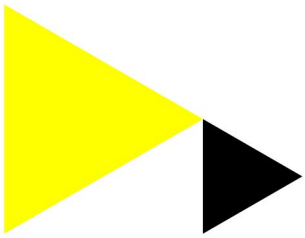
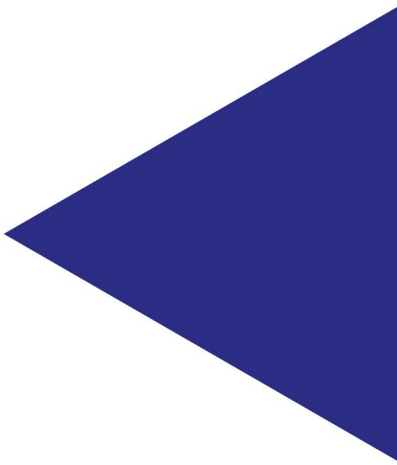
Validating Ontologies for Question Generation

Author(s)
Teitsma, Marten; Sandberg, Jacobijn; Wielinga, Bob; Schreiber, Guus

Publication date
2014

Document Version
Final published version





Validating Ontologies for Question Generation

Marten Teitsma [a] Jacobijn Sandberg [b] Bob Wielinga [c]
Guus Schreiber [c]

[a] *Amsterdam University of Applied Sciences, The Netherlands*
[b] *University of Amsterdam, The Netherlands* [c] *VU University Amsterdam, The Netherlands*

Abstract

In this paper we present an experiment which has been performed to validate a pragmatic-based, expert-based and basic-level ontology. These ontologies were created for use in an application which generates questions for ordinary people with the purpose to determine a crisis situation. All three ontologies have specific characteristics related to their method of creation. This experiment shows that using the basic-level ontology results in the fastest and least ambiguous determination of a crisis situation.

1 Introduction

Making use of humans to gather information is the central subject in the new emerging field of Human-Centered Sensing (HCS) [2]. The application we propose here is typified as a participatory sensor because humans are producing information and not just facilitating the gathering of data as in opportunistic sensing e.g., a mobile device recording background noise. By answering questions the human observers can help making clear what the situation is. Research on crisis situations shows a variety of public involvement during a crisis. Not only experts in crisis management convey information about the crisis at hand but also ordinary people, i.e. people with no specific knowledge of the situation they describe. It becomes more and more accepted to regard members of the public as true 'first responders' [4].

To determine a crisis situation we use the Situation Awareness Question Generator (SAQG) which automatically generates questions from an ontology. During this experiment questions were generated by asking after the specification of a concept, i.e. which of the subordinate concepts is a more specific description of an object in the real world. These questions are presented to ordinary people who, by answering the questions, help to determine the situation. In this paper we show an experiment with participants who determine what kind of object is on fire.

The SAQG is installed on a mobile device and generates questions from an ontology which is received from a server. These questions are presented to the user of the application. The answer given by the user is computed by SAQG and gives rise to another question until a situation is determined by the application. This situation is then communicated to the server. Previous experiments showed that questions generated from an ontology created by knowledge engineers (an expert-based ontology) did not lead to thrustworthy answers while questions generated from a ontology based on pragmatic considerations (a pragmatic-based ontology) were suitable to determine a situation [6, 5].

We discern three sources for the creation of ontologies: *a*) pragmatic classifications found in a particular domain, *b*) existing expert-based ontologies and *c*) natural categorization with basic-level concepts. We developed methods to create these ontologies. To find the ontology most suitable for generating questions we measured the three ontologies using four criteria: *a*) the ontology must have a structure which is useful for the task at hand, i.e. question answering on a mobile device, *b*) the construction of the ontology is efficient, *c*) the ontology must be complete, i.e. all concepts that are relevant should be contained in the ontology and *d*) the ontology should be compliant with human thought.

We used several metrics to compare these ontologies which showed that expert-based ontologies are most easy to construct but lack required cognitive ergonomic characteristics. Basic-level ontologies have structure and concepts which are better in terms of cognitive ergonomics but are most expensive to construct [7]. In this paper we present an experiment in which a simulation of a crisis situations is created and where participants help to determine that situation. With this experiment we identify the most suitable ontology for such a task.

