Study on Automatic Scoring of Descriptive Type Tests using Text Similarity Calculations

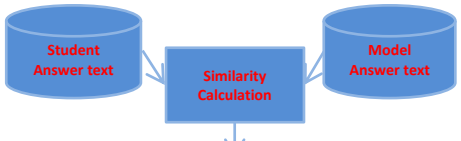
Izuru Nogaito, Keiji Yasuda, Hiroaki Kimura (KDDI R&D Labs, Japan)

Summary

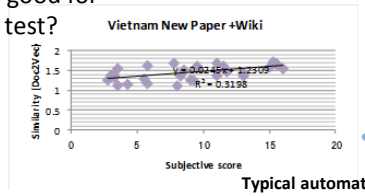
- We evaluate a similarity for an automatic scoring of a descriptive type test.
- We compared 3 similarity measures（BLEU, RIBES, Doc2Vec）.
- BLEU, Ribes are advantageous for summarization question.
  Doc2Vec is better for graph reading questions.
- BLEU/Ribes and Doc2Vec work as a complemental similarity.

Background and Purpose

- One of typical method of auto scoring is using similarity with model answers.



- What similarity is good for a descriptive type test?



Typical automatic scoring by text similarity

Similarity Measures

- **Similarity in surface expression (n-gram, word order)**

+BLEU is an evaluation metric for a machine translations. It uses *n*-gram matching between a reference sentence and a machine translation output.

+RIBES is also an evaluation metric for a machine translation. It inspects the word order for common words based on the rank correlation coefficient.

- **Similarity in distributed expression (Doc2Vec)**

Recently, by using deep learning technology, a word or sentence can be converted into a distributed expression that is a vector of several hundred dimensions. Similarities are defined as vector distance.

Experimental Settings and Test Item

+ Subject students
  More than 20 students, 10 to 16 years old.
+ Human evaluator for subjective evaluation
  4 teachers evaluated students answer, averages are used.
+ Model sample answer is each 4 answer
+Question type
 "Summarization" :
  asked to summarize a given text between 300 to 800 words long.
 "Graph reading" :
  asked to describe a fact that can be read from the given graphs.
+Method of evaluation
  Correlation of subjective evaluation and similarities

Item ID	Topic of question	Question type	Ave. length of student answers (words)	words @ sentence	# of students
ID01	Book	Graph reading	112.2	62.5	21
ID02	Fisherman	Summarization	49.7	33.4	21
ID03	Food	Graph reading	96.4	49.0	24
ID04	Fishery	Graph reading	87.8	53.5	22
ID05	Supermarket	Summarization	101.4	59.7	22
ID06	University	Summarization	110.7	71.6	20
ID07	Japanese	Summarization	77.7	46.8	32
ID08	Mail	Summarization	58.9	44.6	42
ID09	Vietnam	Graph reading	57.5	31.2	29
ID10	Beef	Graph reading	90.2	44.2	24
ID01-10	Average		84.3	49.6	25.7

Study on Automatic Scoring of Descriptive Type Tests using Text Similarity Calculations
(Continue)

Training Corpus for Vec2Doc

	# of words	Lexicon size
Japanese wiki abstract (WIKI)	29,944,313	1,398,558
Mainichi-News-Paper (1991-2014) (NP)	504,844,192	5,578,327
WIKI + NP	534,788,505	6,376,935

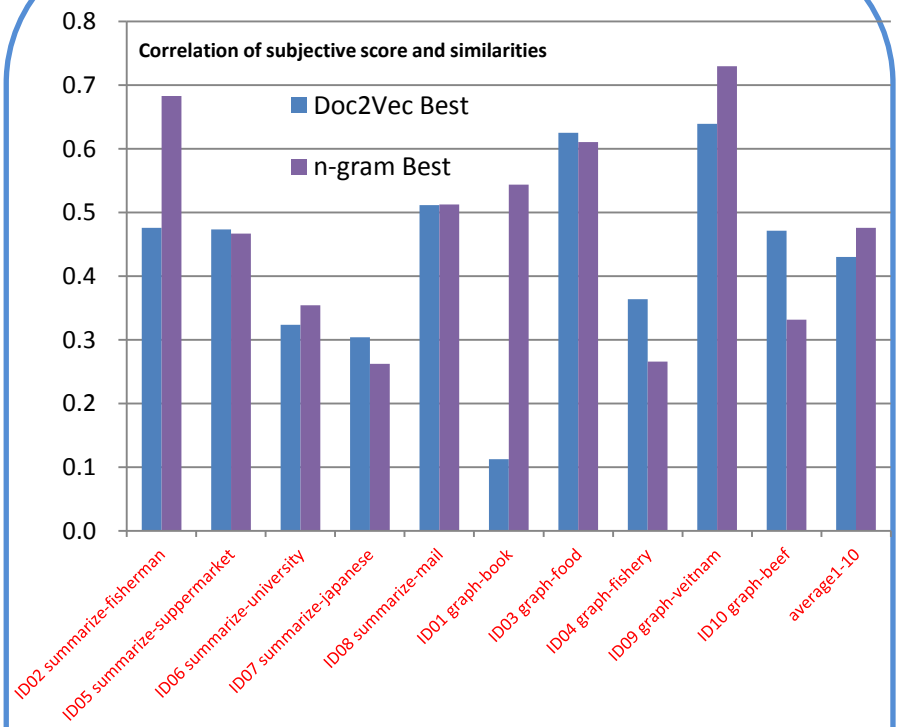
Experimental results and Discussion

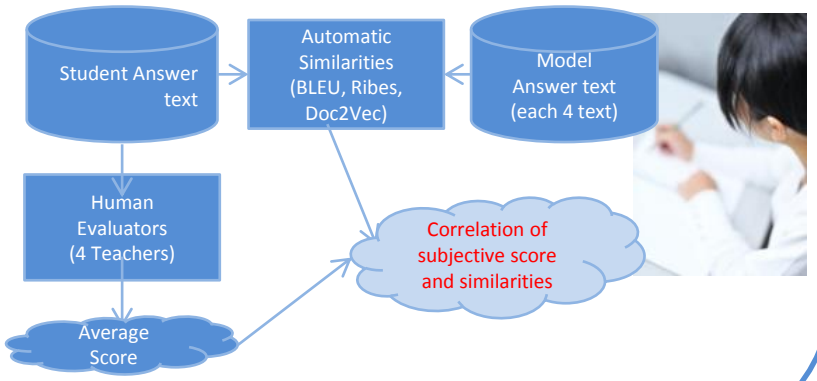
+Similarity of N-Gram type (Bleu, Ribes) is advantageous for summarization question
Students use the expression in question sentence.
ID02, ID05, ID06, ID07, ID08
+Doc2Vec is better for graph reading questions.
Students are free to choose their own words.
ID03, ID04 and ID10.
+ ID01, we understand that the corpus used does not share many similar words with the model answer sentences.
+Doc2Vec similarity sometimes also works as a complemental similarity

Conclusion/ Future Work

+Average correlation of subject score & similarities is 0.4-0.5.
+We will conduct research to use several similarities in a complementary way.
+We will also compare several methods, including the method using cohesion and coherence as a second method.

Similarities





Experimental diagram