Improving Japanese-to-English Neural Machine Translation by Voice Prediction

Hayahide Yamagishi, Shin Kanouchi, Takayuki Sato and Mamoru Komachi
Tokyo Metropolitan University

Abstract

Our previous study [Yamagishi+ Workshop on Asian Translation 2016] presented a voice controlling method to generate sentences for NMT. We demonstrated that the BLEU score improved when the voice of generated sentence was controlled relative to that of the reference. However, it is impractical to use the reference information because we cannot discern the voice of the correct translation in advance. Thus, this study presents a voice prediction method for generated sentences for NMT.

Proposed Method

Voice Controlling Method [Yamagishi+ WAT2016]

[Train Target] The bone density is explained .

① Parsing the English sentences.

② Is there “be-verb + past participle” structure in the main clause?
Yes → <Passive> label
No → <Active> label

③ Adding each label into the end of Japanese sentences as a word.

④ Training an NMT model with labeled Japanese sentences.

[Train Source] 骨 密度 に ついて 解説 し た 。 <Passive>

⑤ Adding the specified voice labels into the end of test sentences.

[Test Source] 骨 密度 に ついて 解説 し た 。 <Active>

⑥ Test and check

[Test Target] We explain the bone density.

Statistics of voice in ASPEC

Sentences (Train data)	Active	Passive	Total
Japanese side	992,902	110,434	1,103,336
English side	604,158	499,178	1,103,336

The NMT model should classify which voice is appropriate.

Voice Prediction

Developing a voice classifier using the following features.
Some of these features are selected and concatenated as a vector used to train a logistic regression model.
(The features marked * are used on the only ASPEC experiments.)

1. Subject in a source sentence.
2. Predicate in a source sentence.
3. Predicate in the previous source sentence.*
4. Voice of the source sentence.
5. Objects in the previous target sentence.*
6. Voice of the output sentences from previous three sentences.*
7. The majority of target voice of each predicate in a source sentence.

ASPEC: Obtaining 67.7% accuracy using feature 2, 6 and 7.
NTCIR: Obtaining 66.0% accuracy using feature 1, 2 and 4.

Experiments

Baseline: Training and testing with the non-labeling data.

Labeling patterns
ALL_ACTIVE: All sentences to active voice.
ALL_PASSIVE: All sentences to passive voice.
REFERENCE: Each sentence to the same voice as that of the reference sentences.
PREDICT: Each sentence to the predicted voice.

Test data: 100 active sentences (not include the intransitive verb) and 100 passive sentences extracted from the official test data.
BLEU score was calculated with this 200 sentences.
Tow comparisons:
Voice Accuracy: The voice of output and that of the reference.
Control Accuracy: The voice of the output and the added label.

Settings

Corpus: (The English sentences were written by the non-natives.)
ASPEC (Abstracts of scientific papers)
- 1,103,336 sentences (329,025 documents)

NTCIR PatentMT Parallel Corpus
- 1,169,201 sentences

Model: NMT with attention [Bahdanau+ ICLR2015]
- Vocabulary: 30,000 (word-level)
- Dimensions of the embedding and hidden unit: 512

Results

Experiment (ASPEC)	Active	Passive	Error	Voice Accuracy	Control Accuracy	BLEU
Test data	100	100	-	-	-	-
Baseline	31	163	6	60.5%	-	20.60
ALL_ACTIVE	147	44	9	57.5%	73.5%	20.22
ALL_PASSIVE	6	189	5	51.0%	94.5%	20.18
REFERENCE	82	113	5	89.5%		22.47
PREDICT	74	118	8	64.0%	89.0%	21.06

Experiment (NTCIR)	Active	Passive	Error	Voice Accuracy	Control Accuracy	BLEU
Test data	100	100	-	-	-	-
Baseline	69	127	3	66.0%	-	31.80
ALL_ACTIVE	127	69	4	71.0%	63.5%	31.91
ALL_PASSIVE	17	186	3	55.0%	93.0%	32.32
REFERENCE	80	116	4	89.5%		33.90
PREDICT	73	122	5	69.5%	83.0%	33.16

Discussion

1. It is easy to generate the passive sentences.
2. The voice information of the previous sentences are good feature.
3. This method is useful for the single sentences.
4. The voice of coordinate clauses tend to become the same as that of main clause.
5. The voice of subordinate clauses tend to be active.

Experiment	Examples of single sentence	Examples of complex sentence
Source	リサイクルに関する最近の話題を紹介した。	また、ドットの形状及び結晶性は温度に依存することもわかった。
Reference	Recent topics on recycling are introduced.	It was also proven that the shape and crystallinity of the dots were dependent on temperatures.
To be Active	This paper introduces recent topics on recycling.	The morphology and the crystallinity of the dots depended on the temperature.
To be Passive	Recent topics on recycling are presented.	It was also found that the shape and crystallinity of the dots depend on the temperatures.