# Measure of Semantic Similarity between Words

Vladimir Jara

# Introduction

- Semantic similarity is a generic issue in a variety of applications in the areas of computational linguistics and artificial intelligence, both in the academic community and industry.

- Examples include word sense disambiguation, detection and correction of word spelling errors, text segmentation, image retrieval, document retrieval, amongst others.

- Similarity between two words is often represented by similarity between concepts associated with the two words.

- Generally, these methods can be categorized into two groups: edge count- ing-based (or dictionary/ thesaurus-based) methods and information theory-based (or corpus-based) methods.

# Semantic Similarity between Words

- There are constraints to the development of similarity measures.

- Semantic similarity is context-dependent and may be asymmetric

- Context

- Similarity between words is influenced by the context in which the words are presented

- For example, if the context is “the outside covering of living objects,” then skin and bark are more similar than skin and hair

- On the other hand, the opposite is true if the context is body parts.

- Asymmetry

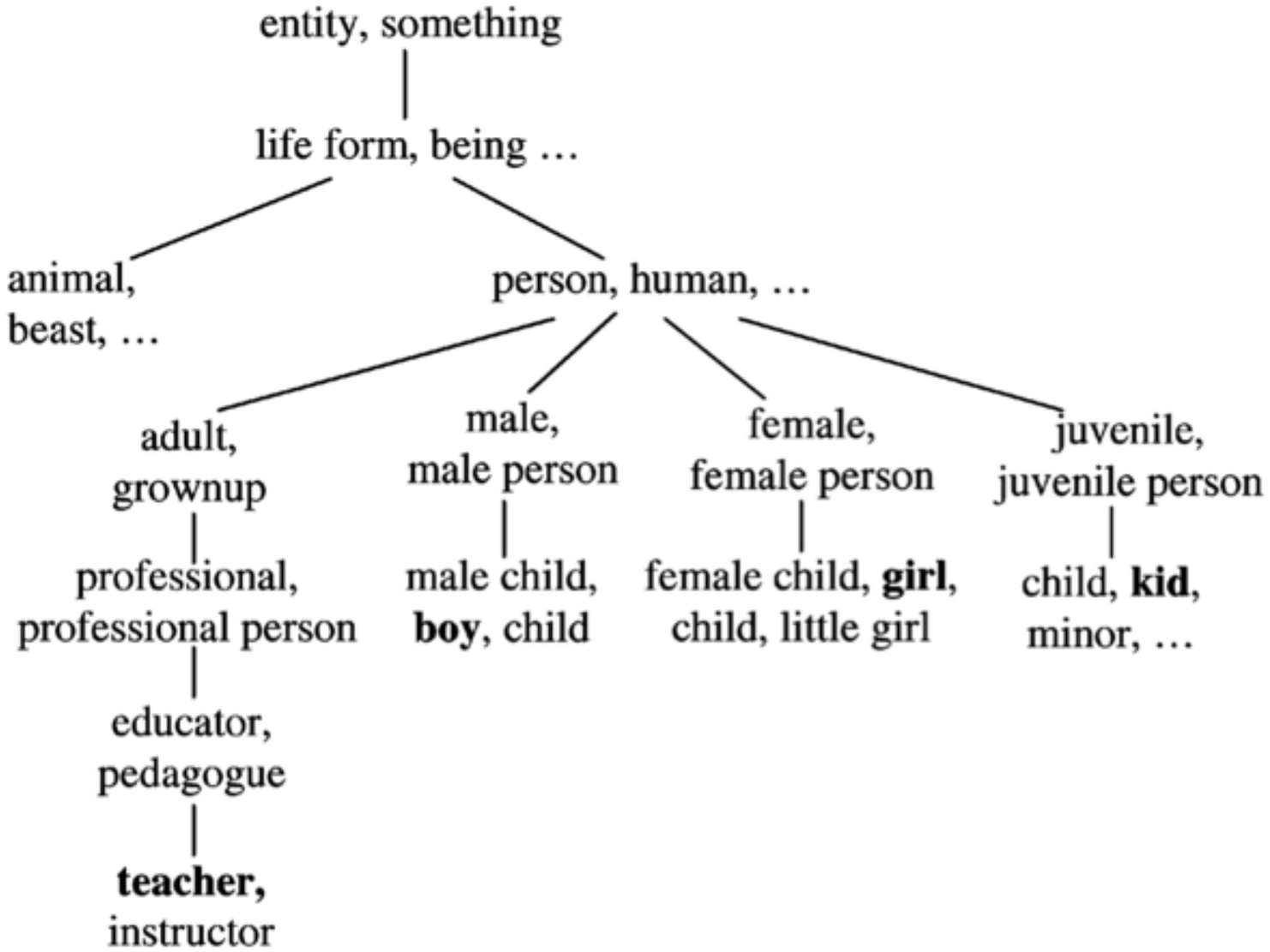
- Similarity may also be asymmetric with respect to direction.

