The minimum number of substitutions, insertions and deletions required to transform the hypothesis into the reference translation .
- **Multi-Reference Position Independent Error Rate (mPER)** : The minimum number of substitutions, insertions and deletions required to transform the unordered set of words into the reference translation .

Results

Rules obtained from	BLEU	m-WER	m-PER
In-domain corpus	11.31	64.67	80.25
Merged data	14.67	59.22	78.57
Merged Tables	15.30	57.66	76.22

- BLEU (Higher the better)
- m-WER & m-PER (Lower the better)

Future Work

- The system has been developed to handle any number of phrase tables together .
- Experiments can be carried out with more than two translation tables and results can be analyzed and generalized for 'n' number of translation tables.