Grammar Induction for Machine Translation

Taro Watanabe
taro.watanabe @ nict.go.jp
Machine Translation

“I fire linguists”

A data-driven approach to MT (or, “crude force of computer”)

The United Inspection Department of Heishantou Port has shortened the procedures for leaving and entering the territory from originally 2 - 3 days to 1 day.

“I’m not crude”

We learn parameters from data assuming a “model”
1. The development of Shanghai's Pudong is in step with the establishment of its legal system.

2. Xinhua News Agency, Shanghai, February 10, by wire (reporters Jinhu Xie and Chijian Zhang)

3. In recent years Shanghai's Pudong has promulgated and implemented 71 regulatory documents relating to areas such as economics, trade, construction, planning, science and technology, culture and education, etc., ensuring the orderly advancement of Pudong's development.

4. Pudong's development and opening up is a century-spanning undertaking for vigorously promoting Shanghai and constructing a modern economic, trade, and financial center. Because of this, new situations and new questions that have not been encountered before are emerging in great numbers.

5. In response to this, Pudong is not simply adopting an approach of "work for a short time and then draw up laws and regulations only after waiting until experience has been accumulated." Instead, Pudong is taking advantage of the lessons from experience of developed countries and special regions such as Shenzhen by hiring appropriate domestic and foreign specialists and scholars, by actively and promptly formulating and issuing regulatory documents, and by ensuring that these economic activities are incorporated into the sphere of influence of the legal system as soon as they appear.

6. Precisely because as soon as it opened it was relatively standardized, China's first drug purchase service center for medical treatment institutions, which came into being at the beginning of last year in the Pudong new region, in operating up to now, has concluded transactions for drugs of over 100 million yuan and hasn't had one case of kickback.
A Process of Translation

I do not want to work

Je ne veux pas travailler
Generative Story

\[ \hat{e} = \arg \max_e Pr(e|f) \]

\[ = \arg \max_e Pr(f|e) Pr(e) \]

\[ = \arg \max_e \sum_d Pr(f, d|e) Pr(e) \]

• \textbf{d}: a derivation which “encodes” a process of translation from \textbf{e} to \textbf{f}

I will skip the Vauquois △
d = Word Alignment

I do not want to work

Je ne veux pas travailler

(Brown et al., 1993)
\[ d = \text{Phrase Pairs} \]

I do not want to work

Je ne veux pas travailler

“No linguistic intuition” (Koehn et al., 2003)
d = Hierarchical Phrases

(Chiang, 2007)
d = Tree Substitution

(Galley et al., 2004)
Research on MT @ NICT

- Model, Search, Optimization
- Today’s Focus: Inducing structural relations (d), i.e. grammars, from the pairs of (f, e)
  - Phrase pair induction
  - Label induction
Phrase Pair Induction

An Unsupervised Model for Joint Phrase Alignment and Extraction
“Traditional” (S)MT

\[
\begin{align*}
\text{bushi yu shalong juxing le huitan} & \iff \text{Bush held a talk with Sharon} \\
\vdots & \quad \vdots
\end{align*}
\]

\[
\begin{align*}
\text{bushi yu shalong juxing le huitan} \\
\downarrow \\
\text{Bush held a talk with Sharon}
\end{align*}
\]

\[
\begin{align*}
\text{bushi} & \quad \text{Bush} \\
\text{yu} & \quad \text{with} \\
\text{shalong} & \quad \text{Sharon} \\
\text{yu shalong} & \quad \text{with Sharon} \\
\text{juxing le huitan} & \quad \text{held a talk} \\
\vdots & \quad \vdots
\end{align*}
\]

\[
\begin{align*}
X & \to \langle \text{bushi } X_1, \text{Bush } X_1 \rangle \\
X & \to \langle X_1 \text{ yu shalong } X_2, X_1 X_2 \text{ with } S \rangle \\
X & \to \langle X_1 X_2 \text{ le huitan}, X_2 \text{ a talk } X_1 \rangle \\
\vdots & \quad \vdots
\end{align*}
\]

GIZA++(zh2en) \quad GIZA++(en2zg)
Simpler

A model to capture many-to-many alignment

(Zhang et al., 2008; DeNero et al., 2008; Blunsom et al., 2010)

bushi yu shalong juxing le huitan $\iff$ Bush held a talk with Sharon

bushi yu shalong juxing le huitan

Bush held a talk with Sharon

bushi $\parallel$ Bush
yu $\parallel$ with
shalong $\parallel$ Sharon
yu shalong with Sharon
juxing le huitan held a talk

$X \to \langle \text{bushi } X_1, \text{Bush } X_1 \rangle$

$X \to \langle X_1 \text{ yu shalong } X_2, X_1 X_2 \text{ with Sharon } \rangle$