Characteristics of the ontologies are presented in Section 2. The experiment we conducted and the results are presented in Section 3. The results are discussed in Section 4 and conclusions are drawn in Section 5.

2 The ontologies

The ontologies used by SAQG are composed of a representational part, a generic part and a domain specific part [8]. The representational and generic part of the ontology are a revised version of the Situation Theory Ontology [3]. The domain specific part captures the relevant knowledge of a particular domain. For all three ontologies we have defined a backbone consisting of the concepts *Streetobject*, *Ship*, *Roadvehicle*, *Railvehicle*, *Nature*, *Industry*, *Building*, *Aircraft* which represent the subjects we are interested in.

The ontologies we use in the experiments were created with three different methods [7]. The pragmatic based ontology (pbo) was developed from a classification used by the fire department at an emergency call center. Relations are heterogeneous and terms are functionally not the same in this classification. To make it suitable for automatic reasoning some knowledge engineering had to be done. The expert-based ontology (ebo) was constructed from existing ontologies created by domain experts. We extracted concepts related to the backbone from AATNed [1] and Cornetto [9] and merged these concepts into an ontology. Because the result was a very large ontology we filtered out less frequently used concepts. The basic-level ontology (blo) is an ontology based on empirical data elicited from ordinary people. In several experiments we asked for concepts related to the backbone and visual properties of objects denoted by these concepts. We then used the properties of the concepts to automatically create a hierarchy of concepts. This algorithm made use of the psychological phenomenon of basic-level concepts. The process to create the blo was rather laborious.

To make the ebo more efficient and the blo suitable for use with SAQG these ontologies were reengineered. The ebo was significantly improved by applying some small changes. The most important of these improvements was the reduction of path length for concepts which only had one subordinate concept. The superordinate concept was then replaced by the subordinate concept. The way we use an ontology in our application, i.e. asking a question and suggesting answers, does not generate information when only one answer is possible. We joined some synonyms and some concepts were replaced as subordinate to another concept when this seemed appropriate. The blo was not suitable due to the method to create this ontology: most superordinate concepts did not have a meaningful label. To create labels for the superordinate concepts we made use of guidelines formulated by van Heijst [8]. It is clear that the most perfect label consists of one word. Unfortunately it was hard to find appropriate words for some of the superordinate concepts and we had to be content with compound labels for some concepts. Only the pragmatic-based ontology could be used without modification by the application. The ontologies created were given new names by adding the prefix 'new': new expert-based ontology (nebo) and new basic-level ontology (nblo).

In the experiment as described in the next section we present participants with situations in three domains: *Buildings*, *Road vehicles* and *Water vehicles*. In Table 1 some characteristics of these subtrees in each ontology are shown. The subtrees of the nebo are largest and also have the highest average path length. The average number of subclasses for the pbo is highest. With respect to entropy the subtrees of the nebo have the highest value compared with the subtrees of the pbo and nblo. The subtrees for *Building* of the nebo and the nblo show a remarkably higher entropy than for other subtrees. The correlation between entropy for the subtrees of all three ontologies and the number of concepts is strong: $r(7) = .97, p < .001$.

		pbo	nebo	nblo
Buildings	number of concepts	15	198	56
	average path length	1.71	3.68	2.72
	average number of subclasses	5.33	3.26	3.56
	entropy	2.67	9.53	5.74
Water vehicles	number of concepts	20	127	21
	average path length	1.77	4.08	3.04
	average number of subclasses	7.00	4.40	2.75
	entropy	2.94	7.50	3.92
Road vehicles	number of concepts	15	97	22
	average path length	1.71	3.59	2.88
	average number of subclasses	5.33	4.59	3.83
	entropy	2.67	7.02	3.80

Table 1: Metrics of the subtrees

3 Experiment

3.1 Method

The experiment was done in eight sessions with a total of 110 participants. The smallest session was done with 5 participants, the largest with 23 participants. All participants were students of the Amsterdam University of Applied Sciences and between 18 and 22 years of age.

We used three videos which showed an object on fire in the domains *Building* (video 1), *Water vehicle* (video 2) and *Road vehicle* (video 3). All the videos were cut to one minute and stripped of sound. Each participant only used one ontology while observing the successively presented situations to prevent interference between the use of different ontologies. The videos were shown to all the participants and each was assigned one of the three ontologies.