- People may give different ratings when asked to judge the similarity of surgeon to butcher and the similarity of butcher to surgeon.

- Thanks to the success of a number of computational linguistic projects, semantic knowledge bases are readily available.

- The lexical hierarchy is connected by following trails of superordinate terms in “is a” or “is a kind of” (ISA) relations.

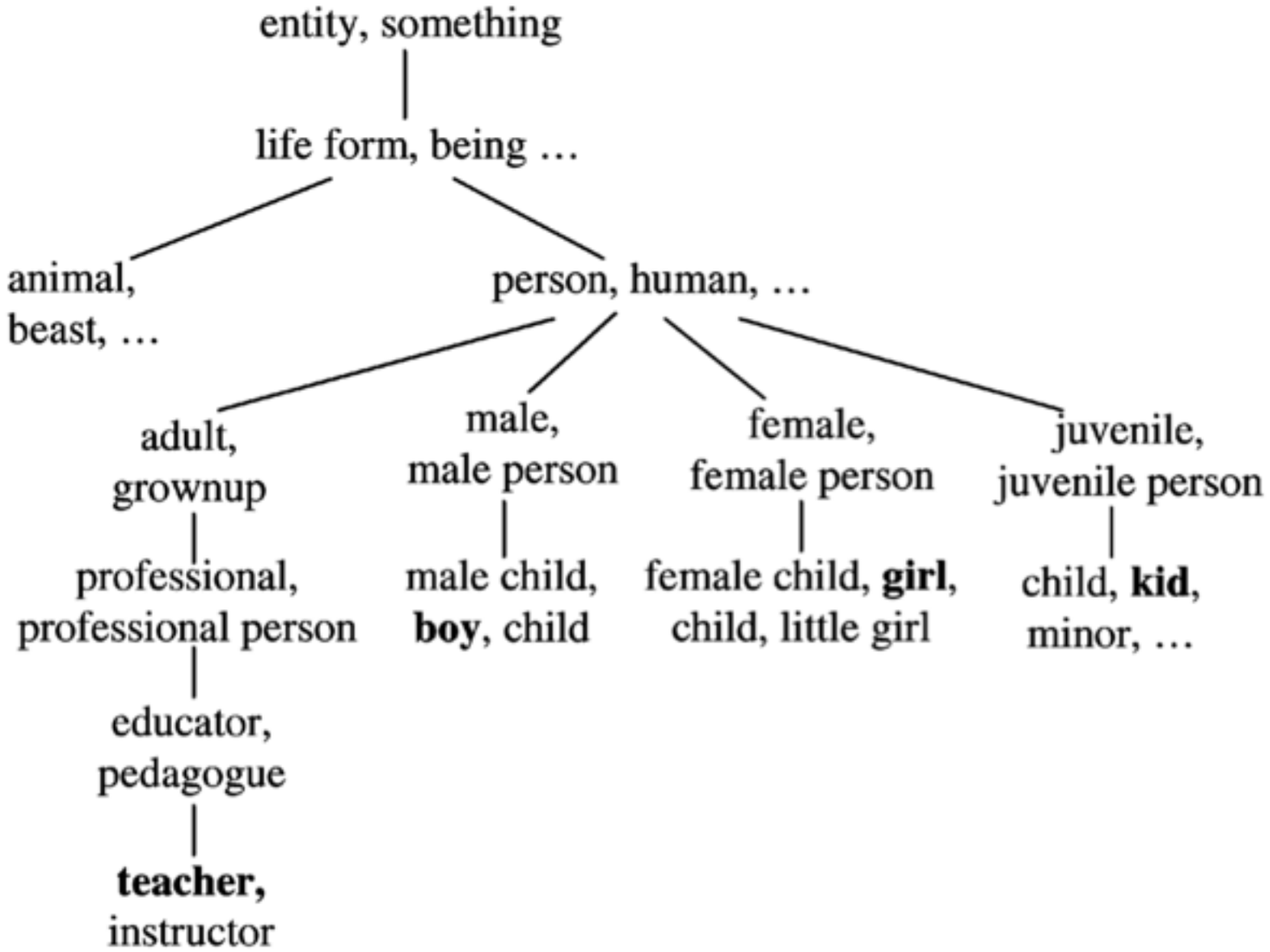
# Wordnet


entity, something
life form, being …
animal,
beast, …
person, human, …
adult,
grownup
male,
male person
female,
female person
juvenile,
juvenile person
professional,
professional person
male child,
boy, child
female child, girl,
child, little girl
child, kid,
minor, …
educator,
pedagogue
teacher,
instructor

- This method of measuring works well on much constrained semantic nets (medical, law).