$X \to \langle X_1 X_2 \text{ le huitan, } X_2 \text{ a talk } X_1 \rangle$
Single, Direct Model

bushi yu shalong juxing le huitan  \iff  Bush held a talk with Sharon

\[
\begin{align*}
\text{bushi} & \quad \mid \quad \text{Bush} \\
\text{yu} & \quad \mid \quad \text{with} \\
\text{shalong} & \quad \mid \quad \text{Sharon} \\
\text{yu shalong} & \quad \mid \quad \text{with Sharon} \\
\text{juxing le huitan} & \quad \mid \quad \text{held a talk}
\end{align*}
\]

\[
\begin{align*}
X & \rightarrow \langle \text{bushi} \ X_1, \text{Bush} \ X_1 \rangle \\
X & \rightarrow \langle X_1 \ yu \ shalong \ X_2, \ X_1 \ X_2 \ \text{with} \ Sharon \rangle \\
X & \rightarrow \langle X_1 \ X_2 \ \text{le huitan}, \ X_2 \ a \ talk \ X_1 \rangle
\end{align*}
\]
Inversion Transduction Grammar (ITG) (Wu, 1997) is a CFG over two languages:

- single non-terminal + regular/inverted production
- single pre-terminal + terminals: phrase pairs
Alignment by Biparsing

$p_x(reg)$

$p_x(term)$

$p_t(bushi/Bush)$

$p_x(reg)$

$p_x(term)$

$p_x(inv)$

$p_x(reg)$

$p_x(term)$

$p_x(term)$

$p_x(term)$

$p_t(yu/with)$

$p_t(shalong/Sharon)$

$p_t(juxing le/held)$

$p_t(huitan/a talk)$

Bush

held

a

talk

with

Sharon

bushi

yu

shalong

juxing

le

huitan
Sampling

Bilingual data with derivations

\{..., \langle f, e, \phi \rangle, ... \}

\[ \phi = \frac{X}{X} \frac{X}{X} \]

Choose data

“parsing” or compute inside probabilities

Choose new \( \Phi' \)

increment for new \( \Phi' \)

Compute \( P(x) \) for all possible derivations

“sampling” by outside computation

Update derivation

decrement for \( \Phi \)
# Parsing

- • = parsed source/target span
- synchronized by the # of words parsed
- continue until: ・・・・・,・・・・・
- no chart, no stack  (Saers et al., 2009)
Block Sampling

\[ Z = P(\Phi_1) + P(\Phi_2) + P(\Phi_3) + P(\Phi_4) + P(\Phi_5) + P(\Phi_6) \]

- Sample new “sentence-wise” derivations \( \Phi' \)
Model

\[ P_t \sim \text{PY}(d, s, P_{\text{base}}) \]
\[ P_x \sim \text{Dirichlet}(\alpha, 1/3) \]

- Bayesian approach:
  - Terminal probabilities for phrase pair \((f, e)\) come from Pitman-Yor Process
  - Branch probabilities \((x = \text{reg}, \text{inv}, \text{term})\) come from Dirichlet distribution

\[
P_t(f, e) = \frac{c(f, e)}{c(\_)} + \frac{s}{c(\_)} + \frac{P_{\text{base}}(f, e)}{c(\_)} + s
\]

\[
P_x(x) = \frac{c(x)}{c(\_)} + \frac{\alpha_x/3}{c(\_)} + \frac{\alpha_x}{c(\_)}
\]

(omit d for brevity)
Minimum Phrases

- Generate from a root until we reach terminals
- Sampled derivations contain only minimum phrases
**Fallback Modeling**

\[ P_t(f, e) = \frac{c(f, e)}{c(\_)} + \frac{s}{c(\_) + s} P_{dac}(f, e) \]

\[ P_{dac}(f, e) = \begin{cases} 
P_x(base)P_{base}(f, e) \\
P_x(str)P_t(f', e')P_t(f'', e'') \\
P_x(inv)P_t(f', e'')P_t(f'', e') 
\end{cases} \]

- Compute terminal (phrase) probabilities, first
- If not in the model, split and divide-and-conquer
Exhaustive ITG Phrases

++ bush yu shalong juxing le huitan/
Bush held a talk with Sharon

++ yu shalong juxing le huitan/
++ bush/Bush
held a talk with Sharon

++ yu shalong juxing le huitan/
with Sharon

++ shalong/Sharon
++ yu/with

++ juxing le/held
++ yu/with

++ huitan/a talk
++ juxing le/held
Experiments

• WMT10 news-commentary: de-en, es-en, fr-en
• NTCIR-8 patent translation: ja-en
• Moses for decoding:
  • Heuristics (GIZA++)
  • ITG (FLAT)
  • Fallback ITG (HIER)
• FLAT and HIER employ phrases in the model
Results

BLEU

Million Phrases

de-en  es-en  fr-en  ja-en

de-en  es-en  fr-en  ja-en

GIZA++  FLAT  HIER
Summary

- Bilingual version of Adaptor Grammars (Johnson et al., 2007)
- Competitive accuracy with GIZA++ baseline with significantly smaller phrase table size
- No more heuristics: a single model captures bilingual relation
- Open sourced: pialign
POS Induction in Dependency Trees for SMT
Akihiro Tamura, Taro Watanabe, Eiichiro Sumita,
Syntax for MT

