Semi-supervised Learning for Neural Machine Translation

Yong Cheng

joint work with Wei Xu, Zhongjun He, Wei He, Hua Wu, Maosong Sun, Yang Liu
Machine Translation

Automated translation using computer software
Machine Translation

- Rule-based Machine Translation 1970s
- Example-based Machine Translation 1984
- Statical Machine Translation (SMT) 1993
- Neural Machine Translation (NMT) 2014

**Trends:** learning to translate from DATA
Parallel corpora are usually limited in

- quantity
- quality
- coverage

Monolingual Corpora

Parallel Corpora
Monolingual Corpora Used in SMT and NMT

- N-gram language model in SMT
  Koehn et al., [2007]

- Monolingual corpora as decipherment
  Ravi and Knight [2011]

- Integrate a neural language model into NMT.
  Gulcchere et al. [2015]

- Additional pseudo parallel corpus.
  Sennrich et al. [2016]
Supervised Training

Parallel Corpus

\[ \mathcal{D} = \{ \langle x^{(n)}, y^{(n)} \rangle \}_{n=1}^{N} \]

Objective

\[ L(\theta) = \sum_{n=1}^{N} \log P(y^{(n)} | x^{(n)}; \theta) \]
Unsupervised Training

Monolingual Corpus

\[ \mathcal{T} = \{ y^{(t)} \}_{t=1}^{T} \]
Our Approach — Autoencoders
Our Approach — Autoencoders

\[ P(y \mid x; \bar{\theta}) \]

bushi yu shalong juxing le huitan

X
Our Approach — Autoencoders

Bush held a talk with Sharon

P(y | x; \theta)

bushi yu shalong juxing le huitan
Our Approach — Autoencoders

Bush held a talk with sharon

\[ P(x | y; \tilde{\theta}) \]

\[ P(y | x; \tilde{\theta}) \]

bushi yu shalong juxing le huitan
Our Approach — Autoencoders

bushi yu shalong juxing le huitan

$x'$

$P(x | y; \tilde{\theta})$

Bush held a talk with sharon

$y$

$P(y | x; \tilde{\theta})$

bushi yu shalong juxing le huitan

$x$
Our Approach — Autoencoders

source autoencoder

bushi yu shalong juxing le huitan

x’

P(x | y; \tilde{\theta})

Bush held a talk with sharon

y

P(y | x; \tilde{\theta})

bushi yu shalong juxing le huitan

x

y’

target autoencoder

Bush held a talk with sharon

P(x | y; \tilde{\theta})

bushi yu shalong juxing le huitan

y

Bush held a talk with sharon

x’
Unsupervised Training (Autoencoders)

Monolingual Corpus \( \mathcal{T} = \{y^{(t)}\}_{t=1}^T \)

\[
\begin{align*}
\text{Bush held a talk with sharon} & \quad \Rightarrow P(y \mid x; \tilde{\theta}) \\
\text{bushi yu shalong juxing le huitian} & \quad \Rightarrow P(x \mid y; \tilde{\theta}) \\
\text{Bush held a talk with sharon} & \quad \Rightarrow P(x \mid y; \tilde{\theta})
\end{align*}
\]

\[
\begin{align*}
\mathbf{y}' & = P(y' \mid y; \overrightarrow{\theta}, \overleftarrow{\theta}) \\
x & = \sum_x P(y', x \mid y; \overrightarrow{\theta}, \overleftarrow{\theta}) \\
y & = \sum_x P(x \mid y; \overrightarrow{\theta}) P(y' \mid x; \overleftarrow{\theta})
\end{align*}
\]
Semi-supervised Training

Training Objective

\[ J(\theta, \tilde{\theta}) = \sum_{n=1}^{N} \log P(y^{(n)} | x^{(n)}; \theta) + \sum_{n=1}^{N} \log P(x^{(n)} | y^{(n)}; \tilde{\theta}) \]

\[ + \lambda_1 \sum_{t=1}^{T} \log P(y' | y^{(t)}; \theta, \tilde{\theta}) + \lambda_2 \sum_{s=1}^{S} \log P(x' | x^{(s)}; \theta, \tilde{\theta}), \]

source-to-target likelihood  target-to-source likelihood  target autoencoder  source autoencoder
## Translation Results

### Compared with Moses (SMT) and RNNSearch (NMT)

| System     | Training Data | Direction | NIST06 | NIST02 |
|------------|---------------|-----------|--------|--------|
|            | CE C E        | C → E     | 32.48  | 32.69  |
| **Moses**  | ✓ × ×         | E → C     | 14.27  | 18.28  |
| **RNNSearch** | ✓ × ×     | C → E     | 30.74  | 35.16  |
|            |               | E → C     | 15.71  | 20.76  |
## Translation Results

### Compared with Moses (SMT) and RNNSearch (NMT)

| System       | Training Data | Direction | NIST06 | NIST02 |
|--------------|---------------|-----------|--------|--------|
|              | CE  | C  | E   | C → E | E → C |        |        |
| **Moses**    | √   | ×  | ×   | 32.48 | 14.27 | 32.69  | 18.28  |
|              | √   | ×  | ✓   | 34.59 | 20.69 | 35.21  | 25.85  |
|              | √   | ✓  | ×   | 30.74 | 15.71 | 35.16  | 20.76  |
| **RNNSearch**| √   | ×  | ×   |       |       |        |        |
## Translation Results