For instruction we used an instruction sheet which presented the participants the goal of the experiment, how SAQG uses an ontology to generate questions and the sequence of steps to make the application work, e.g. making connection to the Internet. During the instruction particular attention was given to the possibility of backward navigation, i.e. return to a previous question and the possibility of choosing a superordinate concept when the subordinate concepts are not known or suitable. All the participants were using the same mobile device: an IDEOS X5 with Android 2.2.1. Because the mobile device we used was the same for all participants and most probably different from their own, the participants were given a small amount of time to get used to the mobile device.

We started with a video about an airplane on fire. The results of the determination done by the participants were not used for the experiment. This was done to get an equal starting point for the participants for each video. Otherwise the participants would have to get used to the question answering after seeing video 1 which they would not have to after seeing video 2 and 3. Because all the ontologies did have the same backbone the first question was 'What is on fire' with the same multiple choices for all the ontologies (the top of the subtrees for all the ontologies was the same (see Figure 1).

We logged the data to gather results and create statistics. To gather additional data we developed a questionnaire. After each video we let the participants answer some questions about the application. We asked the following questions: 'Do you miss a concept which describes the object on fire better?', 'What do you think of the sequence of questions?', 'Do you understand all the used concepts?' and 'What do you think of the Graphical User Interface?'. The first question could be answered by yes or no. The other questions could be answered by choosing from a scale of 1 to 5 a number which reflected their evaluation on this topic, where 1 was a negative evaluation and 5 a positive evaluation.

To measure whether a participant chose the right concept for the object on fire which was shown on the video, we developed three metrics. For the first measure we created a 'gold standard'. We observed, as knowledge experts, independent from each other, the presented situations and chose concepts we thought best represented the object on fire shown on each video independent of any ontology. Then we discussed our own preferences and agreed on one concept for each video as being the best representation of the object on fire.

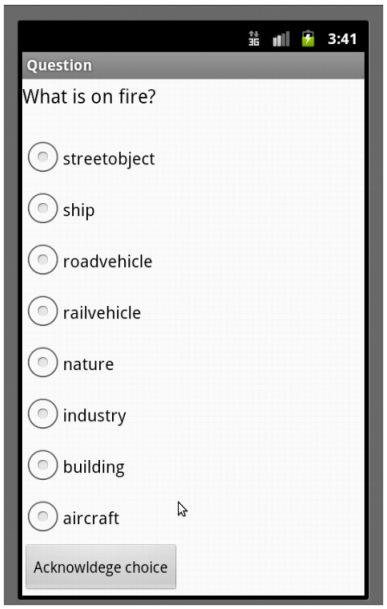


Figure 1: First screen generated from the ontology

For the second measure a concept from each ontology was drawn. This concept was similar to the preferred concept or had a meaning as closely as possible to this concept. We also chose some concepts closely representing the object on fire, i.e. alternatives. In a discussion we determined the perfect answer and the alternative concepts. To compare the results on this measurement per ontology we gave each answer a value: the best answer was given one point and the alternatives half a point.

For the third measure we measured how many different concepts were chosen by the participants, i.e. the variability of the chosen concepts, and how often the most chosen concept was chosen relative to all other choices, i.e. the relative frequency of the most often chosen concept. We expect consensus among the participants about the object on fire for each ontology used after watching the three videos. The level of agreement among the participants indicates the suitability of the ontology for this task.

3.2 Results and analysis

How long it took for the participants to determine the object on fire is shown in Table 2. The participants which used the nebo on average needed much more time than the participants which used the pbo or the nblo. The difference between the time needed to determine the object on fire when using the pbo or the nblo was not significant (two-sample $t(210) = .6, p = .53$). The correlation between the time needed to determine the object on fire and the entropy of the subtree as shown in Table 1 is significant: $r = .73, n = 9, p < .05$. The correlation with the number of concepts is even a bit stronger: $r = .77, n = 9, p < .05$.

	pbo	nebo	nblo
n	36	35	39
video 1	29.54(20.97)	64.73(26.59)	40.49(16.72)
video 2	43.55(34.75)	80.59(39.53)	37.97(31.65)
video 3	25.97(16.35)	34.25(21.15)	26.77(10.48)
mean	33.02(26.16)	59.86(35.52)	35.08(22.17)

