ARPA: Armenian Paraphrase Detection Corpus and Models

Arthur Malajyan
malajyanarthur@ispras.ru

Speaker: Karen Avetisyan
karavet@ispras.ru

Tsolak Ghukasyan
tsggukasyan@ispras.ru

Ivannikov Laboratory for System Programming at Russian-Armenian University, Yerevan

Motivation

- Create paraphrase corpora for the Armenian language
- Establish baseline results for paraphrase detection

Main Contributions

- Train paraphrase corpus
- Test paraphrase corpus
- Baseline results

Plan

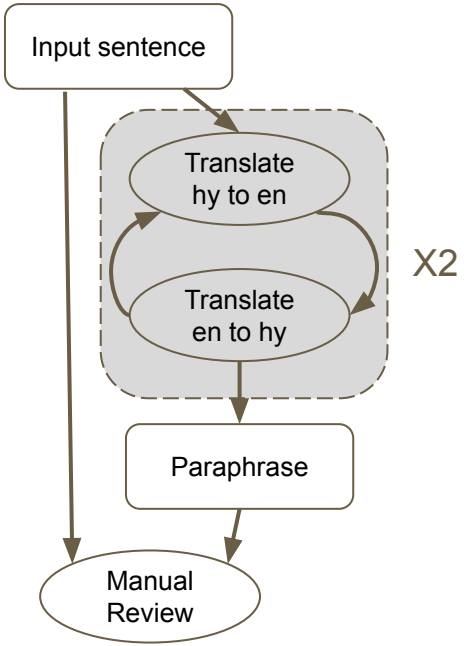
- Datasets
  - Automatic paraphrase generation
  - Train set annotation
  - Test set annotation
- Evaluation of paraphrase detection BERT-based models

Paraphrase generation

- Sentence Selection and Filtration
  - Initial sentences were crawled from news websites (Hetq and Panarmenian)
  - Filtered meaningless sentences
  - Filtered sentences containing fewer than 6 and more than 22 tokens

Paraphrase generation

- Sentence Pair Generation Using Double Back Translation



Sentence set annotation labels

- Manually filtered syntactically/semantically incorrectly translated sentences
- Annotation guideline follows the 2012 SemEval's Semantic Textual Similarity degrees

  Paraphrase
  - 5 - Completely equivalent
  - 4 - Mostly equivalent, but some unimportant details differ

  Not Paraphrase
  - 3 - Roughly equivalent, but some important information differs/missing
  - 2 - Not equivalent, but share some details
  - 1 - Not equivalent, but are on the same topic
  - 0 - On different topics

Sentence set annotation examples

	Input Sentence:	Paraphrase:
Paraphrase	Կոռուպցիան չարիք են համարում բոլորը՝ չինովնիկից մինչև բանվոր: Corruption is considered evil by everyone, from deputies to worker.	Կոռուպցիան բոլորի համար չարիք է համարվում ՝ պաշտոնյաներից մինչև աշխատակիցներ: Corruption is considered bad for everyone, from officials to employees.
Near Paraphrase	Այսօր 100%-ով վերականգնվել է էլեկտրամատակարարումը - հայտարարել է նախարար Խորխե Ռոդրիգեսը: The power supply has been fully restored today, said Minister Jorge Rodriguez.	Այսօր էլեկտրաէներգիան վերականգնվել է 100% -ի չափով: Today, electricity is 100% restored.
Not paraphrase	Այլ կերպ ասած՝ ինչից շատ ունենք, դա էլ ցույց ենք տալիս: In other words, we show that which we have most.	Այլ կերպ ասած, մենք ցույց ենք տալիս ավելին, քան ունենք: In other words, we show more than what we have.

Test and Train Annotation Results

- Sentence pair were divided into 2 subsets
  - 1573 for training
  - 1382 for testing
- 1339 out of 1573 train examples were considered as “paraphrase” (85%)
- 1021 out of 1382 test examples were considered as “paraphrase” (74%)
- Inter-annotator agreement varies from 0.55 to 0.65 (3 annotators)

Paraphrase diversity level

- Diversity verified by computing the average number of word-level edits between the source sentence and its paraphrase

Dataset	***Paraphrase diversity***	
	Train set	***Test set***
MRPC	6.79	7.01
ParaPhraser.ru	5.02	5.51
ARPA	**8.70**	**8.66**

Label distributions

Dataset	***Paraphrase***		***Non-Paraphrase***		***Total***
	Examples	***Average Jaccard Similarity***	***Examples***	***Average Jaccard Similarity***	
Test					
MRPC	1147	0.438	578	0.322	1725
ParaPhraser.ru	1137	0.317	762	0.169	1899
ARPA	1021	0.327	661	0.172	1682
Train					
MRPC	2753	0.444	1323	0.325	4076
ParaPhraser.ru	4255	0.306	2947	0.119	7202
ARPA	1339	0.320	2894	0.056	4233

Evaluated Models

1. Multilingual BERT, fine-tuned on the following datasets
    a. MRPC, translated to Armenian
    b. ParaPhraser.ru, translated to Armenian
    c. ARPA dataset, proposed in this work
    d. All of the training sets above combined
2. DeepPavlov's RUBERT, tested on ARPA google-translated to Russian
3. BERT trained on MRPC and tested on ARPA google-translated to English

Evaluation on ARPA test

Models	*Scores (95% confidence interval)*			
	F1	*Accuracy*	*Recall*	*Precision*
1.a. tr-MRPC	0.801 ± 0.014	0.699 ± 0.028	0.993 ± 0.005	0.672 ± 0.021
1.b. tr-Parahraser	0.838 ± 0.002	0.771 ± 0.002	0.977 ± 0.005	0.734 ± 0.002
1.c. ARPA	0.837 ± 0.003	0.775 ± 0.003	0.952 ± 0.009	0.747 ± 0.002
1.d Combined	0.840 ± 0.002	0.776 ± 0.002	0.971 ± 0.006	0.741 ± 0.001
2. RUBERT	0.837	0.764	0.998	0.721
3. BERT	0.779	0.656	1.0	0.638

Performance on near-paraphrase examples

- Calculated accuracy on examples which were difficult to differentiate from paraphrase

Model	Accuracy on near-paraphrases
1.a. tr-MRPC	3.00%
1.b. tr-Parahraser	4.17%
1.c. ARPA	9.05%
1.d Combined	4.55%

State-of-the-art models comparison

Dataset	BERT Model	F1	Accuracy	Recall	Precision
MRPC	BERT-Base	88.9	83.5	99.38	80.39
ParaPhraser.ru	RUBERT	87.9	84.9	91.60	84.48
	BERT-Base Multilingual	83.4	79.3	86.84	80.22
ARPA	BERT-Base Multilingual	83.7	77.5	95.20	74.70

Thank You!