Evaluation of semantic similarity measures

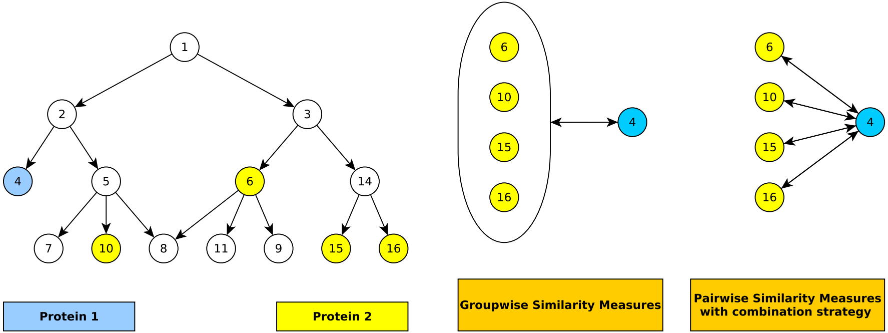
Maxat Kulmanov, Robert Hoehndorf

King Abdullah University of Science and Technology, Saudi Arabia
Computational Bioscience Research Center
Bio-Ontologies Research Group

Semantic Similarity Measures

- Semantic similarity measures capture the strength of interaction between concepts based on their meaning.
- Widely used in bioinformatics
  - Protein-protein interaction identification
  - Gene-Disease associations
  - Patient diagnoses

Semantic Similarity Measures



1
2
3
4
5
6
14
7
10
8
11
9
15
16
Protein 1
Protein 2
6
10
15
16
4
Groupwise Similarity Measures
6
10
15
16
4
Pairwise Similarity Measures with combination strategy

Motivation and Aim

- Large number of semantic similarity measures has been developed:
  - 21 groupwise
  - 38 pairwise with 7 different combination strategies
  - Available in Semantic Measures Library `http://www.semantic-measures-library.org/`
- Classify semantic similarity measures by their sensitivity to the:
  - number of annotated classes
  - difference of the number of annotated classes

Materials

- Gene Ontology (GO)
- 6,108 gene annotations from Yeast Genome Database. Annotation sizes vary from 1 to 55
- 5,500 randomly generated annotations
    - 55 groups with 100 genes in each:
        - 1st group annotated with 1 GO class
        - 2nd group annotated with 2 GO classes
        - 3rd group annotated with 3 GO classes
        - and so on

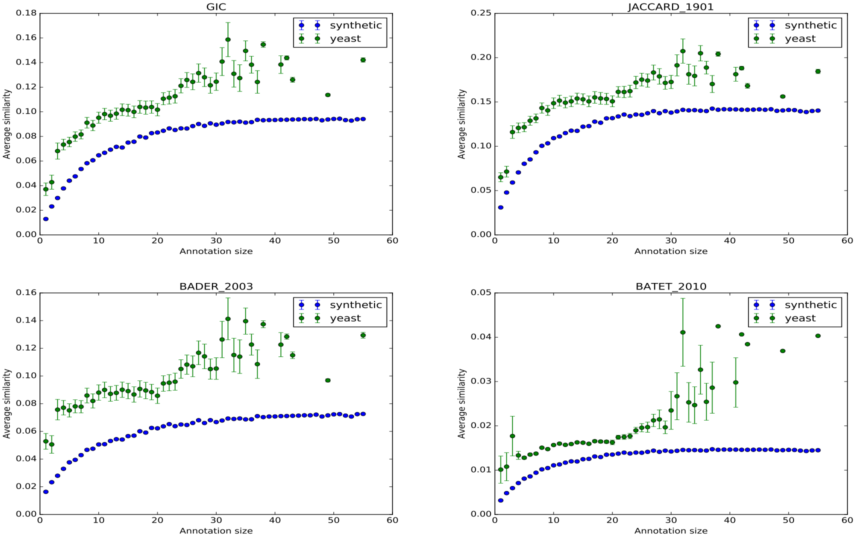
Methods

- Compute similarity between each pair of genes
  - 18,656,886 similarity values for yeast annotations ((6108 + 1) / 2 * 6108)
  - 15,127,750 similarity values for random annotations
- Group similarities by annotations size
- Group similarities by annotations size difference
- Take average similarities for all groups