Table 2: Mean duration of determination and standard deviation for each video and ontology

Table 3 shows how the ontologies relate to each other when the mean duration of determination for each video is corrected by the average path length of the subtree which was used. The correlation of the mean duration of determination for each video corrected by the average path length with the time needed

to determine an object on fire is not significant. The correlation between the path length as shown in Table 1 and the time needed to determine an object is significant: $r = .68, n = 9, p < .05$.

	pbo	nebo	nblo
video 1	17.28(21.05)	17.47(18.78)	14.89(15.62)
video 2	24.61(30.72)	19.72(21.79)	12.49(16.12)
video 3	15.19(17.43)	9.60(11.21)	9.30(9.88)
mean	19.02(23.07)	15.59(17.26)	12.22(10.54)

Table 3: Mean duration of determination for each video and ontology corrected by the average path length of the subtree

The time needed to determine the object on fire corrected by the number of subclasses is shown in Table 4. The correlation with the time needed to determine an object on fire is rather high: $r = .86, n = 9, p < .05$. The correlation of the number of subclasses of a subtree and the time needed to determine an object using that subtree is not significant: $r = .17, n = 9, p = .66$.

	pbo	nebo	nblo
video 1	5.54(6.75)	19.72(21.20)	11.37(11.94)
video 2	6.22(7.77)	18.28(20.21)	13.81(17.82)
video 3	4.87(5.59)	7.51(8.77)	6.99(7.43)
mean	5.56(6.70)	15.17(16.73)	10.72(12.40)

Table 4: Mean duration of determination for each video and ontology corrected by the number of subclasses of the subtree

In Table 5 the results from the questionnaire are shown. The participants did not miss concepts in the determination process in one ontology more than in another. Taken all ontologies together nearly 63% of the participants did not miss a concept to describe the object on fire. The pbo, which is the smallest ontology, scores worst on this question.

		1	2	3	4
pbo	average	0.59	4.22	4.15	2.83
n = 36	video 1	0.611	4.36	4.36	2.83
	video 2	0.81	3.94	3.44	2.83
	video 3	0.36	4.36	4.64	2.83
nebo	average	0.67	4.03	3.51	3.26
n = 35	video 1	0.69	4.20	3.40	3.29
	video 2	0.49	3.57	3.03	3.17
	video 3	0.83	4.31	4.11	3.31
nblo	average	0.68	4.30	4.44	3.29
n = 39	video 1	0.72	4.23	4.36	3.28
	video 2	0.90	4.36	4.26	3.28
	video 3	0.41	4.31	4.69	3.31

Table 5: Results from questionnaire. 1: 'Do you miss a concept which describes the object on fire better?' (no(1)/yes(0)), 2: 'What do you think of the sequence of questions?' (1-5), 3: 'Do you understand all the used concepts?' (1-5), 4: 'What do you think of the Graphical User Interface?' (1-5).

Comparing the mutual results of the questionnaire we see a difference in mean for the comprehension of concepts between the nebo (3.51) and nblo (4.44) which is significant ($F = 5.77, p < .05$). The difference in mean for the GUI between the pbo (2.83) and nebo (3.26) is significant ($F = 4.71, p < .05$).

When we compare the results as shown in Table 2 with the results on the question whether the participant understands the concepts a significant correlation occurs: $r = .91, n = 9, p < .001$. When

comparing the time needed to determine an object with the results on the evaluation of the sequence of questions also a significant correlation can be found: $r = .83, n = 9, p < .05$. The correlation between the two questions just mentioned is also significant: $r = .84, n = 9, p < .05$. A comparison between the values for the question after the Graphical User Interface and the number of subclasses for each subtree shows a significant correlation: $r = .84, n = 9, p < .05$.

	independent of ontologies			dependent on ontologies		
	pbo	nebo	nblo	pbo	nebo	nblo
video 1	0	6	0.90	31.11	10.50	16.71
video 2	13.61	0	24.23	21.88	6.50	27.37
video 3	0	29	0	35.00	29.00	33.21
total	13.61	35	25.13	87.99	46.0	77.29

Table 6: Validation of the ontologies using the gold standard (normalized) independent and dependent on ontologies

