Building a Monolingual Parallel Corpus for Text Simplification Using Sentence Similarity Based on Alignment between Word Embeddings

https://github.com/tmu-nlp/sscorpus

Tomoyuki Kajiwara and Mamoru Komachi Tokyo Metropolitan University kajiwara-tomoyuki@ed.tmu.ac.jp

1. Japan is an island country in East Asia.
2. Japan is a stratovolcanic archipelago of 6,852 islands.
3. The four largest are Honshu, Hokkaido, Kyusyu and Shikoku, which make up about ninety-seven percent of Japan's land area.

Complex Corpus

	1	2	3	・・・
1	0.63	0.15	0.33	
2	0.50	0.30	0.25	
・・・	Sentence Similarity Matrix			

1. Japan is a country in Asia.
2. It has many islands.
3. Four of them are big, and the biggest is the 7th largest island in the world.
4. This island is called Honshu.

Simple Corpus

Woody Bay Station was ***purchased*** by the Lynton and Barnstaple Railway Company in 1995 and, after much effort, a short section of railway reopened to passengers in 2004. Woody Bay Station was ***bought*** by the Lynton and Barnstaple Railway Company in 1995 and, after much effort, a short section of railway reopened to passengers in 2004. (similarity: 0.9)
Mir ***has been*** a significant influence on late 20th-century art, in particular the American abstract expressionist artists ~~such as Motherwell, Calder, Gorky, Pollock, Matta and Rothko, while his lyrical abstractions and color field paintings were precursors of that style by artists such as Frankenthaler, Olitski and others.~~ Mir ***was*** a significant influence on late 20th-century art, in particular the American abstract expressionist artists. (similarity: 0.7)

Parallel Corpus

①Computing sentence similarity
②Extracting sentence pairs by a threshold
③Training PBSMT model using parallel corpus
④Translating sentence from complex to simple

Mozart's Clarinet Concerto and Clarinet Quintet are both in A major, and generally Mozart was more likely to use clarinets in A major than in any other key besides E-flat major.

→ PBSMT model →

Mozart's Clarinet Concerto and Clarinet Quintet are both in A major, and Mozart used clarinets in A major often.

Sentence Alignment by Word Alignment

Song and Roth (NAACL-2015) Unsupervised Sparse Vector Densification for Short Text Similarity

1. Average Alignment

We averaging the similarities between all pairs of words taken from the two sentences

$$S_{ave}(x,y) = \frac{1}{|x||y|}\sum_{i=1}^{|x|}\sum_{j=1}^{|y|}\phi(x_i,y_j)$$

2. Maximum Alignment

Average Alignment is a noisy method → We utilize only accurate alignments from the most similar word y_j for each word x_i

$$S_{\max}(x,y) = \frac{1}{|x|}\sum_{i=1}^{|x|}\max_j \phi(x_i,y_j)$$

3. Hungarian Alignment

We represent two sentences in a weighted complete bipartite graph with words as nodes and word similarities as edges. The one-to-one word alignment that maximizes sentence similarity is obtained by finding the maximum matching of the graph.

$$S_{hun}(x,y) = \frac{1}{\min(|x|,|y|)}\sum_{i=1}^{|x|}\phi(x_i,h(x_i))$$

Related Works: Building a monolingual parallel corpus for text simplification from English Wikipedia and Simple English Wikipedia

- Zhu et al. (2010) paired normal and simple sentences represented as TF-IDF vectots using cosine similarity
- Coster and Kauchak (2011) extended Zhu et al.'s work by considering the order of the sentences
- Hwang et al. (2015) computed sentence similarity taking account of word-level similarity using a dictionary

Our Work: Building a text simplification corpus using sentence similarity based on alignment between word embeddings

We also compute sentence similarity considering word-level similarity but using word embeddings

Experimental Setup (Hwang et al.'s dataset)

Hwang et al. (2015) built a benchmark dataset for text simplification extracted from English Wikipedia and Simple English Wikipedia.

67,853 complex and simple sentence pairs with 4 labels.

- Bidirectional Entailment: 277 sents.
- Unidirectional Entailment: 281 sents.
- Related: 117 sents.
- Unrelated: 67,178 sents.

Building an English Text Simplification Corpus

① We paired 126,725 articles from English Wikipedia and Simple English Wikipedia by an exact match of titles

② We paired 492,993 normal and simple sentences by Maximum Alignment
- Threshold for word alignment: > 0.49
- Threshold for sentence alignment: > 0.53

Binary Classification between Parallel and Nonparallel Sentences	Bidirectional Entailment vs. Others		Unidirectional Entailment vs. Others	
	MaxF1	AUC	MaxF1	AUC
Zhu et al. (2010)	0.550	0.509	0.431	0.391
Coster and Kauchak (2011)	0.564	0.495	0.415	0.387
Hwang et al. (2015)	0.712	0.694	0.607	0.529
Additive Embeddings	0.691	0.695	0.518	0.487
1. Average Alignment	0.419	0.312	0.391	0.297
2. Maximum Alignment	0.717	0.730	0.638	0.618
3. Hungarian Alignment	0.524	0.414	0.354	0.275

※ Additive Embeddings: Comparative method without word alignment
Cosine similarity of sentence embeddings composed by adding word embeddings

PBSMT-based English Text Simplification

	#sents.	Avg. #words complex	Avg. #words simple	BLEU Bi-Ent.	BLEU Uni-Ent.
Baseline (None)				42.1	22.3
Zhu et al. (2010)	107,516	21.2	17.4	42.0	22.1
Coster and Kauchak (2011)	136,862	23.6	21.1	44.3	23.8
Hwang et al. (2015)	284,238	26.0	19.8	43.9	23.1
Ours	492,493	25.3	17.9	47.5	26.3

※ Avg. #words per sent. of the entire English Wikipedia: 25.1
※ Avg. #words per sent. of the entire Simple English Wikipedia: 16.9

100,000 sents.	BLEU Bi-Ent.	BLEU Uni-Ent.
Zhu et al.	41.8	22.1
Coster and Kauchak	43.8	23.4
Hwang et al.	42.9	22.7
Ours	43.2	23.6

200,000 sents.	BLEU Bi-Ent.	BLEU Uni-Ent.
Hwang et al.	43.1	22.7
Ours	45.7	24.8

Tomoyuki Kajiwara and Mamoru Komachi. Building a Monolingual Parallel Corpus for Text Simplification Using Sentence Similarity Based on Alignment between Word Embeddings. COLING-2016, P12, December 14th, 2016.