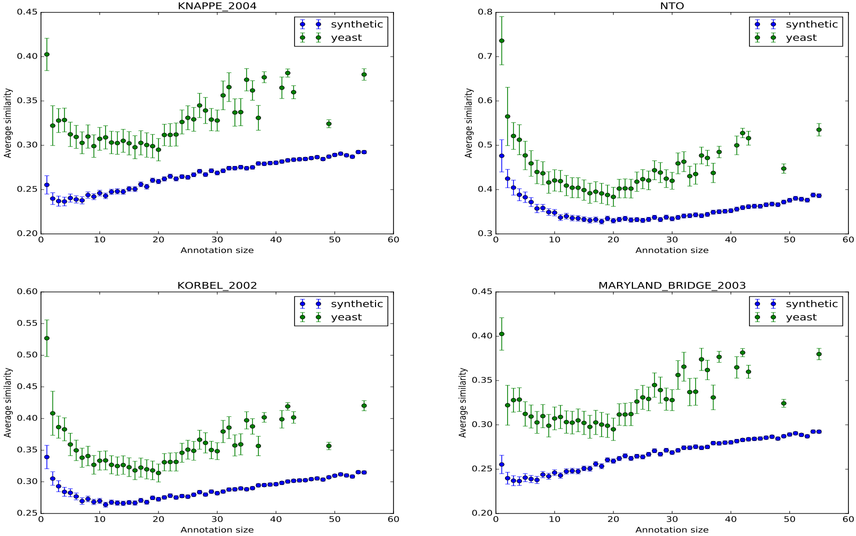
Results

- Sensitive
  - Similarity value increases when annotation size (difference) increases
  - Similarity value decreases when annotations size (difference) increases
- Not sensitive

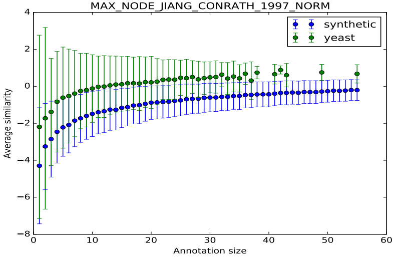
Annotation size - Groupwise Similarity Measures

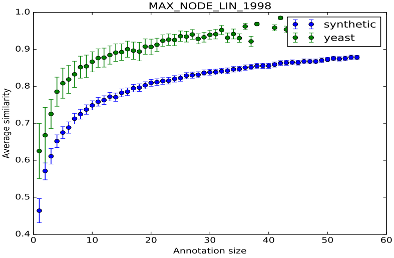


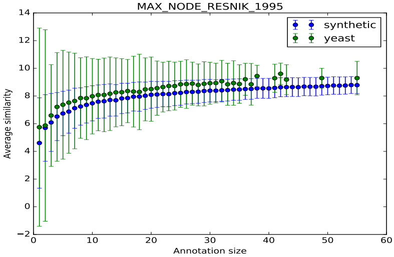
Annotation size - Groupwise Similarity Measures

