Building a Non-Trivial Paraphrase Corpus Using Multiple Machine Translation Systems

首都大学東京 TOKYO METROPOLITAN UNIVERSITY Yui Suzuki Tomoyuki Kajiwara Mamoru Komachi

TMUP Corpus https://github.com/tmu-nlp/paraphrase-corpus

Background

- Generating paraphrases manually is costly.
- Primitive paraphrase acquisition tends to result in a biased corpus.
- Resources in languages other than English are not often available.

Contributions

- Proposing an automatic sentential paraphrase candidate generation method.
- Collected **non-trivial** Paraphrases & Non-Paraphrases.
- Releasing Japanese sentential paraphrase corpus for evaluation.

Examples

Jaccard score

Non-trivial Paraphrase 0.07

It is rarely used.

めったに 使わ れる こと は あり ませ ん 。

まれ に 使用 さ れ ます 。

- low word overlap rate
- semantically equivalent

Trivial Paraphrase 0.60

He was a member of the Republican Party.

彼 は 共和党 の メンバー でし た 。

彼 は 共和党 の 一員 だっ た 。

Non-trivial Non-Paraphrases 0.90

There is also a strong Roman Catholic presence.

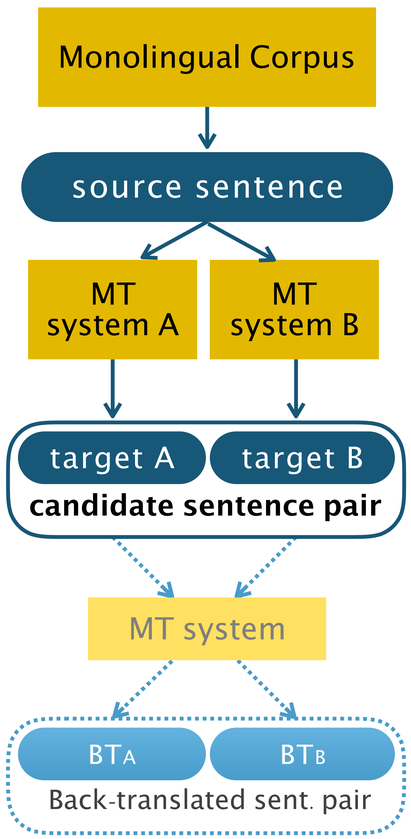
強力 な ローマ カトリック の 存在 感 も あり ます 。

強力 な ローマ カトリック の 存在 も あり ます 。

- high word overlap rate
- semantically inequivalent

Automatic Candidate Generation

Paraphrase Generation using machine translation systems



Quality Estimation #1

To acquire better translation outputs, calculate language model probabilities of source sentences, and translate only sentences with high probabilities.

Quality Estimation #2

To make sure to collect adequate outputs, back-translate the output sentences, and calculate overall sentence BLEU of each pair.

$$QE_i = \mathrm{SBLEU}(\mathrm{source}_i, \mathrm{BT_A}(e_i)) \times \mathrm{SBLEU}(\mathrm{source}_i, \mathrm{BT_B}(e_i))$$

Balance Control #1

To sample pairs with balance, calculate word overlap rate (Jaccard score) of each pair, and segment them into 11 ranges based on the score.

Balance Control #2

Non-Trivial Non-Paraphrase Extraction from a monolingual corpus

To actively acquire non-trivial non-paraphrases, randomly collect sentence pairs with high word overlap rate from a monolingual corpus written in the target language.

Manual Annotation

- **Equivalency**: Both *translated* and *extracted* candidate pairs
- **Fluency**: Only *translated* candidate pairs

Japanese Paraphrase Corpus

Source Language	English
Target Language	Japanese
Monolingual Corpora	Wikipedia (EN) (JA)
Language Model	5-gram LM from English Gigaword Fifth Edition
MT Systems	Google SMT and Google NMT
# of sampled pairs for annotation	2,000 (200 each from 10 ranges except the exact match)
# of Annotators	2
Inter annotator agreement	κ = 0.60

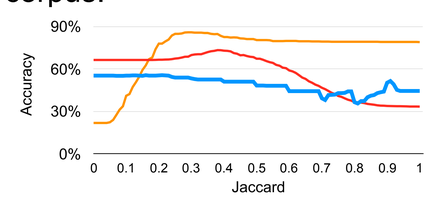
Table. 1: Details of Japanese paraphrase corpus creation.

Jaccard	translated sentence	sentence length (av.)			Annotation result		
		source	PBMT	NMT	Para-phrase	Non-Para-phrase	others
[0.0, 0.1)	228	19.42	20.65	19.75	2	1	197
[0.1, 0.2)	2,117	21.56	24.81	22.01	11	14	175
[0.2, 0.3)	14,080	21.56	26.50	23.37	20	9	171
[0.3, 0.4)	51,316	23.48	29.69	26.29	24	15	161
[0.4, 0.5)	100,674	24.40	31.35	28.08	27	16	157
[0.5, 0.6)	134,101	23.16	29.90	27.26	34	16	150
[0.6, 0.7)	100,745	21.04	27.32	25.30	38	13	149
[0.7, 0.8)	55,610	18.83	24.57	23.04	53	12 (40)	135
[0.8, 0.9)	26,884	16.23	21.31	20.24	81	3 (80)	116
[0.9, 1.0)	8,071	13.79	18.40	17.55	73	3 (70)	124
[1.0, 1.0]	6,174	10.10	13.07	12.96	-	-	-
Total	500,000	19.42	24.32	22.35	363	102 (190)	1,535

Table. 2: Statistics on Tokyo Metropolitan University Paraphrase (TMUP) Corpus. Inside the parentheses in the annotation result is the number of added non-trivial non-paraphrases.

Corpus Analysis

- Binary classification using word overlap rate on:
  - **TMUP** (our corpus)
  - **MSRP** (Microsoft Research Paraphrase Corpus)
  - **TPC** (Twitter Paraphrase Corpus)
- **TMUP** is suitable for evaluation: The lower the accuracy is, the more difficult to solve the corpus.



Accuracy (max)

TPC	86.0%
TMUP	**55.7%**
MSRP	73.4%

Figure. 1: Accuracy of Paraphrase identification on three corpora.

The 55th annual meeting of the Association for Computational Linguistics (ACL) Student Research Workshop (SRW) July 31, 2017. Vancouver, Canada.