Online Proceedings, Midwest Artificial Intelligence and Cognitive Science Society Conference (MAICS96), Bloomington, IN, April 1996. http://www.cs.indiana.edu/event/maics96/Proceedings/old.

Synonymy and Word Equivalence

L. John Old

University of Arkansas at Little Rock, jold@ai.ualr.edu

Abstract

Two words which share a sense are commonly called synonyms. They are equivalent in the context of that sense, and can be used interchangeably in a sentence which utilizes that sense. Two words are rarely synonyms for each other for all of their senses. With the advent of on-line-accessible lexicons it is possible to identify those words which are, in all contexts, interchangeable or `equivalent'. This paper describes an analysis of synonym sets (`semi-colon groups') in Roget's International Thesaurus (RIT3) in regard to such equivalence.

Introduction

Roget's Thesaurus is organized as a hierarchy of concepts. The lowest level in the hierarchy are sets of words, commonly called synonyms, which describe the same concept at that point. These lowest level concepts may be referred to as senses of the words (or phrases) found amongst the sets. For example, the set: {above, higher than, over, superior to} describes one sense of `over'. It also describes one sense of `above'.

There are twenty-two senses of `over' and seven senses of `above' in RIT. All but one of the seven `above'-senses are shared by `over'. `Over' and `above' are two words commonly thought of as synonyms, regardless of context. When describing the position of a light fixture attached to a ceiling, in relation to a table below it, the two words are interchangeable and almost impossible to discriminate.

Consider, on the other hand, the phrase ``the house is over (beyond) the hill." Here, `above' cannot be substituted. Similarly, the word `over' cannot be substituted for the word `above' in the sentence ``It is above me." (as in, ``It is beyond my comprehension.").

There is obviously a strong relationship between `above' and `over' but, as demonstrated in the examples, they are not equivalent in all contexts. The question of which synonyms are equivalent in all contexts, then arises. These words will be called *word equivalents*, and are the focus of this discussion.

Analysis of Word Equivalences in RIT

The word equivalents derived for this study always occur together, in all synonym sets in RIT--they share all of their senses. Of about 114,000 distinct word forms (or strings) in RIT there are 5,612 distinct word forms which are word equivalents to at least one other word. They form 8,294 pairs of word equivalents, which is more than 5,612 pairs because, in some cases, they form larger sets of equivalent words, not just pairs. The largest set contains eight word equivalents.

Scrutiny suggests that word equivalents are of several types. Below are examples from the data which have been grouped under different headings in order to make explicit the author's interpretation of these types.

Abbreviated:
amidst, amid, midst, and 'mid; abaft, abaff, and baff; capability and ability; accompanyist and accompanist
Coined abbreviation: (rare)
A-bomb and atomic bomb; H-bomb and hydrogen bomb
Strict abbreviation: (very common)
F.B.I and Federal Bureau of Investigation; I.Q. and Intelligence quotient
Foreign equivalent:
North Star and l'Etoile du Nord; pigeon post and Taubenpost; headfirst and a corps perdu; afterthought and arri`ere-pens'ee
Foreign - foreign:
Gaucho and vaquero; regards and devoirs; ad rem and apropos
Rough equivalent:
Anglicism and Briticism; Asian and Asiatic; Satanism and diabolism
Spelling variant:
Fü hrer and Fuehrer; Odin and Woden; Papist and papist; airplane and aeroplane
Pseudonym:
Cupid and Eros
Indirect reference:
Azrael and death's bright angel; theory of relativity and Einstein's theory
Same-idea-different-cultural-history:
Caesar and czar; John Bull and Uncle Sam
Technical versus lay equivalents:
Insecta and insects; alexia and auditory amnesia
Slang equivalent:
Caucasian and paleface
Slang - slang:
Fritz and Jerry
Shared quality:
Lilliputian and Tom Thumb; Brobdingnagian and Goliath; snooper and Paul-Pry; Mister and Master
Qualifier representative:
Shetland and Shetland pony
Commonly associated:

Dark Ages and Middle Ages; Washingtonese, federalese, and officialese absolute-item-, limited-, pocket-, qualified-, and suspensory veto

Antipodals:

North Pole and South Pole; autumnal equinox and vernal equinox

Equivalent idiom:

I assure you and believe me; an eye for an eye and tit for tat; abreast of the times and up-to-the-minute; automatic factory and push-button plant

Word - phrase:

arrive and get there; autism and dereistic thinking

Rephrasing:

aching heart and heartache; nonplussed and at a nonplus

Regular, normal synonyms:

V-shaped and crotched; abolishment and abolition; abjurement and abjuration; absorbent and adsorbent; absurdly and ridiculously; accountant and bookkeeper

Equivalence and Synonymy

Are equivalent words, synonyms? Strictly speaking they are the only `true' synonyms in RIT. Intuitively, however, they are not (except for the last group) what we usually call synonyms. It appears that, if they are included as synonyms, then the synonymy relation should be somehow graded to enable discrimination.

A first pass is to define a relevancy metric R(w1,w2) between words, defined as the number of senses shared by any two words or sets of words, w1 and w2, in relation to all of either word's senses. (s(w) denotes the set of senses of word w.)

$$R(w1,w2) = \frac{|s(w1) \text{ intersected with } s(w2)|}{|s(w1)|}$$

This has the advantage that the values range between zero and one, where R(w1,w2)=0 indicates two semantically disjoint words and R(w1,w2)=R(w2,w1)=1 indicates two equivalent words--with degrees in between.

This metric has the drawback that the relevance of w1 to w2 is often different from the relevance of w2 to w1 because their polysemy is often different. For example, the relevance of over to above is R(above,over)=6/7 = 0.86 while the relevance of above to over is R(over,above)=6/22= 0.27

An alternative metric is to define the denominator as the total of the senses of both words:

$$R(w1,w2) = \frac{|s(w1) \text{ intersected with } s(w2)|}{|s(w1) \text{ unioned with } s(w2)|}$$

For both metrics, synonyms that share many senses and have few independent senses have a higher value, while synonyms that share few senses in relation to their

independent senses score low. The first metric can identify such situations as when the senses of w1 are a subset of the senses of w2 because R(w1,w2) is equal to 1 in that case.

Discussion

This study demonstrates one advantage of on-line lexicons. It would have been mostly impractible to identify all equivalent words in RIT prior to its automation.

The results suggest that word equivalents are a very small group. But note that many possible sets of equivalents will have been lost from the data because of homography. For example, any equivalents to lead (to guide) will have been corrupted by synonyms of lead (the soft metal)--`lead' and `guide' can never be equivalent because guide will never occur as a synonym to the metal.

Further study will focus on identifying patterns amongst synonym sets such as were found amongst word equivalents--it is possible that all synonyms fall under these suggested or similar 'type' headings. Etymologists will recognize that some of the culprits in the sample were caused by words of common ancestry arriving in the English language at different times or from different sources e.g. Woden and Odin; Caesar and czar (and Kaiser, if you will).

The metrics suggested are inadequate to account for the richness of relationships between synonyms. A more topological or graphical approach might be preferred. Other approachs are suggested by Kozima and Ito (1996) and Harper (1965).

Finally, the dual of synonymy, polysemy should be explored. Equivalent senses--senses described by the same words--may shed some light on this area of semantics.

Bibliography

Kozima, Hideki and Ito, Akira (1996). Context-Sensitive Measurement of Word Distance by Adaptive Scaling of a Semantic Space, Archive at THE COMPUTATION AND LANGUAGE E-PRINT ARCHIVE

Harper, Kenneth E., (1965). Measurement of Similarity between Nouns, in Proceedings of the International Conference on Computational Linguistics.

RIT3: A relational database version of Roget's International Thesaurus. Edited by W. A. Sedelow, S. Y. Sedelow, L. J. Old, University of Arkansas at Little Rock.