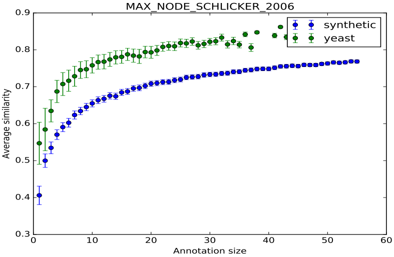


Annotation size - Pairwise Similarity Measures

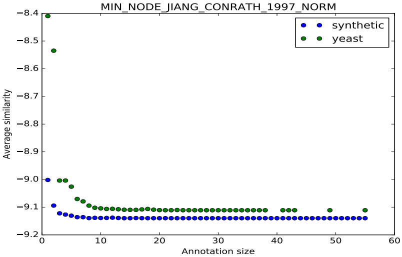


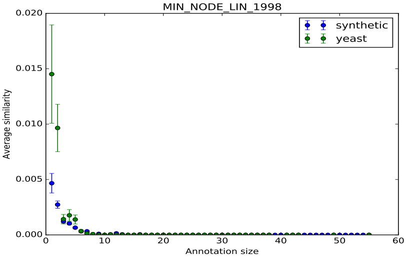


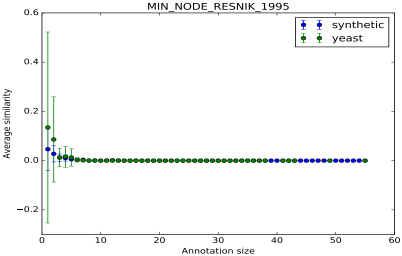


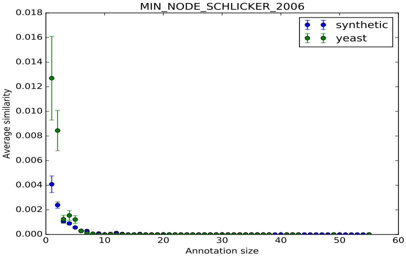


Annotation size - Pairwise Similarity Measures

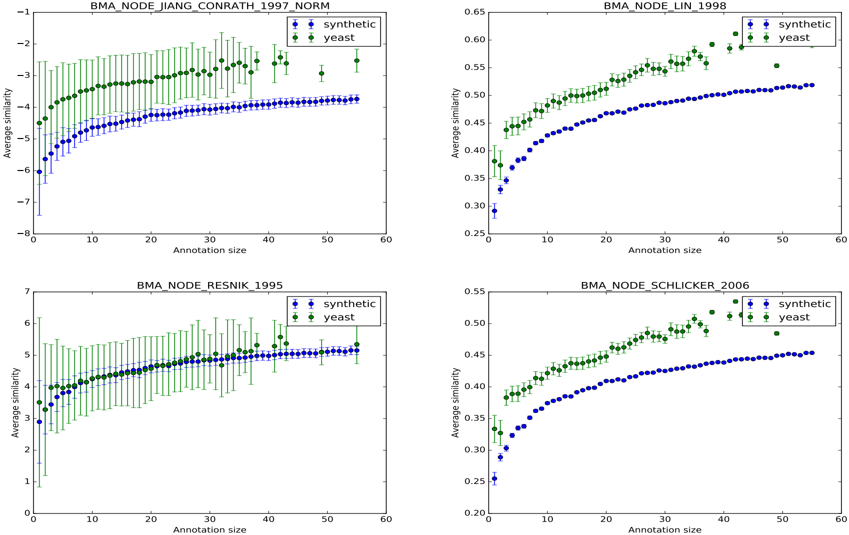




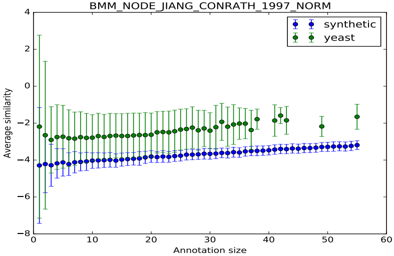


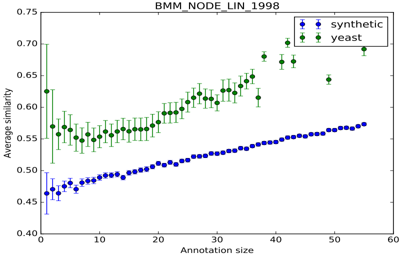


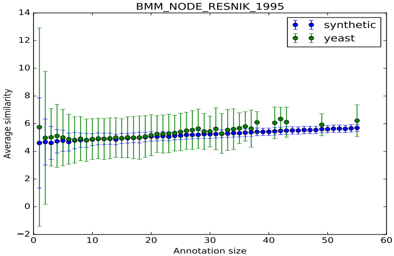
Annotation size - Pairwise Similarity Measures

