Talking Rabbit: a User Evaluation of Sentence Production

Paula Engelbrecht[1] Glen Hart[1] and Catherine Dolbear[1]

[1] Ordnance Survey, Romsey Road, Southampton, Hampshire SO16 4GU United Kingdom
{Paula.Engelbrecht, Glen.Hart, Catherine.Dolbear}@ordnancesurvey.co.uk

1 Introduction

Ordnance Survey, the national mapping agency of Britain is researching the development of topographic ontologies to describe our spatial data, motivated by the business need of improving ease of understanding and reuse of our data, as well as assisting its integration with third party datasets. Through our experience of authoring topographic ontologies, we have found that the family of Web Ontology Languages (OWL), the W3C standard for ontology authoring, are difficult for domain experts to master. This insight led to the development of a controlled natural language, Rabbit [1], which is designed to make it easier for domain experts – people with expertise in a particular subject area, but not versed in knowledge modeling or Description Logics – to understand and author their own ontologies. With the possible exception of Sydney Syntax (SOS) [2], Rabbit differs from other controlled natural languages such as ACE [3] and CloNE [4] by limiting sentence complexity to ensure that sentences are simple and describe single concepts. Further more its development differs from other similar CNLs in two main ways. Firstly, its development has been informed by the building of several medium-scale (approximately 500-concept) ontologies, with updates to Rabbit being based on the requirements that have arisen from these real-world knowledge modeling needs. Secondly, we have modified Rabbit based on user evaluations: testing domain experts' understanding and authoring of Rabbit. Rabbit comprehension tests were reported in [5], and this paper now describes the results of our user evaluations of Rabbit authoring.

2 Experimental Design

Twenty-one Ordnance Survey employees volunteered to participate in the evaluation. The materials they were given included a short introductory slide show outlining the basic principles of Rabbit and explaining the evaluation task. A crib sheet that succinctly outlines 17 essential Rabbit keywords and expressions was also provided. For example, the crib sheet outlines the use of 'is a kind of' to denote that

one class is a subclass of another (e.g. "Every Beck is a kind of Stream"). Another example is the use of 'one or more of' to specify that the subject can be associated to more than one of the concepts in the list (e.g. " Every River flows into one or more of a River, a Lake or a Sea"). Finally, a short text describing an imaginary domain (the planet Zog and its inhabitants) was provided. This fictitious domain was chosen to control for pre-existing domain knowledge. The participants were then given 45 minutes to write as much about planet Zog as they could using Rabbit. Two sets of analyses were performed. The first analysis consisted of a comparison between the information content participants attempted to capture in their Rabbit sentences and that of a Rabbit language expert, who completed the study under test conditions. The sentences this expert generated were broken down into 48 units of information, which could be expressed as "subject predicate object" triples of varying complexity (e.g. Bimbles [subject] only exists [predicate] on Zog [object]; Bimbles that breed [subject] always breed [predicate] in a lake [object]; Adult Bimbles [subject] can not [predicate] both work and breed [object]). If a participant managed to capture all 48 units of information then they were given a score of 100%. The second analysis consisted of an evaluation of the types of errors novice users make when writing Rabbit sentences. The Rabbit sentences were scored by three expert users of Rabbit. In cases where there was disagreement about the presence or type of an error a consensus was reached through discussion. The proportion of sentences that contained errors was calculated for each participant by dividing the number of erroneous sentences by the total number of sentences.

3 Results

The amount of information conveyed ranged widely between participants with an average score of 60% (StD = 16%), a low score of 29% and a high score of 79%. Performance on individual sentences also varied widely, the highest scoring units of information being mentioned by all participants (100%) and the lowest only being mentioned by 4 participants (19%). On average, 51% (StD = 22%) of the sentences generated contained at least one Rabbit language error. Again, individual performance varied widely, with the most accurate participant producing errors in 16% of sentences and the lowest scoring participant producing errors in 94% of sentences. Participants' overall accuracy was quite poor with 50% of sentences generated containing at least one error. Some specific issues were identified:

Core versus secondary concepts
In general, the participants were very good at capturing information relating to the main protagonists of the text. For example, all participants mentioned that Bimbles have three eyes and two arms. In fact, all of the sentences that were in the top 25% correct were about Bimbles. It was found that the less related information is to the main protagonists, the more likely it is to be left out. For example, although 67% of participants mentioned that Bimbles don't eat fish from Lake Quish, only 38% of participants mentioned that they are not eaten because they are poisonous. Even fewer participants mention the chain of causality that causes fish in Lake Quish to be

poisonous. An improvement to the experiment might have been to include a scope and purpose for the ontology being built, to ensure that participants knew what in the text was considered important. However, these results provide supporting evidence for our use of "Core" and "Secondary" concepts within Rabbit, as the test shows that people naturally categorise information as being more or less important to the central purpose of the ontology.

