Evaluating the Stability and Credibility of Ontology Matching Methods

Xing Niu, Haofen Wang, Gang Wu, Guilin Qi, and Yong Yu

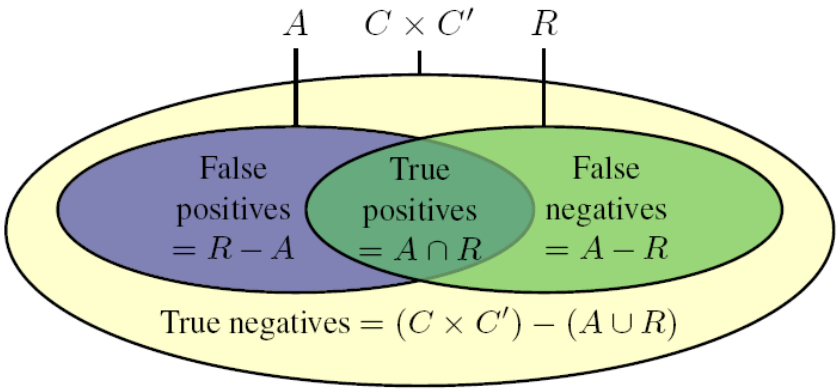
2011.5.31

Agenda

- Introduction
- Basic Concepts
  - *Confidence Threshold*
  - *Test Unit*
- Evaluation Measures and Their Usages
  - *Comprehensive F-measure*
  - *STD score*
  - *ROC-AUC score*
- Discussion and Future Work

Introduction

- Stability
  - Reference matches are scarce
  - Training proper parameters (e.g. confidence threshold)
  - High **stability**: a matching method performs consistently with the trained parameter on the data of different domains or scales
- Credibility



  - Candidate matches sorted by their matching confidence values
  - High **credibility**: a matching method generate true positive matches with high confidence values while return false positive ones with low values.

Introduction (con't)

- Judging the basic compliance is not sufficient
  - Precision
  - Recall
  - F-measure

- Measurements
  - Stability: Comprehensive F-measure and STD (STandard Deviation) score
  - Credibility: ROC-AUC (Area Under Curve) score