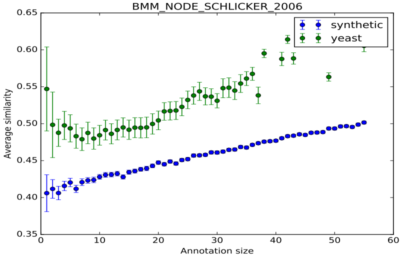


Annotation size - Pairwise Similarity Measures

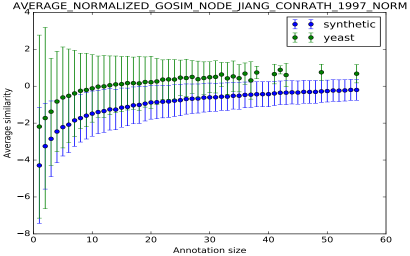


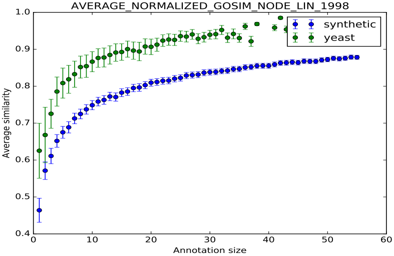


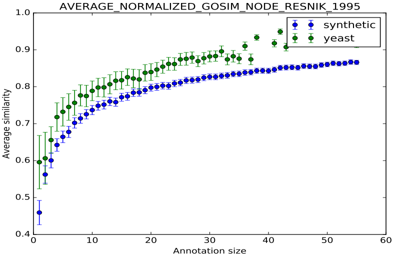


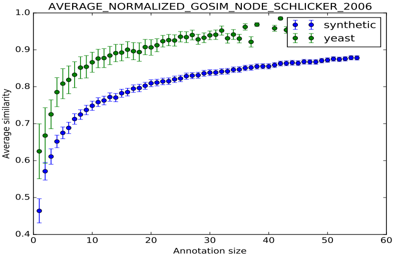


Annotation size - Pairwise Similarity Measures

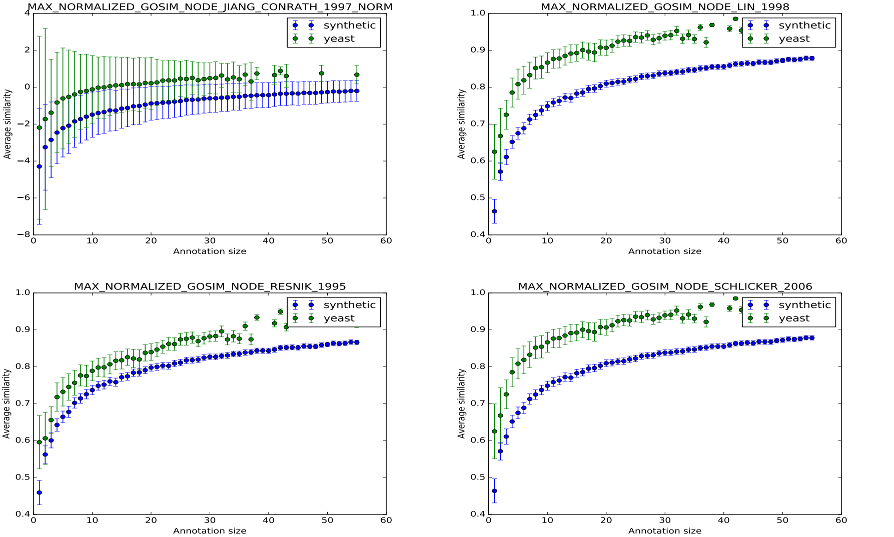




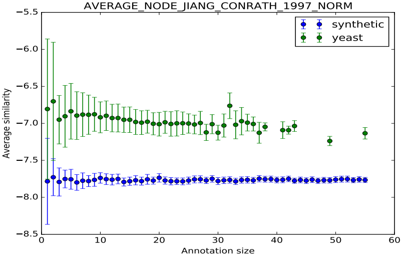


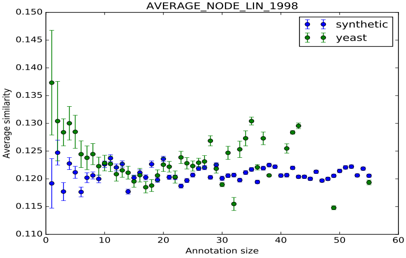


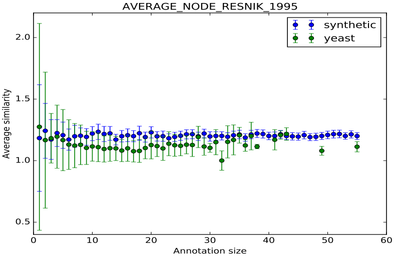
Annotation size - Pairwise Similarity Measures

