Words Constituting Phrases in Ontology

Min Ding   Fumihiro Matsuo
Graduate School of Information Science and Electorical Engineering
Kyushu University, Hakozaki, Fukuoka, 812-81, Japan

1. Introduction

In recent years, the AI community has been evolving a notion of ontologies as artifacts that play significant roles in knowledge representation and reasoning. An ontology is a description of concepts and their inter-relationships. It may take a variety of forms, and a very simple case would be a general-to-specific hierachy.

In constructing ontology, identifying and defining the concepts is its first phase. Terms to represent the concepts can be a single word or a phrase. It is observed that there exist some phrases of which stemmed content word is the direct superordinate of that phrase itself. Such phrases are to be called ISA phrase below. For example, the word "engine" is the direct superordinate of the phrase "aircraft engine", and the word "work" is the direct superordinate of the phrase "work in progress". Here, terms like "aircraft", "in progress" modify the superordinate words, and we call them attaching terms in this paper. The attaching terms seem to have contribute to making a word from general to specific.

We focuse on the characteristics of words constituting phrases in Ontology, and our initial step is to discover the common characteristics of those attaching terms. In this paper, the ISA phrases in WordNet are examined.

2. ISA Phrases in WordNet

WordNet is a sematic word database based on psycholinguistic principles. To use the authors' words,

WordNet "... is an attempt to organize lexical imformation in terms of word meaning, rather than word forms...". To this end, word forms, single words or phrases with synonymous meanings are grouped in sets, called synsets. A phrase in WordNet is a string of two or more single words, connected by spaces or hypens. WordNet noun synsets are organized hierarchically according to ISA relation. The synsets, which appear at the lowest level of the hierarchies, are defined as minimum synsets, and the phrases in the minimum noun synset are examined in this paper.

In the noun category, WordNet contains about 87642 noun word forms, including 42256 single noun words and 45386 noun phrases, organized into about 60588 synsets. The number of the minimum synsets is 47110, while the number of the phrases among them is 40327.

3. Related Work

3.1 Retrieve the ISA phrases

Our first step is to retrieve all the ISA phrases in the minimum synset, there are total 11425 ISA phrases in the minimum synsets of WordNet. The next is to identify the superordinate word and the attaching terms in these phrases.

Let X be a superordinate word, Y(or Y1,Y2) be an attaching term, then the phrase is one of the following forms, X-Y, Y-X, or Y1-X-Y2 (Table1).

Table1. Attaching Terms in ISA Phrases

	ISA Phrase		Different Y
X-Y	820	7.2 %	742
Y-X	10530	92.2 %	5648
Y1-X-Y2	75	.6 %	/
Total	11425	/	/

In the form X-Y, the number of different attaching terms Y is 742, and in the form Y-X, it is 5648. The

case Y1-X-Y2 is not studied yet.

3.2　Analysis of the Attaching Terms

In the forms of both X-Y and Y-X, the attaching terms are devided into two parts. The first part includes terms retrievable in WordNet and it is analysed below. The second part, however, includes terms unretrievable in WordNet, and it is characterized by special forms of these terms. For example, the possission form(poss.) (eg. *driver's* license), the preposition form(prep.)(eg. judgement *by default*, *flight of* stairs), or special words(spec.) (eg.*Bedford* cord) (Table2)

Table2. Attaching Terms

	X-Y		Y-X	
Part I	387		4198	
Part II	355	poss.87	1450	prep.175
		spec.216		other 1275
		prep.14		/
		other 41		/
total	742		5648	

In WordNet, synsets are organized into 45 lexicographer files based on part of speech and logical groupings. Referring to it, the attaching terms in Part I which are retrievable in WordNet noun and verb categories are taxonomized into 29 groups. As for those retrievable in adjective and adverb categories, they need further studing and have just been classified into the group "other".

Since the same word form may appear in more than one synset, it may be ambiguous. So while taxomizing, the first problem is the determination of which sense the present attaching term reprsents. We have made use of an integer sense number in WordNet. It is given for sense disambiguation, and for every word forms, senses are generally ordered from most to least frequently used, with the most common sense numbered 1. Here it is supposed that all the attaching part represent the sense numbered 1 in WordNet. Table3 presents the results of our taxonomy.

4.　Concluding Remarks

Since we have not yet dealed with the verb, adjective and adverb phrases, the result is still very limited. There are several items on our immediate agenda.

Besides, WordNet also presents some phrases of which stemmed content word is the direct subordinate of the phrase. There are about 75 such phrases in WordNet. For example, "baseball" is the direct subordinate of phrase "baseball equipment", "estate" is the direct subordinate of phrase "personal estate". So such cases are at the same time need to be noticed.

Table3. Taxonomy of Attaching Terms

classification	Lexifile no.	X-Y	Y-X
animals	05	121	169
plants	20	74	280
body parts	08	16	88
man-made objects	06	42	566
natural objects	17	6	96
natural phenomena/events	11 19 22	6	115
substance	27	17	199
social activities/events	41	0	3
people	18	11	288
attributes of people/objects	07 24	7	98
groupings of people/objects	14	7	102
acts/actions	04 35 38	6	322
communication	10 32	16	261
eating/drinking	13	5	159
change/intensifying	30	0	7
creation/performing	36	0	2
cognitive processes/contents	09	7	78
feelings/emotions	12 37	0	20
states of affairs	26 42	2	78
spatial position	15	6	170
shape	25	1	33
time/temporal relations	28	6	79
quantities/units of measure	23	8	53
motive/goals	16	0	4
possession	23	3	70
other	0 1 2 44	18	849
total	/	387	4198

Reference

1) Uschold,M, M.Gruninger(1996). Ontologies: Principles, Methods and Application. Knowledge Engineering Review, 11(2).

2) Miller,G.A, R.Beckwith, C.Fellaum, D.Gross,& K.J.Miller(1990). Introduction to WordNet: An on-line lexical database. Journal of Lexicography, 3(4):235-244.

3) WordNet 1.6 Reference Manual. http://www.cogsci.princeton.edu/~wn/.