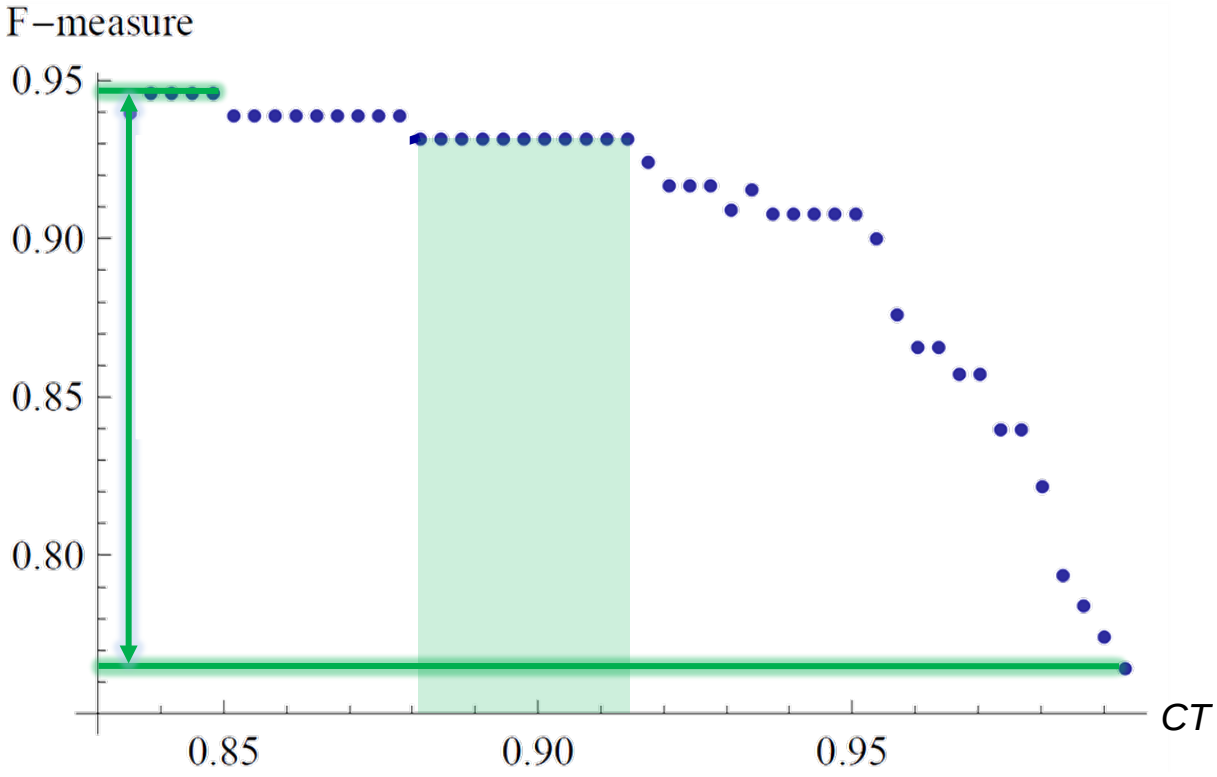
Agenda

- *Introduction*
- Basic Concepts
  - *Confidence Threshold*
  - *Test Unit*
- Evaluation Measures and Their Usages
  - *Comprehensive F-measure*
  - *STD score*
  - *ROC-AUC score*
- Discussion and Future Work

Confidence Threshold

- A match can be represented as a 5-tuple
  - $< id, e_1, e_2, r, n >$
  - $< m1, person, human, =, 0.9 >$

- Confidence Threshold (CT)



Test Unit

- Test Unit
  - A set of similar datasets describing the same domain and sharing many resemblances, but differing on details.

- Examples
  - **Benchmark 20X:** structural information remains • another language, synonyms, naming conventions
  - **Conference Track:** conference organization domain • built by different groups
  - **Cyclopedia:** same data source • different categories
  - ...
  - **Others:** same data source • random N-folds

Agenda

- *Introduction*
- *Basic Concepts*
  - *Confidence Threshold*
  - *Test Unit*
- Evaluation Measures and Their Usages
  - *Comprehensive F-measure*
  - *STD score*
  - *ROC-AUC score*
- Discussion and Future Work

Comprehensive F-measure

- Maximum F-measure
  - maxF-measure reflects the theoretical optimal matching quality of matching methods

$$maxF\text{-}measure = \max F\text{-}measure(CT)$$

- Uniform F-measure
  - uniF-measure simulates the practical application and evaluates such stability of matching methods.

$$maxFCT = \arg\max_{CT} F\text{-}measure(CT)$$

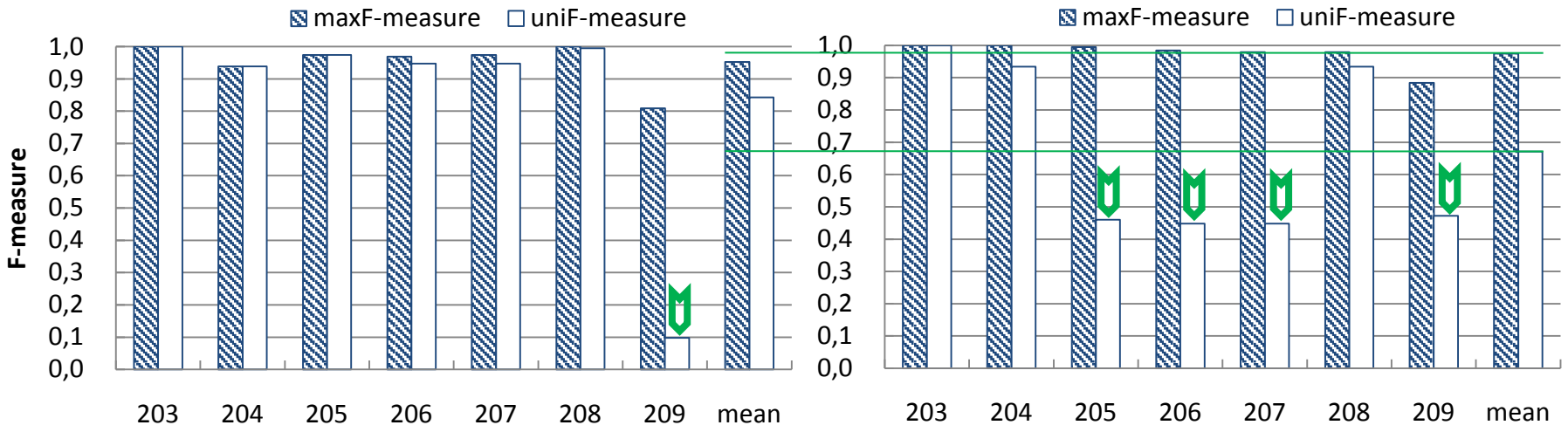
$$uniF\text{-}measure = F\text{-}measure(\text{average}(maxFCTs))$$

- Comprehensive F-measure

$$comF\text{-}measure = (maxF\text{-}measure + uniF\text{-}measure)/2$$

Usage of comF-measure

- Compare mapping systems according to uniF-measure.