Table 6 shows the score for each ontology using the gold standard created independent of the ontologies. For each time a participant chose the preferred concept an ontology scores a point and when an alternative was chosen half a point was given. To compare the ontologies a normalization has been applied because the number of participants using the pbo was 36, the nebo 35 and the nblo 39, the result was corrected by respectively $\frac{35}{36}$, 1 and $\frac{35}{39}$. It is clear that nebo scores best and pbo worst using this measurement. For participants which used the pbo it was not possible to choose the right concept when seeing video 1 and 3. The same holds for the nebo with video 2 and the nblo with video 3. The participants which used the nblo with video 1 who wanted to choose *Bungalow*, had to choose *Holiday accommodation* (*Vakantieverblijf*) first instead of *Residence* (*woning*). Most participants chose *Residence*. Table 6 also shows the score for each ontology using the gold standard and alternatives created with use of concepts which are part of the ontologies, i.e. dependent on ontologies. The lowest score is for nebo. The best score is for pbo. Due to the strong variability of choices made by participants using the nebo, this ontology does not score high on video 1 and 2. Only with video 3 the nebo scores nearly as high as pbo and nblo. The variability of choices for all ontologes after seeing video 3 is much smaller.

Subject	pbo		nebo		nblo	
	#	%	#	%	#	%
Buildings	2	77.78	11	22.86	5	89.74
Water vehicles	6	38.89	12	37.14	5	69.23
Road vehicles	1	100.00	4	80.00	2	94.87
Average	3.00	72.22	9.00	46.67	4.00	84.61

Table 7: Variability (#) and relative frequency of the most often chosen concept (%)

Table 7 shows how many different choices were made by the participants (variability) and the mode relative to the total number of choices, i.e. how often the participants chose for the most frequently chosen concept, relative to the total number of choices made. The pbo scores best with respect to the variability. The best score is for nblo and the worst score for nebo with respect to the relative mode.

4 Discussion

We found five characteristics of the ontologies which have an influence on the the time needed to determine an object: path length, comprehension of concepts, sequence of questions, number of concepts and entropy. The path length of an ontology has a direct influence on the time needed to determine an object because it is a measure of the number of questions asked. Whether participants comprehend concepts has a less direct influence and is probably dependent on the familiarity of the participant with a particular domain, which can vary. The reason why the sequence of questions has influence is less clear. It might be a derivative of the number of questions and thus the path length. The number of concepts

and entropy have a strong correlation although the measurement of entropy also incorporates path length and number of subclasses.

The analysis of the choices made by the participants which was done using a gold standard shows rather confusing results. Using a gold standard set up by knowledge experts independent of the ontologies shows a best score for nebo, using a gold standard of concepts drawn from the ontologies shows a best score for pbo. To make matters even more confusing, when only the choices made by the participants are taken into account the best score is for nblo.

The gold standard independent of the ontologies is, to our opinion, a measurement for evaluating the completeness of an ontology, i.e. the number of concepts from an ontology also found in a corpus relative to the total number of concepts in that corpus, i.c. natural language. It makes clear the difference of determination of an object on fire by knowledge engineers on the one hand and ordinary people on the other. The gold standard drawn from the ontologies does measure the suitability of the ontologies for use by ordinary people much better. The outcome would even be similar to the outcome of Table 7, which shows the consensus among the participants, when for video 1 (*Building*) the gold standard would have been *Residence* (*Woning*) instead of *Bungalow*. The choice for *Bungalow* was again due to our own knowledge engineering. Erroneously omitting the concept *Cruiseship* from nebo did not have an effect on the overall results. The variability, as shown in Table 7, is the lowest for the pbo and highest for the nebo showing that the last ontology possibly offered too many concepts for the participants to be clear about which object was on fire. The pbo did offer much less concepts and scored better in this respect. The nblo scored nearly as good as the pbo on variability. The measurement of the relative frequence, as is shown in Table 7, shows for the nblo the best score although the pbo scored on video 3 an ideal score and was overall nearly as good as the nblo. With this measurement nebo scores worst, again. A measurement of the relative frequency is, in our opinion, most informative about the suitability of the ontologies for the task we envisage for it measures the unanimity among the participants about the object on fire better than the variability because this only shows how many different concepts are chosen and not how many times the most chosen concept was chosen.