- However, this method may be not so accurate if it is applied to larger and more general semantic nets such as WordNet

# Wordnet


entity, something
life form, being …
animal,
beast, …
person, human, …
adult,
grownup
male,
male person
female,
female person
juvenile,
juvenile person
professional,
professional person
male child,
boy, child
female child, girl,
child, little girl
child, kid,
minor, …
educator,
pedagogue
teacher,
instructor

- To address this weakness, the direct path length method must be modified by utilizing more information from the hierarchical semantic nets.

- It is intuitive that concepts at upper layers of the hierarchy have more general semantics and less similarity between them, while concepts at lower layers have more concrete semantics and stronger similarity.

- Therefore, the depth of concept in the hierarchy should be taken into account.

# The Benchmark Data Set

- The quality of a computational method for calculating word similarity can only be established by investigating its performance against human common sense.

- In evaluating all methods, it is necessary to compute word similarity on a benchmark word set with human ratings.

- Researchers Rubenstein and Goodenough gave a group of 51 human subjects 65 word pairs and asked the subjects to rate them for similarity in meaning on a scale from 0 (no similarity) to 4 (perfect synonymy).

- Rubenstein-Goodenough's 65 word pairs were divided into two sets: One contains the commonly used 28 word pairs for training, and another contains the remainder, which has 37 word pairs for learning of parameters.

# Shortest path length

- Similarity measure is linear and exclusively based on the shortest path length between the two words.

$$S_1(w_1, w_2) = f_0(l) = 2 \cdot M - l$$

- This strategy does not have any parameters to tune

- We calculate the similarities for word pairs in the test set.

- The correlation coefficient between $S_1$ and human similarity judgments of Rubenstein-Good-enough's was 0.664

# Shortest path length + depth

- Similarity measure is a linear combination of shortest path length and depth.

$$S_2(w_1, w_2) = \alpha S_1(w_1, w_2) + \beta d$$

- This strategy is plausible because the depth of the subtree carries useful information about where the two words possess the same features.

- The higher the subtree is in the semantic hierarchy, the more abstract meaning the two words share and vice versa.

- It is possible to combine this information with the shortest path length in calculating the semantic similarity of words.

- Using the optimal parameters $\alpha=0.05$ and $\beta= 1$, the similarities for word pairs in the test set were calculated.

- The correlation coefficient between this method and human similarity judgments is 0.8315

# Nonlinear shortest path length

- The similarity measure is a nonlinear function of the shortest path length.

$$S_3(w_1, w_2) = f_1(l) = \mathrm{e}^{-\alpha l}.$$

- It is observed that the strongest correlation is reached at α=0.25.

- Using this optimal α, we have that the correlation coefficient between this method and human similarity judgments is 0.8911

- This strategy illustrates that a simple transformation of the shortest path length using a nonlinear function can significantly increase the accuracy of the similarity measure.

# Transferred depth nonlinear function

- Similarity measure is the transferred depth of the subtree through a nonlinear function

$$S_{10}(w_1, w_2) = \frac{\mathrm{e}^{\beta h} - \mathrm{e}^{-\beta h}}{\mathrm{e}^{\beta h} + \mathrm{e}^{-\beta h}}$$

- The strongest correlation against human similarity judgments is at $\beta = 0.15$

- The correlation coefficient between this method and human similarity judgments is 0.8356

# Conclusions

- The similarity measure can be improved by a suitable combination of information sources.

- The similarity measure can be improved by nonlinearly transferring information sources.

- The depth of the subtree is more similar to human ratings than the shortest path length.

# References

* “An Approach for Measuring Semantic Similarity between Words Using Multiple Information Sources”, Yuhua Li, Zuhair A. Bandar, and David McLean

* “Semantic Distance in WordNet: An Experimental, Application-Oriented Evaluation of Five Measures”, A. Budanitsky and G. Hirst

* “A Comparison of WordNet and Roget’s Taxonomy for Measuring Semantic Similarity”, M. McHale

Thank you