Classification and Clustering of Arguments with Contextualized Word Embeddings

Nils Reimers, Benjamin Schiller, Tilman Beck,
Johannes Daxenberger, Christian Stab, Iryna Gurevych
Ubiquitous Knowledge Processing Lab (UKP-TUDA)
Department of Computer Science, Technische Universität Darmstadt
www.ukp.tu-darmstadt.de

A UKP ASPECT Corpus: Amazon Mechanical Turk Guidelines and Inter-annotator Agreement

The annotations required for the UKP ASPECT Corpus were acquired via crowdsourcing on the Amazon Mechanical Turk platform. Workers participating in the study had to be located in the US, with more than 100 HITs approved and an overall acceptance rate of 90% or higher. We paid them at the US federal minimum wage of $7.25/hour. Workers also had to qualify for the study by passing a qualification test consisting of twelve test questions with argument pairs. Figure 1 shows the instructions given to workers.

B AFS Corpus: Detailed Results

Table 1 shows the full results of the (un)supervised methods for the argument similarity calculation on the AFS dataset (all topics).

References

Amita Misra, Brian Ecker, and Marilyn A. Walker. 2016. Measuring the similarity of sentential arguments in dialogue. In *Proceedings of the SIGDIAL 2016 Conference, The 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue, 13-15 September 2016, Los Angeles, CA, USA*, pages 276–287.

	Gun Control		Gay Marriage		Death Penalty		Avg.	
	r	ρ	r	ρ	r	ρ	r	ρ
Human Performance	.6900	-	.6000	-	.7400	-	.6767	-
Unsupervised Methods								
Tf-Idf	.6266	.5528	.4107	.3778	.3657	.3589	.4677	.4298
InferSent - fastText	.3376	.3283	.1012	.1055	.3168	.2931	.2519	.2423
InferSent - GloVe	.3757	.3707	.1413	.1435	.2953	.2847	.2708	.2663
GloVe Embeddings	.4344	.4485	.2519	.2741	.2857	.2973	.3240	.3400
ELMo Embeddings	.3747	.3654	.1753	.1709	.2982	.2663	.2827	.2675
BERT Embeddings	.4575	.4460	.1960	.1999	.4082	.4072	.3539	.3507
Supervised Methods: Within-Topic Evaluation								
SVR (Misra et al., 2016)	.7300	-	.5400	-	.6300	-	.6333	-
BERT-base	.8323	.8076	.6255	.6122	.7847	.7768	.7475	.7318
BERT-large	.7982	.7592	.6240	.6137	.7545	.7149	.7256	.6959
Supervised Methods: Cross-Topic Evaluation								
BERT-base	.6892	.6689	.4307	.4236	.6339	.6245	.5849	.5723
BERT-large	.6895	.6749	.5071	.4866	.6641	.6486	.6202	.6034

Table 1: Pearson correlation r and Spearman's rank correlation ρ on the AFS dataset. Within-Topic Evaluation: 10-fold cross-validation. Cross-Topic Evaluation: System trained on two topics, evaluated on the third.

Read each of the following sentence pairs and indicate whether they argue about the same aspect with respect to the given topic (given as "Topic Name" on top of the HIT). There are **four options**, of which one needs to be assigned to each pair of sentences (arguments). Please read the following for more details.

- **Different Topic/Can't decide**: Either one or both of the sentences belong to a topic different than the given one, or you can't understand one or both sentences. If you choose this option, you need to very briefly explain, why you chose it (e.g. "The second sentence is not grammatical", "The first sentence is from a different topic" etc.). For example,

  Argument A: "*I do believe in the death penalty, tit for tat*".

  Argument B: "*Marriage is already a civil right everyone has, so like anyone you have it too*".

- **No Similarity**: The two arguments belong to the same topic, but they don't show any similarity, i.e. they speak about completely different aspects of the topic. For example,

  Argument A: "*If murder is wrong then so is the death penalty*".

  Argument B: "*The death penalty is an inappropriate way to work against criminal activity*".

- **Some Similarity**: The two arguments belong to the same topic, showing semantic similarity on a few aspects, but the central message is rather different, or one argument is way less specific than the other. For example,

  Argument A: "*The death penalty should be applied only in very extreme cases, such as when someone commands genocide*".

  Argument B: "*An eye for an eye: He who kills someone else should face capital punishment by the law*".

- **High Similarity**: The two arguments belong to the same topic, and they speak about the same aspect, e.g. using different words. For example, Argument A: "*An ideal judiciary system would not sentence innocent people*".

  Argument B: "*The notion that guiltless people may be sentenced is indeed a judicial system problem*".

Your rating should not be affected by whether the sentences attack (e.g. "*Animal testing is cruel and inhumane*" for the topic "*Animal testing*") or support (e.g. "*Animals do not have rights, therefore animal testing is fair*" for the topic "*Animal testing*") the topic, but only by the aspect they are using to support or attack the topic.

Figure 1: Amazon Mechanical Turk HIT Guidelines used in the annotation study for the Argument Aspect Similarity Corpus.