5 Conclusion

In this paper an experiment is presented which was conducted to measure whether one of three ontologies is most suitable for the task of answering of automatically generated questions by ordinary people for determination of a situation. The three ontologies were each developed in a different way. Before we used the ontologies, two of the ontologies had to be re-engineered.

The participants understand the concepts used in the pragmatic-based ontology but evaluate the Graphical User Interface, relative to the nebo, negative. This is probably due to the large average number of subclasses of the pbo. The pbo scores best on the time needed to determine an object on fire (but the difference with the time needed when using the nblo is not significant). Using the 'gold standard' independent of the ontologies the pbo scores worst but when using the 'gold standard' with concepts drawn from the ontologies the pbo scores best. The pbo shows the smallest variability of choices when determining an object on fire. The average relative frequence of the most often chosen concept of the pbo is rather high but not the highest.

The concepts used in the new expert-based ontology are the least understood relative to the nblo. Using the nebo it took the participants the most time to determine an object on fire. Using the 'gold standard' independent of the ontologies the nebo scores best but when the concepts are drawn from the ontologies the nebo scores worst. The nebo shows the largest variability and the highest relative mode.

The participants do understand the concepts used in the new basic-level ontology rather well. When comparing the time needed to determine an object on fire the nblo scores nearly as good as the highest scoring ontology. The measurement using the 'gold standard' shows for nblo a mediocre result. The variability is rather low but not the lowest. The average relative mode is the highest of the three ontologies.

When predicting the time needed to determine an object on fire the average path length and the number of concepts used in an ontology are good indicators. Path length has a rather strong correlation with the time needed to determine an object. The number of subclasses does not have a correlation with the time needed to determine an object. Comprehension of the concepts also has a strong influence on the time needed to determine an object but is dependent on the specific knowledge of a domain users

have.

We conclude that each of the ontologies has its own merits and idiosyncrasies. When a very accurate determination of an object is required and users have expert knowledge about the domain, one should use the nebo. A longer time to determine such an object has to be accepted. The nblo scores nearly as good or better than the pbo on most respects, such as time needed to determine an object and consensus about the concept which denotes the object on fire best. The nblo is evaluated more positive than other ontologies with respect to the user interface. A great disadvantage is the high costs to develop a nblo. When an application such as SAQG is deployed on a large scale for many people such an effort could be worthwhile. The concepts and their categorization represented in the ontology used by SAQG to generate questions and possible answers should maximally comply with the language used by ordinary people. The new basic-level ontology has the best results in this respect.

References

[1] AATNed. *http://www.aat-ned.nl/*, 2012.

[2] M. Jiang and W.L. McGill. Participatory Risk Management: Managing Community Risk Through Games. In *Social Computing (SocialCom), IEEE Second International Conference on*, pages 25–32. IEEE, 2010.

[3] M.M. Kokar, C.J. Matheus, and K. Baclawski. Ontology-based situation awareness. *Information Fusion*, (10):83–98, 2009.

[4] L. Palen, K.M. Anderson, G. Mark, J. Martin, D. Sicker, M. Palmer, and D. Grunwald. A vision for technology-mediated support for public participation & assistance in mass emergencies & disasters. In *Proceedings of the 2010 ACM-BCS Visions of Computer Science Conference*, page 8. British Computer Society, 2010.

[5] M. Teitsma, J. Sandberg, M. Maris, and B. Wielinga. Using an Ontology to Automatically Generate Questions for the Determination of Situations. In *Database and Expert Systems Applications*, pages 456–463. Springer, 2011.

[6] M. Teitsma, J.A.C. Sandberg, M. Maris, and B.J. Wielinga. Automatic question generation to determine roles during a crisis. In *SOTICS 2011, The First International Conference on Social Eco-Informatics*, pages 37–42, 2011.

[7] Marten Teitsma, Jacobijn Sandberg, Guus Schreiber, Bob Wielinga, and Willem Robert van Hage. Engineering ontologies for question answering. *Applied Ontology*, 2014.

[8] G.A.C.M. van Heijst. The role of ontologies in knowledge engineering. 1995.

[9] P. Vossen, I. Maks, R. Segers, and H. van der Vliet. Integrating lexical units, synsets and ontology in the Cornetto Database. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, 2008.