QuantML: a comprehensive study of the annotation scheme for quantification and its future applications

Mahzabin Yasmin Binte Amin, Annajiat Alim Rasel, Md Humaion Kabir Mehedi
Brac University, Dhaka, Bangladesh
mahzabin.yasmin.binte.amin@g.bracu.ac.bd, annajiat@gmail.com,
humaion.kabir.mehedi@g.bracu.ac.bd

Abstract

Structural ambiguity, the application of arguments to multiple argument sets, is a frequent occurrence in sentences. A phenomenon that can almost always be attributed to quantification expressions that are very common in both spoken and written texts, its proper interpretation in applications connected to natural language processing is essential. Despite the widespread occurrence and implications of quantification, no in-depth annotation scheme has been put forward thus far. Existence of annotation schemes like the ISO-TimeML that primarily deals with temporal quantifications falls short in dealing with everything related to quantification expressions.

In logical studies, quantification has been limited to the universal and existential quantifiers. The universal and existential quantifiers are sets of all the elements and at least one element in a domain respectively. The study of generalized quantifiers such as “less than half of”, “most” “seven”, etc. gave birth to the generalized quantifier theory (GQT), a theory that played an integral part in highlighting the distinction between quantification in logic and that in natural language. Even though words equivalent to “all” and “some” in English may seem to exist in the same group as the universal and existential quantifiers of formal logic, and words like “five”, and “most” with general quantifiers, it is besides the case. In natural language, we cannot express quantifications that range over all individual objects in a universe of discourse unlike in logic. This is because saying something is true “for all” or “for some” implies something is true for all or some conceivable object and in natural language it is impossible to make this assumption. We require a more specific domain that the quantification can be restricted to. Quantifiers in natural language are more often noun phrases or adverbial expressions (temporal and spatial quantifiers) and not determiners.

The neo-Davidsonian approach makes the semantic roles of the participants in an event more explicit. This allows representation of some quantifications from the point of view of the set of participants involved in an event. In natural language, a noun phrase comprises two sections, a noun or head, and one or more determiners.

Both these approaches are integral in addressing the issues relating to ambiguity in quantification. We will look into how this is achieved using a relatively new but prospective annotation scheme, QuantML.

The restrictor is what fundamentally differentiates quantification in natural language from quantification in logic. Quantification in NL is restricted to a source domain, the set of entities referred to by the restrictor, instead of all entities in a given universe of discourse. Quantification in NL is further restricted to a specific part of a domain. Ambiguity is generated in quantification in one of three ways: specificity, distributivity and individuation.

Automatic annotation processes also fail to foresee all possible readings of quantifications due to its lack of understanding and assessment when it comes to general world knowledge and information that is based on a situation or context. Hence, it is of utmost importance that an automatic annotation, much like a manual automation, is not forced to make very particular choices that are not supported by the available information, skills or resources. The multiplicity of possible readings of quantifications

can pose a colossal challenge for language understanding systems. On the flip side, if sufficient information and skills are indeed available, there should not be an issue for the system to make precise annotations. The ability to allow specifications with varying degrees of granularity is desirable for a useful annotation scheme.

Another source of ambiguity comes about due to quantifier distributivity that may diminish logical precision in interpretation. Ambiguity in natural language quantification is mostly considered in terms of the number of logically precise interpretations.

Another source of ambiguity arises from the 'individuation' of a quantification. The count/mass distinction is often characterized semantically in terms of 'individuation'. This relates to count nouns that have a domain of reference made up of individuals, a mass noun is made up of entities or 'quantities'.

The paper not only showcases how QuantML that is based on both GQT and the neo-Davidsonian approach tackles such varying levels of ambiguity but also outlines the annotation scheme that has been designed according to the ISO principles of semantic annotation (ISO standard 24617-6). The scheme has a three-part definition including i) an abstract syntax that identifies and states the possible annotation structures at a conceptual level, ii) a semantics that details the meaning of the annotation structures that have been previously defined by the abstract syntax and iii) a concrete syntax, that defines a representation format using XML expressions.

We will also delve into the uses of QuantML and its prospects in improving interpretations in natural language processing. We will also look into the loopholes that exist in the scheme in the status quo and the possible changes that can be made to bring improvements in this regard.