Using Information Content to Evaluate Semantic Similarity in Taxonomy

Section I. Overview

A. Reader Interest

1. Which category describes this manuscript?
   This is a research paper on finding similarity between two concepts in the taxonomy.

B. Content

1. Please explain how this manuscript advances this field of research and/or contributes something new to the literature.
   They tackle the problem of semantic similarity based on the information content the concepts share. They combine the taxonomic structure with the empirical problem estimates which provides better way of adapting knowledge to multiple context concepts. They also try to find a solution to the multiple inheritance problems.

   Their approach is to associate probabilities with taxonomic concepts. They define a probability function for a concept c, based on the probability of encountering an instance of concept c. They claim that information content of a concept can be expressed as $-\log p(c)$ which makes sense.

   Thus, as we go higher in the hierarchy or taxonomy, the information content of concepts decreases. Thus the probability of a super class concept is higher than subclass concept. Thus information content of super class concept is less than that of its subclass. Thus, information content is defined as $-\log p(c)$. Thus, their result would have a common concept that has the max information content and that concept must subsumes both concepts (concepts for which we are trying to find similarity).

   There have been many approaches to find semantic similarity. Approaches like edge distance, finding common ancestors, statistical methods, hybrid approach of combining statistical and knowledge based taxonomic approach. Their approach is fairly new and might have a say in later research.

2. Is the manuscript technically sound?
   Yes.

C. Presentation

1. Are the title, abstract, and keywords appropriate?
   Yes.
   Abstract of the paper seems to be less explanatory and doesn’t explain their approach in detail. They don’t define any keywords.

2. Does the manuscript contain sufficient and appropriate references?
   They have a very detailed section of references of related work and they brief out various approaches other researchers have taken to solve the problem of semantic similarity of concepts in taxonomy and corpus based determination of word similarity. They also give some references of other researchers working on word senses disambiguation.
3. Does the introduction state the objectives of the manuscript in terms that encourage the reader to read on?
No. The paper is relatively dull. It’s been written at a very abstract level and doesn’t really focus on providing a decent introduction.

4. How would you rate the organization of the manuscript? Is it focused? Is the length appropriate for the topic?
Could be improved. The paper is focused. The length of paper might be a little less. I think the author could have explained their approach & introduction much better. He could have separated the motivation from the introduction.

5. Please rate and comment on the readability of this manuscript.
Readable - but requires some effort to understand. This paper is written at a very theoretical level and has been loaded with formulas. I think the introduction should provide enough detail to the user about their approach in the simplest words so that the reader or reviewer can get to know their work with out having to go through other sections.

Section II. Evaluation

Please rate the manuscript. Explain your choice.
Good. I think their approach is more dependent on the text information they have used. They use this information to find concept probabilities which are dependent on the frequency of words set that are consumed by concept c, and frequency is dependent on the noun frequencies in the information text.

Section III. Detailed Comments

As explained in earlier sections this paper isn’t very well formed. The idea seems to be good, never know if this would work if the information text is changed. They conduct testing by comparing their similarity methods with other techniques like edge counting, probability based approach and their approach. Evaluations of these kinds of problems are generally take human subjects consensus. Most of the times it’s based on the opinion of subjects who are trying to find whether given two words were similar. The most interesting thing is that they propose new way of associating word senses while finding semantic similarity.

I think there are many approaches for solving this problem. I think it’s hard to compare different approaches before claiming that a particular approach is good. The authors of this paper seem to have achieved a slightly higher correlation than probability based approach. The approach they have used doesn’t really indicate thoroughly that they had achieved a good correlation measure.