Annotation Study Concept Formation

TECHNISCHE UNIVERSITÄT DARMSTADT

OntoWiktionary – Constructing an Ontology from the Collaborative Online Dictionary Wiktionary
by Christian M. Meyer and Iryna Gurevych

Goal

Concepts are the basic building blocks for ontologies. Each concept should denote a single entity of world that is modeled by the ontology. Entity thereby includes real world objects, abstract ideas, processes, states, etc. Besides a textual description, a concept can be represented by lexicalizations, i.e. certain terms or expressions that directly refer to the concept. The concept 'DOG' could, e.g., be modeled for representing all instances that are denoted by the word 'dog' in our world. The noun 'dog' (in the animal sense) thus serves as a lexicalization of 'DOG', which might also be represented by a second lexicalization using the noun 'hound'.

The goal of this annotation study is to validate the consistency of semi-automatically learned concepts. The concepts are represented by different lexicalizations that are explained by a short textual definition. The annotation study is intended to analyze the overall quality of the creation approach and if errors rather occur in the lexicalizations or their definitions (i.e. their meaning).

Setup

- The dataset is given as an Excel sheet.
- The data is organized in sections.
    - Each section starts with lexicalizations in the first column and their textual definition in the second column – one pair of lexicalization and definition per row.
    - After the list of lexicalizations/definitions, you're asked to answer the question "Is the above concept consistent?" (see further explanation below). Please type your answer in the first column (before the question).
    - You can also leave a comment for each lexicalization/definition in the third column.
- Concepts whose lexicalizations AND definitions are consistent, i.e. they all belong to the same entity, should be marked with "1" (see Example #1).
- Concepts whose lexicalizations but NOT definitions are consistent, i.e. there is a meaning for all lexicalizations that belong to the same entity (although at least one of them is associated with the wrong definition), should be marked with "2" (see Example #3 – there is another meaning for the word 'bass' that is commonly used in the English language).
- Inconsistent concepts should be marked with "0" (see Example #2).

Please note

- Do not imitate an algorithm!
- Sometimes lexicalizations from different parts of speech are put together. You should accept ("1" or "2") these concepts if the meaning of these lexicalizations belong to the same entity (see Example #4).
- Reject concepts ("0") that have at least one lexicalization that is very broad or very narrow. In Example #2, the 'bass' is a certain type of a 'singer', so one would expect two different concepts here.
- You should ignore very subtle differences in the lexicalizations. The concept "statement that does not conform to the truth" can, e.g., be lexicalized as 'lie' and 'misrepresentation'. A lie, for instance, usually infers deceiving someone, while a misconception can be simply due to ignorance. This kind of subtle difference should be accepted ("1" or "2").
- Sometimes the textual definitions are trimmed or normalized by our extraction method. Additionally, there are some special characters or format commands within the definitions. This should be ignored for the judgment.
- You are allowed to use any additional resource for grounding your judgment, including dictionaries, lexicons, the Web, and particularly Wiktionary (http://www.wiktionary.org), which was used as a source for the textual definitions.
- Synsets within the Princeton WordNet are similar to the concepts that are to be judged in our study (besides the restriction to one part of speech). Experience in the work with WordNet can thus be used to judge the concepts. Caveat: Not every lexicalization in our study is part of WordNet or is always in a consistent synset. Do not directly compare to WordNet, but stick to your own judgment.
- The annotation process should take between 1-3 hours. Please write the time you needed for the study at the end of the sheet.

Lemma	Gloss	Comment
bass	[N] A male singer who sings in the bass range.	
basso	[N] A bass singer, especially in opera.	
1	**Is the above synset consistent?**	***Example #1***
bass	[N] A male singer who sings in the bass range.	
basso	[N] A bass singer, especially in opera.	
singer	[N] person who sings, is able to sing, or earns a living by singing.	*WRONG: too broad*
0	**Is the above synset consistent?**	***Example #2***
bass	[N] The perch; any of various marine and freshwater fish resembling the perch, all within the order of Perciformes.	*WRONG SENSE*
basso	[N] A bass singer, especially in opera.	
2	**Is the above synset consistent?**	***Example #3***
singer	[N] person who sings, is able to sing, or earns a living by singing.	
sing	[V] To produce harmonious sounds with one's voice.	
1	**Is the above synset consistent?**	***Example #4***