Omitting implicit information

Participants tended to leave out information which is implicit in the name of an entity. For example, only a third of the participants mentioned that Lake Quish is a kind of lake. This information would not need to be explained if one were to tell another person about planet Zog. Participants were also more likely to mention information which holds for the majority of cases than information which holds for only a few cases. For example, although all of the participants mentioned that "Bimbles are usually red", only 42% mentioned that blue Bimbles exist. Participants also tended to omit restrictions. For example although the facts that Zoggian fish are killed by fish rot, predation and old age was mentioned by 67%, 47% and 67% of participants respectively, only 33% of participants mentioned that these are the only things that can kill Zoggian fish.

Leaving out Every

The most common Rabbit error found was the omission of the word "Every" at the beginning of sentences. In fact, if one recalculates the error scores omitting all those sentence in which omission of the word "every" is the only mistake, then the average error rate drops to 43% (StD = 18%). A T test shows that the difference in error percentage between the set of scores that omits "every only" errors and the set of scores that includes them, is statistically significant, $t(19) = 3.11$, $p < .01$. Based on this finding, whether to continue to require the use of the "Every" keyword, is an ongoing area of debate.

Instance versus subclass

The second most common error was to confuse instance declarations and assertions about one concept being the subclass of another, or to mix elements of both. Participants frequently used "is a" where they should have used "is a kind of". For example "Lake Quish is a Lake" would be the correct way to declare that Lake Quish is an instance of the "Lake" class. Removing sentences which only contain this kind of error from the error analysis reduces the overall error rate further from 43% (StD = 18%) to 39% (StD = 18%). This difference was significant, $t(19) = 3.34$, $p = .01$.

Open world assumption

Participants found it difficult to model knowledge under the open world assumption and understand the meaning of Rabbit keywords "only" and "only ... or nothing". In Rabbit "only" is used to denote that the object of a sentence has to apply to the subject, and that it will be the only object that can be linked to the subject via that relationship. The use of "only .. or nothing" in a sentence, denotes the same than "only" with the concession that the relationship does not have to apply to the subject.

The words "or nothing" have been erroneously omitted from the sentence "Every Bimble only breeds in a lake" (Participant 1). This is an error because not all Bimbles breed. Conversely, the sentence "Every Bimble only eats Zoggian Fish or nothing (Participant 1) is also incorrect, because every Bimble eats. In fact the text given to the participants specifies that every Bimble has a natural imperative to eat.

4. Conclusions

The aim of the current study was to establish what type of errors novice Rabbit users would make. An analysis of these errors can be used to inform both changes to the Rabbit language itself as well as to the ROO Protégé plugin (an ontology authoring tool that implements Rabbit) [6]. The use of a software tool is supported by the result that participants varied widely in the amount of information they included in their text. Such a tool would ensure that those who tend to be more restrained in the amount of information they express, include all the information necessary. For example, the tool could ensure that concepts which are introduced as objects in sentences are also defined, if they are considered to be core concepts based on the scope and purpose. Furthermore, a supporting tool that completes things that need to be expressed repetitively (e.g. a relationship that is used over and over again) might encourage these individuals to express more information.

Similarly, tool support can again help to avoid confusion between instance and subclass statements, for example by automatically suggesting the alternatives ".. a kind of" or "... a" whenever the keyword "is" is typed in by the user, and prompting them to choose one of these alternatives. A similar solution could be introduced for other reserved words such as "exactly", "no", and "at most". Specifically, the application could store synonyms for the most common keywords and make the correct suggestion when these are typed in. For example, the software could suggest the use of "exactly" when the word "precisely" is entered.

This article has been prepared for information purposes only. It is not designed to constitute definite advice on the topics covered and any reliance placed on the contents of this article is at the sole risk of the reader.

References

1. Hart, G., Dolbear, C., & Goodwin, J.: Lege Feliciter: Using Structured English to represent a Topographic Hydrology Ontology. In: Proceedings of the OWL Experiences and Directions Workshop (2007). Horridge, M., Drummond, N., Goodwin, J., Rector, A., Stevens, R., & Wang, H. H.: The Manchester OWL Syntax. In: Proceedings of the OWL Experiences and Directions Workshop (2006).
2. Cregan, A., Schwitter, R., & Meyer, T.: Sydney OWL Syntax—towards a Controlled Natural Language Syntax for OWL 1.1. In: Proceedings of the OWL Experiences and Directions Workshop (2007).
3. Kaljurand,K. and Fuchs, N.E. Mapping Attempto Controlled English to OWL DL. Poster demo at the Third European Semantic Web Conference (2006)

4. Funk, A., Tablan, V., Bontcheva, K., Cunningham, H., Davis, B., & Handschuh, S.: CLOnE: Controlled Language for Ontology Editing. In: Proceedings of the 6th International Semantic Web Conference (ISWC07) (2007).
5. Hart, G., Johnson, M. & Dolbear, C.: Rabbit: Developing a controlled natural language for authoring ontologies. In S Bechhofer, M. Hauswirth, J. Hoffmann & M. Koubarakis (Eds). The Semantic Web: Research and Applications, LNCS, 5021, 348-360, (2008)
6. Denaux, R., Dimitrova, V. Cohn, A.G, Dolbear, C. and Hart, G ROO: Involving domain experts in authoring OWL ontologies In Proceedings of the Seventh International Semantic Web Conference (2008)