Syntax-based Rewriting for Simultaneous Machine Translation

He He, Alvin Grissom II, John Morgan, Jordan Boyd-Graber and Hal Daumé III
Simultaneous Machine Translation

- Translate segments instead of sentences
- Trade-off between
  - Translation quality
  - Translation delay
- Intelligent segmentation
  - Grissom II et al., 2014 EMNLP
  - Oda et al., 2014 ACL
  - Fujita et al., 2014 Interspeech
- Challenges
  - Divergent word order
  - SOV vs. SVO

Skype Translator demo from WPC 2014
我々は
We-TOP
我々は
We-TOP

We
例 I

我々は 政府の
We-TOP government-GEN

We

… of the government
Example I

我々は 政府の 構造 や 組織を
We-TOP government-GEN structure and composition-ACC
Example I

我々は 政府の 構造 や 組織を 変更 すべき だ
We should change the government's structure and composition.

We
We should change the structure and composition of the government.
我々は政府の構造や組織を変更すべきだ
We should change the structure and composition of the government
Example I

我々は 政府の 構造 や 組織を 変更 すべき だ
We should change the structure and composition of the government

・Divergent word order → large segment → delay
・Use word orders similar to the source language
我々は
We-TOP
Example II

我々は
We-TOP

HOLD ON
We
Example II

我々は 政府の
We-TOP government-GEN

HOLD ON
We

… of the government → the government’s
Example II

我々は 政府の
We-TOP government-GEN

HOLD ON the government’s

We

… of the government → the government’s
Example II

我々は 政府の 構造 や 組織を
We-TOP government-GEN structure and composition-ACC

HOLD ON the government’s

We

… of the government → the government’s
Example II

我々は 政府の 構造 や 組織を
We-TOP government-GEN structure and composition-ACC

HOLD ON the government’s structure and composition
We

… of the government → the government’s
Example II

我々は 政府の 構造 や 組織を 変更 すべき だ
We TOP government GEN structure and composition ACC change should COP

HOLD ON the government's structure and composition

We

… of the government → the government’s
change … → … change → be changed
Example II

我々は 政府の 構造 や 組織を 変更 すべき だ
We should change the structure and composition of the government

HOLD ON  the government’s  structure and composition  should be changed
We should change the structure and composition of the government

… of the government → the government’s
change … → … change → be changed
Example II

We should change the structure and composition of the government.

… of the government → the government’s
change … → … change → be changed
Example II

我々は 政府の 構造 や 組織を 変更 すべき だ
We should change the structure and composition of the government.

… of the government → the government’s
change … → … change → be changed
Example II

We should change the structure and composition of the government

... of the government → the government’s
change ... → ... change → be changed

• Similar word order → smaller segment → reduce delay
• Rewrite target / reference sentence during training
Roadmap

Better speed-accuracy trade-off

More monotone / simultaneous translation

Target word order similar to the source

Hand-crafted rules for rewriting training data
The Problem of Delay Reduction

• Rewrite the training target translation to
  \[ \min \text{ translation delay} \]
  \[ \text{s.t. obey the grammar} \]
  \[ \text{keep the original meaning} \]

• Delay
  • \# words to observe between two (segment) translations

• Data-driven approach
  • Lack of parallel translation-interpretation data
  • Flaw of human interpretation

• Rule-based approach
  • Linguistically inspired
  • Syntactical structure (parse trees)

1. Perfect text input
2. Segmented into sentences
3. Comes in word by word
4. Japanese to English
We should change the structure and composition of the government
the structure and composition of the government should be changed by us

The talk was denied by the boycott group spokesman
Transformation Rules: Voice

\[ NP_1 \text{ verb} \ NP_2 \ \xrightarrow{\text{passivization}} \ NP_2 \text{ verb} \ NP_1 \]

the structure and composition of the government should be changed by us

the boycott group spokesman denied the talk

• Useful when \( \text{len}(NP_2) > \text{len}(NP_1) \)
• Accommodate verbs coming at the end
They announced that the president will restructure the division
Transformation Rules: Verb+Clause

- **Quotative**: say, announce, conclude ...
- **(Anti-)Factive**: know, realize, observe, doubt, deny ...
- **Belief**: think, believe, suspect ...

the president will restructure the division, they announced
the structure and composition of the government
Transformation Rules: Noun Phrases

NP₁ \textit{of} \ NP₂ \rightarrow \ NP₂ \ 's / ' \ NP₁

the government 's the structure and composition
Transformation Rules: Noun Phrases

\[
\text{It is ADJP SBAR } \quad \rightarrow \quad \text{SBAR is ADJP}
\]

\[
\text{It is important to remain watchful}
\]

\[
\text{that we remain watchful is important}
\]
We should march because winter is coming
Transformation Rules: Conjunction Clauses

• Clausal: *because (of), in order to*
• Contrastive: *despite, even though, although*
• Conditional: *even (if), as a result (of)*
• Misc: *according to*

winter is coming ; because of this , We should march

It was cold outside, even though this is the case, they went to swim
Rewriting Evaluation

We love the new world

We new world the love

delay = 2.5 words / seg

The new world is loved by us

我々は新しい世界を愛して

我々は新しい世界を愛して

延迟减少！

delay = 2.5 words / seg

delay = 1.7 words / seg
Roadmap

Better speed-accuracy trade-off

More monotone / simultaneous translation

Target word order similar to the source

Hand-crafted rules for rewriting training data
この措置は、台湾の企業や投資家に、台湾ドルと他の外国通貨の間の為替オプション取引の機会を提供することになるという。

The move will offer options trading for Taiwan companies and investors in exchanges between the Taiwan dollar and other foreign currencies.