Falcon-AO and RiMOM in Benchmark-20X Test

*another language, synonyms

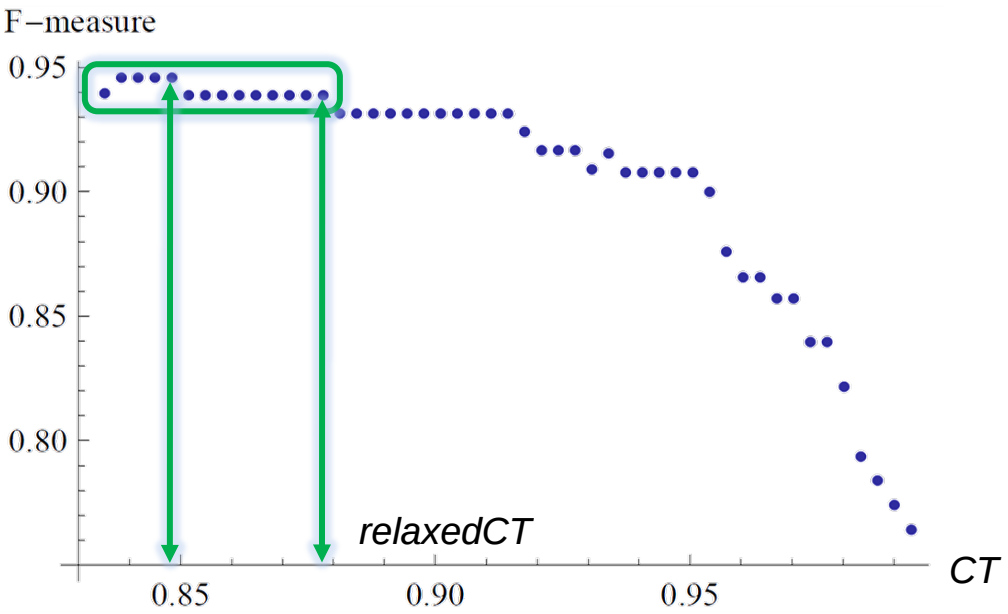
Usage of comF-measure (con't)

- Use the comF-measure value
  - as a indicator of matching quality
    - because it reflects both theoretical and practical results
- Use the comF-measure function
  - as the objective function of a optimization problem
    - because it conceals a multi-objective optimization problem