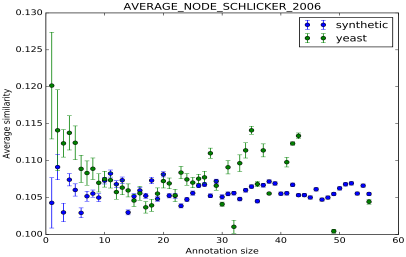


Annotation size - Pairwise Similarity Measures

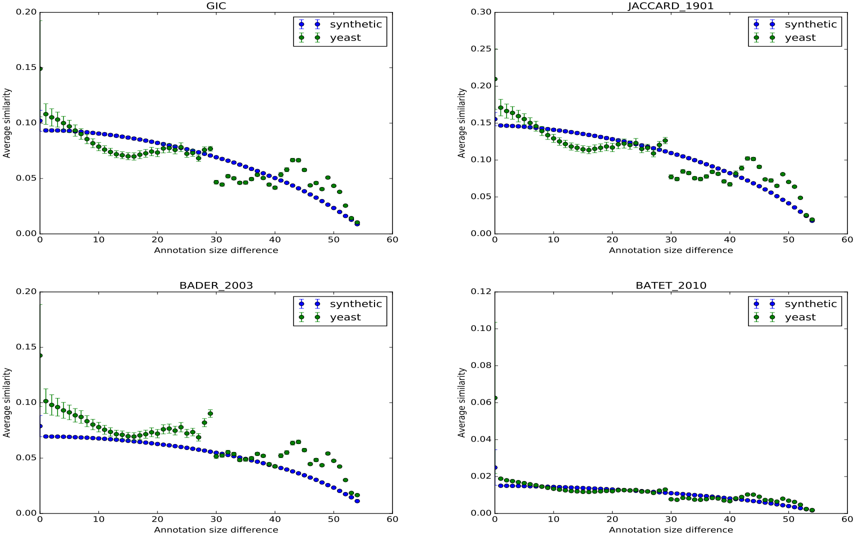




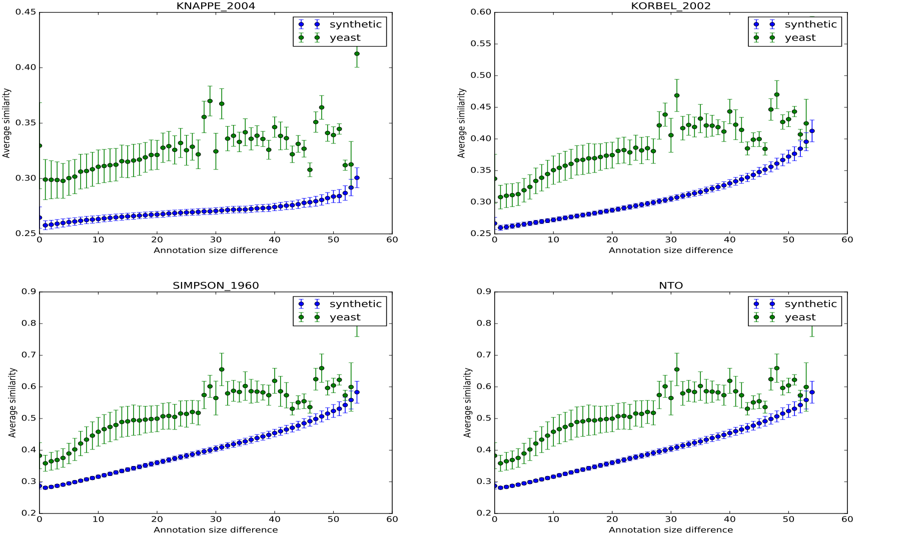




