Human Baseline Evaluation - Results

Patrick Huber, Jan Niehues, Alex Waibel
Institute for Anthropomatics and Robotics, Karlsruhe Institute of Technology (KIT)
uhejt@student.kit.edu, jan.niehues@kit.edu, alex.waibel@kit.edu

In Extension of the 2018 LREC Paper:

Automated Evaluation of Out-of-Context Errors

The Task

The human baseline evaluation is conducted to assess the performance of human subjects on the task of semantic error detection. Therefore, seven subjects have been presented with ten mutually unrelated, modified sentences from the 2016 TEDTalk dataset (Cettolo et al., 2012). The task given to the participants is to find the word(s) that do not/least fit into the presented text passage. The evaluation is conducted fully anonymous and is only used to determine the human-baseline for the 2018 LREC conference paper. An example for a sentence is shown in figure 1 below.

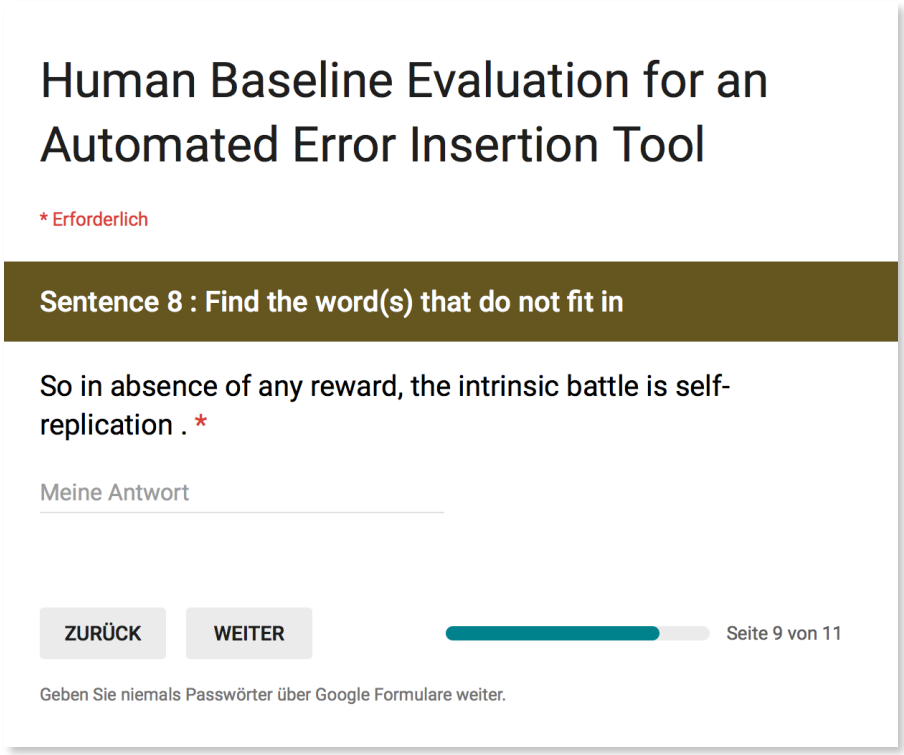

Figure 1: Sentence Example

Sentences

The sentences presented to the participants are shown in the following table alongside with the replaced word used as the ground truth.

#	Sentence	Modified Word
Sentence 1	Today, in the last few years, there's been an explosion in research on happiness.	Today
Sentence 2:	These clouds are bombing along, but from all the way down here, they appear to be moving gracefully, slowly, like most expectations.	expectations
Sentence 3	So we'll release the findings this September for the first time, and then next year, we'll poll again, and we'll take the additive step this time of ranking the 1,000 largest U.S. companies from number one to making 1,000 and everything in between.	making
Sentence 4	And watch it now slip, and see what it does with its conservation.	conversation
Sentence 5	The answer is that what you do, and the details are not terribly important here, is to make work more elaborate.	work
Sentence 6	You know what your word is, right?	word
Sentence 7	This is Iran government data, a completely independent site, "Where Does My Money Go?“, the way they called it: a betrayal to the Saudi country and the Saudi things, and they even started a hashtag called #OsloTraitor on Twitter.	Iran, things
Sentence 8:	So in absence of any reward, the intrinsic battle is self-replication.	battle
Sentence 9	I felt, when I had my options, that I was holding on to something true, regardless of agendas or politics.	options
Sentence 10	And this is one of these scientists that we captured in what we call gigapixel technology.	scientists

Results per Participant

The following table shows the individual results per participant on the task.

#	Correctly Classified	Wrongly Classified	Precision	Recall	F-Score
Subject 1	2	1	0.67	0.18	0.29
Subject 2	6	4	0.67	0.54	0.57
Subject 3	2	8	0.2	0.18	0.19
Subject 4	2	8	0.2	0.18	0.19
Subject 5	2	9	0.18	0.18	0.18
Subject 6	6	14	0.3	0.55	0.39
Subject 7	2	8	0.2	0.18	0.19

Results per Sentence

The final results per sentence are displayed in the following table.

#	Correctly Classified	Wrongly Classified
Sentence 1	5	2
Sentence 2	2	8
Sentence 3	4	6
Sentence 4	4	3
Sentence 5	0	17
Sentence 6	2	10
Sentence 7	3	9
Sentence 8	0	11
Sentence 9	2	7
Sentence 10	2	8

Reported Results

The following table shows the consolidated results published in the paper.

Words Replaced	Number of Words	Number of Sentences	Average F-Score
11	232	10	0.28

References

Cettolo, M., Girardi, C., and Federico, M. (2012). Wit^3: Web inventory of transcribed and translated talks. pages 261–268, May.