**Compared with Moses (SMT) and RNNSearch (NMT)**

| System   | Training Data | Direction | NIST06 | NIST02 |
|----------|---------------|-----------|--------|--------|
| **MOSES** |               |           |        |        |
|          | CE            | C         | E      |        |
| ✓        | ✗             | ✗         | ✗      |        |
|          | C → E         | E → C     |        |        |
|          | 32.48         | 14.27     |        |        |
|          | 32.69         | 18.28     |        |        |
|          | ✓             | ✗         | ✗      |        |
|          | C → E         | E → C     |        |        |
|          | 34.59         | 20.69     |        |        |
|          | 35.21         | 25.85     |        |        |
|          | ✓             | ✓         | ✗      |        |
|          | C → E         | E → C     |        |        |
|          | 30.74         | 15.71     |        |        |
|          | 35.16         | 20.76     |        |        |
| **RNNSEARCH** |             |           |        |        |
|          | ✓             | ✗         | ✗      |        |
|          | C → E         | E → C     |        |        |
|          | 35.61+++      | 17.59++   |        |        |
|          | 38.78+++      | 23.99++   |        |        |
|          | ✓             | ✗         | ✓      |        |
|          | C → E         | E → C     |        |        |
# Translation Results

## Compared with Moses (SMT) and RNNSearch (NMT)

| System    | Training Data | Direction | NIST06 | NIST02 |
|-----------|---------------|-----------|--------|--------|
|           | CE | C | E | C → E | E → C | C → E | E → C | C → E | E → C |
| **Moses** | ✓  | ✗ | ✗ | 32.48  | 14.27  | 32.69  | 18.28  | 35.21  | 25.85  |
|           | ✓  | ✗ | ✓  | 34.59  | 20.69  |        |        |        |        |
|           | ✓  | ✓  | ✗ | 30.74  | 15.71  | 35.16  | 20.76  |        |        |
| **RNNSearch** | ✓  | ✗ | ✓  | 35.61**** | 17.59++ | 38.78**** | 23.99 ++ | 38.20**** | 29.52**** |
|           | ✓  | ✓  | ✗ | 35.01++ | 21.12+++ | 38.20**** | 29.52**** |        |        |
### Translation Results

**Compared with Moses (SMT) and RNNSearch (NMT)**

| System        | Training Data | Direction | NIST06 | NIST02 | NIST03 | NIST04 | NIST05 |
|---------------|---------------|-----------|--------|--------|--------|--------|--------|
|               | CE C E       | C → E     |        |        |        |        |        |
| **Moses**     | ✓ × ×        | C → E     | 32.48  | 32.69  | 32.39  | 33.62  | 30.23  |
|               | ✓ × √        | C → E     | 34.59  | 35.21  | 35.71  | 35.56  | 33.74  |
|               | ✓ √ ×        | E → C     | 20.69  | 25.85  | 19.76  | 18.77  | 19.74  |
| **RNNSearch** | ✓ × ×        | C → E     | 30.74  | 35.16  | 33.75  | 34.63  | 31.74  |
|               | ✓ × √        | C → E     | 35.61**+ | 38.78**+ | 38.32**+ | 38.49**+ | 36.45**+ |
|               | ✓ √ ×        | E → C     | 17.59++ | 23.99++ | 18.95++ | 18.85++ | 17.91++ |
|               | ✓ √ ×        | C → E     | 35.01++ | 38.20***+ | 37.99***+ | 38.16***+ | 36.07***+ |
|               | ✓ √ ×        | E → C     | 21.12***+ | 29.52***+ | 20.49***+ | 21.59***+ | 19.97++ |
## Translation Results

Compared with *Sennrich et al.* [2015a]

| Method            | Training Data | Direction | NIST06 | NIST02 | NIST03 | NIST04 | NIST05 |
|-------------------|---------------|-----------|--------|--------|--------|--------|--------|
|                   | CE | C | E       |       |        |        |        |
| *Sennrich et al.* [2015a] | √  | √ | ×       | √     | C → E  | 34.10  | 36.95  | 36.80  | 37.99  | 35.33  |
|                   |   |   |         |       | E → C  | 19.85  | 28.83  | 20.61  | 20.54  | 19.17  |
| *this work*       | √  | √ | ×       | √     | C → E  | 35.61**| 38.78**| 38.32**| 38.49* | 36.45**|
|                   |   |   |         |       | E → C  | 17.59  | 23.99  | 18.95  | 18.85  | 17.91  |
|                   | √  | √ | ×       |       | C → E  | 35.01**| 38.20**| 37.99**| 38.16  | 36.07**|
|                   |   |   |         |       | E → C  | 21.12**| 29.52**| 20.49  | 21.59**| 19.97**|
| Monolingual       | hongsen shuo, ruguo you na jia famu gongsi dangan yishenshifa, name tamen jiang zihui qiancheng. |
|-------------------|-------------------------------------------------------------------------------------------------|
| Reference         | hongsen said, if any *logging companies* dare to defy the law, then they will *destroy their own future*. |
| Translation       | hun sen said, if any of *those companies* dare defy the law, then they will *have their own fate*. [iteration 0] |
|                   | hun sen said if any *tree felling company* dared to break the law, then they would *kill themselves*. [iteration 40K] |
|                   | hun sen said if any *logging companies* dare to defy the law, they would *destroy the future themselves*. [iteration 240K] |

\[
y^* = \arg\max_x \left\{ P(y|x; \hat{\theta}) P(x'|y; \hat{\theta}) \right\}
\]
Conclusion

- Monolingual corpora is an important resource for neural machine translation.

- We have proposed a semi-supervised approach to training bidirectional neural machine translation models for exploiting monolingual corpora.

- As our method is sensitive to the OOVs present in monolingual corpora, we plan to integrate Jean et al. (2015)’s technique on using very large vocabulary into our approach.
Thank You!
Effect of Sample Size

ZH-EN

EN-ZH
Effect of OOV ratio

ZH-EN

EN-ZH