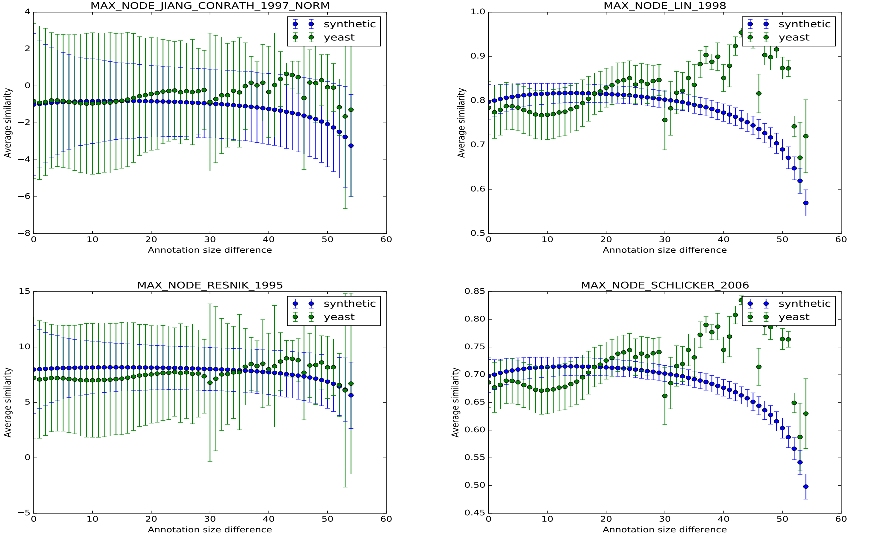
Annotation size difference - Groupwise Similarity Measures



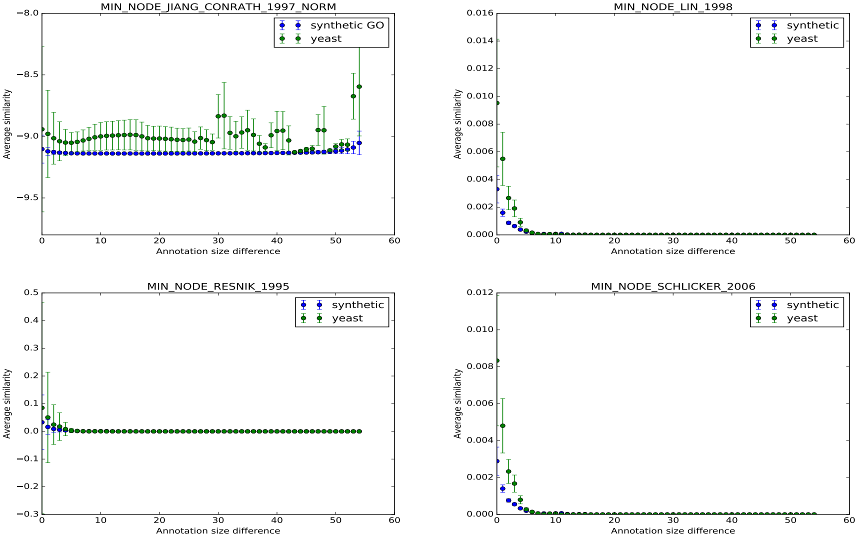
Annotation size difference - Groupwise Similarity Measures