Options trading for Taiwan companies and investors will be offered by the move in exchanges between the Taiwan dollar and other foreign currencies.

さらに同議⾧長は、商品価格はかつてほどにはインフレ指標として有効ではなくなっただ、と述べた。

Greenspan also said that commodity prices were not as good an indicator of inflation as they once were.

Commodity prices were not as good an indicator of inflation as they once were, Greenspan also said.
Lee said the central bank believes exchange rates should be left to the market, although he expressed concern about sharp fluctuations.
Lee said the central bank believes exchange rates should be left to the market, although he expressed concern about sharp fluctuations.

The central bank believes exchange rates should be left to the market, although he expressed concern about sharp fluctuations, Lee said.
Lee said the central bank believes exchange rates should be left to the market, although he expressed concern about sharp fluctuations.

The central bank believes exchange rates should be left to the market, although he expressed concern about sharp fluctuations, Lee said.

He expressed concern about sharp fluctuations, although this is the case, the central bank believes exchange rates should be left to the market, Lee said.
Lee said the central bank believes exchange rates should be left to the market, although he expressed concern about sharp fluctuations.

The central bank believes exchange rates should be left to the market, although he expressed concern about sharp fluctuations, Lee said.

He expressed concern about sharp fluctuations, although this is the case, the central bank believes exchange rates should be left to the market, Lee said.

Concern was expressed by him about sharp fluctuations, although this is the case, exchange rates should be left to the market, the central bank believes, Lee said.
Error Analysis

• Word usage error

ASEAN currently comprises Indonesia, Malaysia, Vietnam, Thailand, Singapore, the Philippines and Brunei.

*Indonesia, Malaysia, Vietnam, Thailand, Singapore, the Philippines and Brunei currently is comprised by ASEAN.*

• Parsing error

The London Stock Exchange closes at 1230 GMT today.

*GMT today is closed at 1230 by the London Stock Exchange.*
Statistics of Rewritten Sentences

- Dataset
  - Reuters Ja-En corpus of news articles
- Most widely applicable
  - Voice changing
- Most effective
  - Possessive case, *It is* + clause
- 95.5% words have the same semantic role labels in the corresponding rewritten sentences
Roadmap

Better speed-accuracy trade-off

More monotone / simultaneous translation

Target word order similar to the source

Hand-crafted rules for rewriting training data
Machine Translation Settings

- Baseline MT system
  - Phrase-based model trained by Moses
  - Trigram language model, GIZA++ for word alignment, MIRA for tuning
- Segmentation (Fujita et al., 2013)

Pr(translation of two consecutive phrases being monotone)
Threshold controls how aggressive the segmentation is

low RP: 少年の花 \[\xrightarrow{\text{the flower}}\text{of the boy}\]

high RP: 少年の花 \[\xrightarrow{\text{the boy's flower}}\]
Machine Translation Models

• OriginalOnly
  • Only use original reference translations
• RewriteOnly
  • Only use rewritten reference translations
• Combined helps alleviate errors introduced by rewriting
  • Interpolate language models of OriginalOnly and RewriteOnly
  • Combine translation models of OriginalOnly and RewriteOnly by fill-up (Bisazza et al., 2011)
    • Tune on rewritten reference translations
• RP threshold
  • Controls length of segments
  • 0.0 (phrase), 0.2, 0.4, 0.6, 0.8, 1.0 (batch)
Experiments: Machine Translation

• Better trade-off against the rewritten reference
Better trade-off against the original reference
Especially with shorter segments
Bonus: Help Keep the Verbs

- Verb dropping problem in SOV-to-SVO translation, e.g., Ja-En, De-En
  - Long range distortion

- Putting verb at the end helps
  - Similar to source-side pre-reordering
Ref
• He also said that the real dangers for the euro lay in the potential for divergences in the domestic policy needs among the various participating nations of the single currency.

OriginalOnly
• He also for the euro, is a real danger to launch a single currency in many different countries and domestic policies to the need for the possibility of a difference.

RewriteOnly
• He also for the euro is a real danger to launch a single currency in many different countries and domestic policies to the need for the possibility of a difference, he said.
Conclusion

Better segmentation methods: +cache

Data selection based on syntax

Learning from human interpretation

...