- How syntax can help MT?
Forest-to-String MT

A very strong syntax based MT: CYK binarized forest-to-string (Zhang et al., 2011)

- dependency
- constituency
- CYK binarize (Zhang et al., 2011)
- "I pay the fare" (Huang et al., 2006)
- translation forest (Mi et al., 2008)

"I pay the fare"
“Monolingual” Labels

私の地図を使って教える

I teach it using a map

I used the pen and returned it
POS Induction

- Hidden states as POS labels

$Z_5 = \text{POS}1, \text{POS}2, \ldots, \text{POS}C$

(Gao et al., 2008)
Infinity

- No limit in the # of states

Hierarchical Dirichlet Process

\[ z_5 = \text{POS}_1, \text{POS}_2, \ldots, \text{POS}_{\infty} \]

(Gael et al., 2009)
∞ Over Tree

\[ Z_5 = \text{POS1, POS2, \ldots, POS}_\infty \]

(Finkel et al., 2007)

• Instead of HMM, we assume dependency tree is given
Monolingual Induction

私は地図を使って教える

私はペンを使って返した
Think Bilingually

I teach it using a map

I used the pen and returned it
Bilingual Induction

• Jointly emit both of the source and target terminals
Independent Model

\[ \gamma \rightarrow \beta \rightarrow \pi_k \rightarrow \phi_k \]

\[ \alpha_0 \rightarrow \beta \rightarrow \pi_k \rightarrow \phi_k \]

\[ H \rightarrow \phi_k \]

\[ H' \rightarrow \phi'_k \]

\[ \z_1 \rightarrow \z_2 \rightarrow \z_4 \rightarrow \z_7 \rightarrow \text{map} \rightarrow \text{NULL} \]

\[ \z_2 \rightarrow \z_6 \rightarrow \text{using} \rightarrow \text{NULL} \]

\[ \z_1 \rightarrow \z_3 \rightarrow \z_5 \rightarrow \text{teach} \rightarrow \text{NULL} \]

A diagram showing the relationships between various variables and concepts, including labels like "map", "NULL", "using", and "teach". The diagram illustrates dependencies and connections between these elements.
Variation in Emission

Joint Model

Independent Model
Refinement

$\mathbb{Z}_5 = \text{POS}_1, \text{POS}_2, \ldots, \text{POS}_\infty$

The tags, i.e. $V$, come from the original parse trees

$\mathbb{Z}_5 = V_1, V_2, \ldots, V_\infty$
Inference

forward filtering

\[ p(z_t | x_{\sigma(t)}, u_{\sigma(t)}) \propto p(x_t | z_t) \sum_{z_{d(t)}} p(z_{d(t)} | x_{\sigma(d(t))}, u_{\sigma(d(t))}) \]

\( Z_5 = \text{POS1, POS2, ..., POS}^{\infty} \)

how to handle this?

i.e. \( Z_5 = \text{POS7} \)

backward sampling

\[ p(z_t | z_{c(t)}, x_{1:T}, u_{1:T}) \propto p(z_t | x_{\sigma(t)}, u_{\sigma(t)}) \prod_{t' \in c(t)} p(z_{t'} | z_t, u_{t'}) \]
Limit Infinity: Beam Sampling

\[ u_t \sim \text{Uniform}(0, \pi_{zd(t)} z_t) \]

\[ Z_3 = \text{POS1}, \text{POS2}, \ldots, \text{POS}_{\infty} \]

\[ Z_5 = \text{POS1}, \text{POS2}, \ldots, \text{POS}_{\infty} \]

\[ \prod_{z_3 z_5} > u_5 \]

- \( u_t \): an auxiliary variable to limit the infinity
- Choose pairs which satisfies: \( \prod_{zt' z_t} > u_t \)
Experiments

• NTCIR-9 patent translation: ja-en
• Learn labels on subset (10K)
  1. Tagging/parsing/alignment (MeCab/CaboCha/GIZA++)
  2. Learn new labels
  3. Learn a joint tagger/parser (corbit)
  4. Parse all the data (3M)
• Forest-to-string translation (cicada)
Moses
base
mono

+surf
+POS
both

Induction
Refinement

Joint
Independent

BLEU

Results
Summary

• Induce labels which consider bilingual relation in the other language

• Improved performance:
  • Better than monolingual induction
  • Independent model alleviates the sparsity problems
Conclusion
Grammar Induction

- Currently, limited to phrase-pairs under ITG
- Future work: more complex arbitrary synchronous-CFG with arbitrary labels

Very good work by (Xiao et al, 2012; Xiao and Xiong, 2013)
Dependency Induction

• Currently, limited to POS given dependency trees

• Future work: jointly induce tree structures + POS