Annotation size difference - Pairwise Similarity Measures



Annotation size difference - Pairwise Similarity Measures

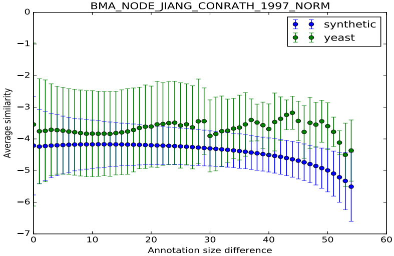

MIN_NODE_JIANG_CONRATH_1997_NORM

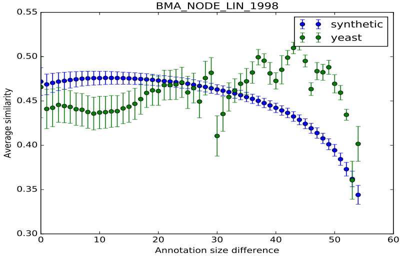
MIN_NODE_LIN_1998

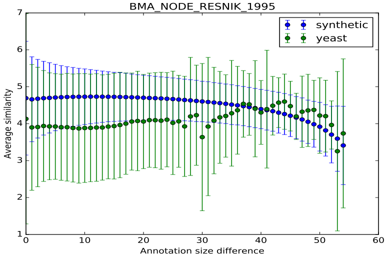
MIN_NODE_RESNIK_1995

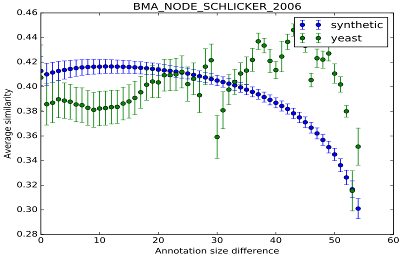
MIN_NODE_SCHLICKER_2006

Annotation size difference - Pairwise Similarity Measures

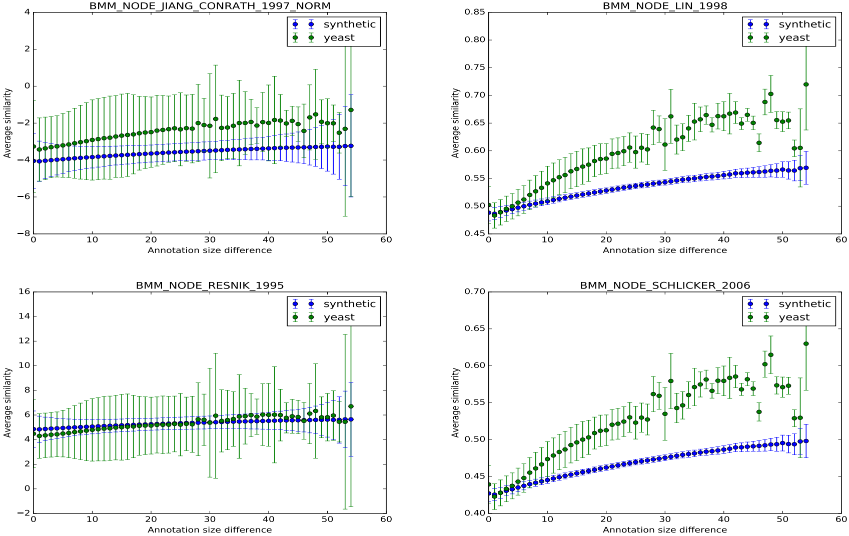




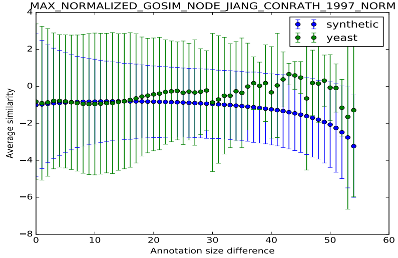




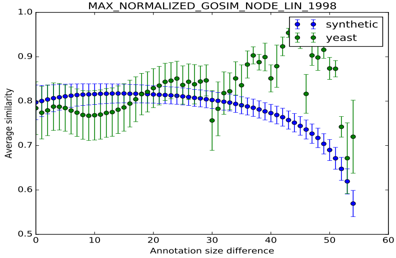
Annotation size difference - Pairwise Similarity Measures



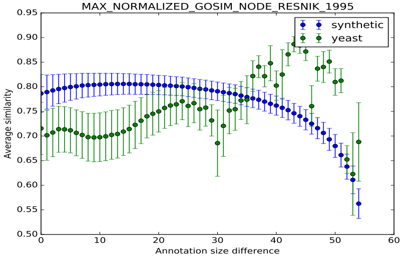
Annotation size difference - Pairwise Similarity Measures



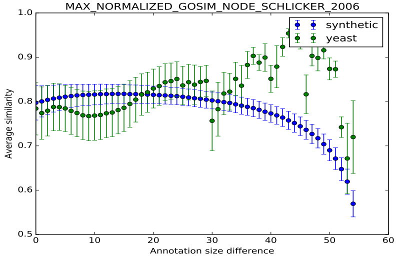
MAX_NORMALIZED_GOSIM_NODE_JIANG_CONRATH_1997_NORM



MAX_NORMALIZED_GOSIM_NODE_LIN_1998

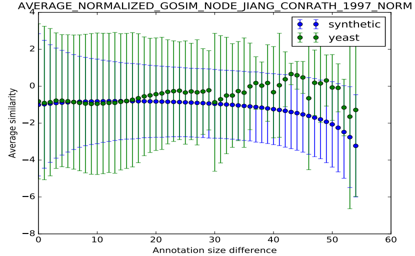


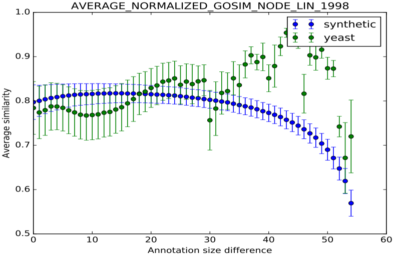
MAX_NORMALIZED_GOSIM_NODE_RESNIK_1995

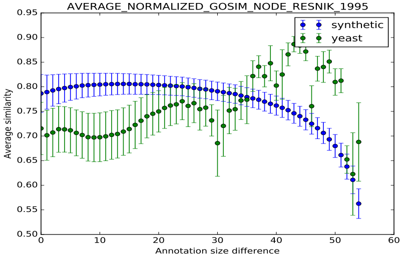


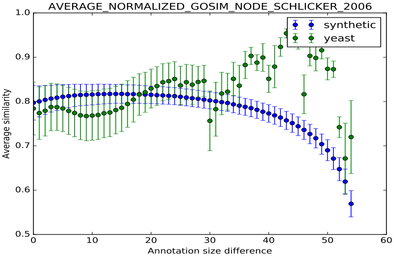
MAX_NORMALIZED_GOSIM_NODE_SCHLICKER_2006

Annotation size difference - Pairwise Similarity Measures

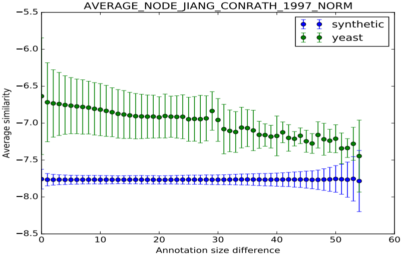


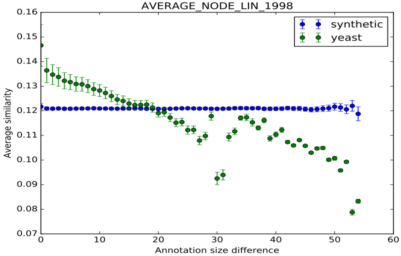


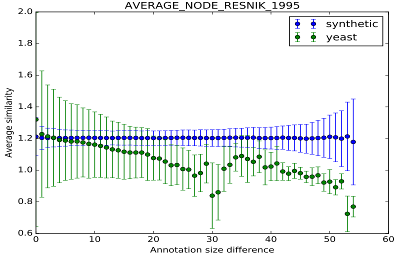


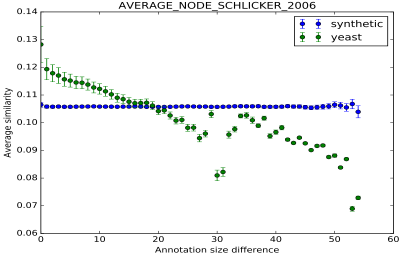


Annotation size difference - Pairwise Similarity Measures


AVERAGE_NODE_JIANG_CONRATH_1997_NORM


AVERAGE_NODE_LIN_1998


AVERAGE_NODE_RESNIK_1995


AVERAGE_NODE_SCHLICKER_2006

Summary

- Most of the similarity measures are sensitive to the annotation size
- Pairwise measures depend on combination
- Well annotated entities get higher similarities
- Studies which use similarity measures may be biased by annotation size

Recommendations

- If annotations size variance is high, use pairwise similarity measures with average strategy

Thanks!

http://www.cbrc.kaust.edu.sa/onto/sim-eval/