STD Score

Relaxed CT

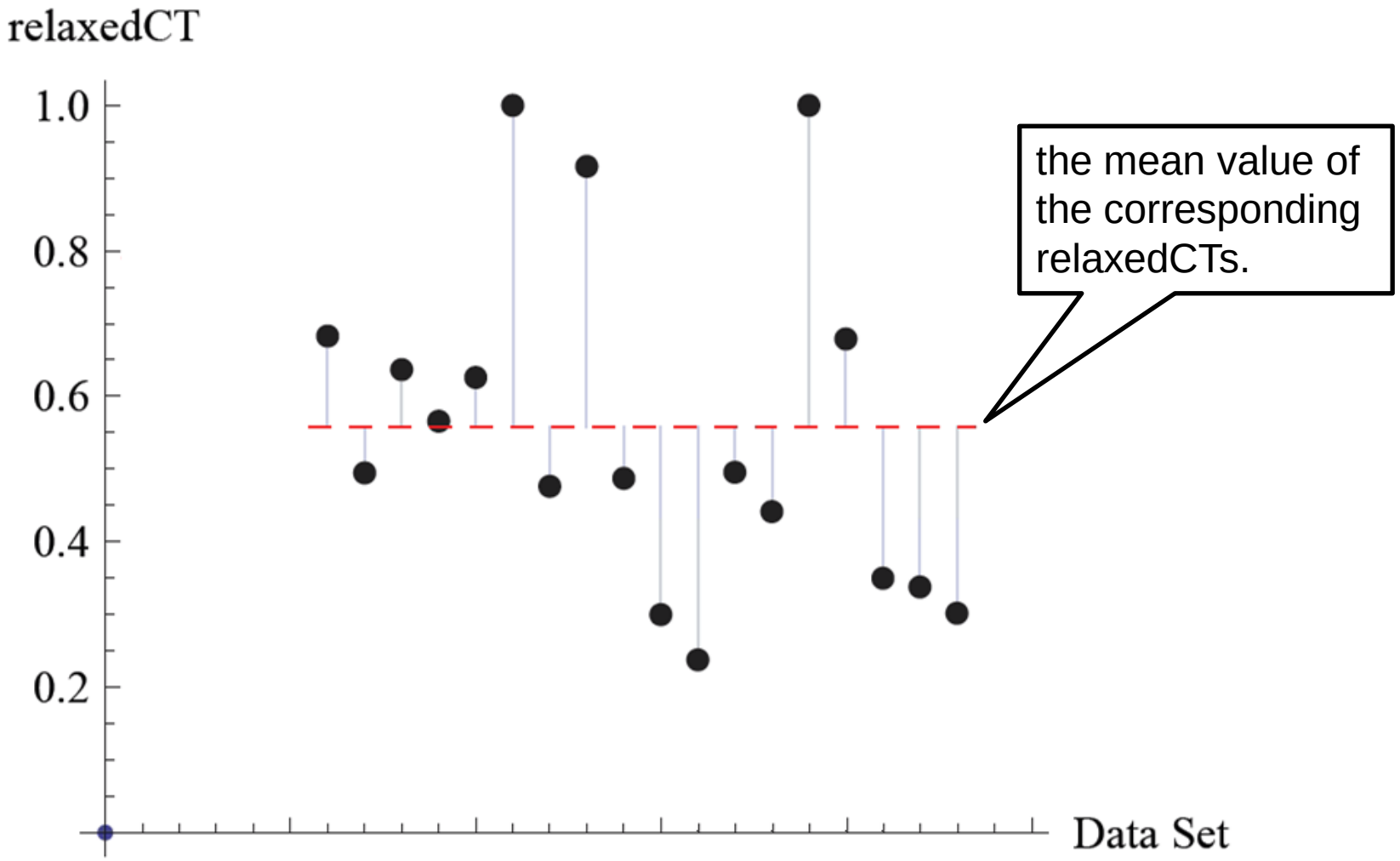
- Only accepting best performance is somehow too strict.
- Accept those CT values that cause a matching method to have F-measure close to maximum F-measure.



STD Score: a standard deviation to measure the dispersion of relaxed CT.

$$STD\,score = 1 - \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(relaxedCT_i - \overline{relaxedCT}\right)^2}$$

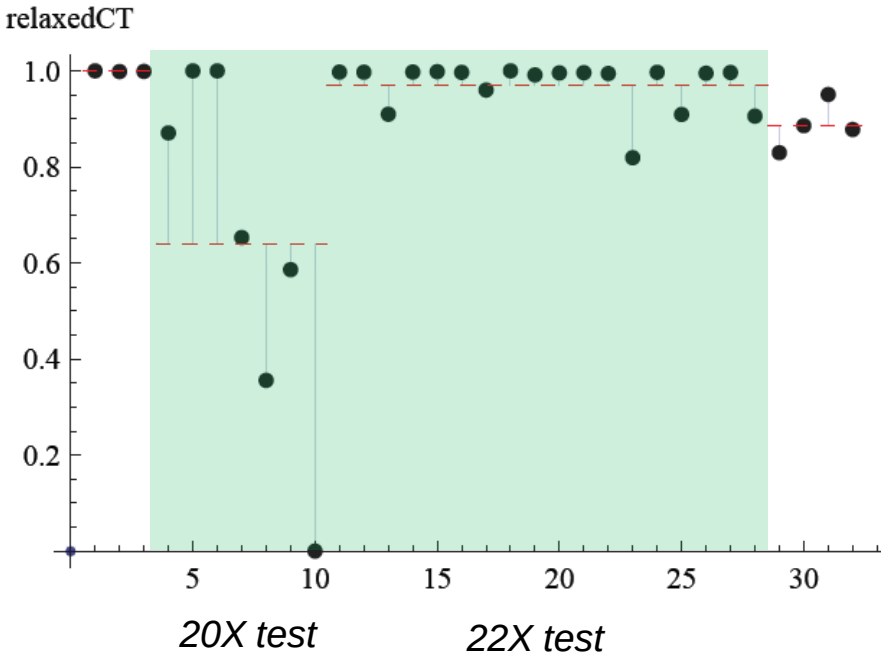
STD Score

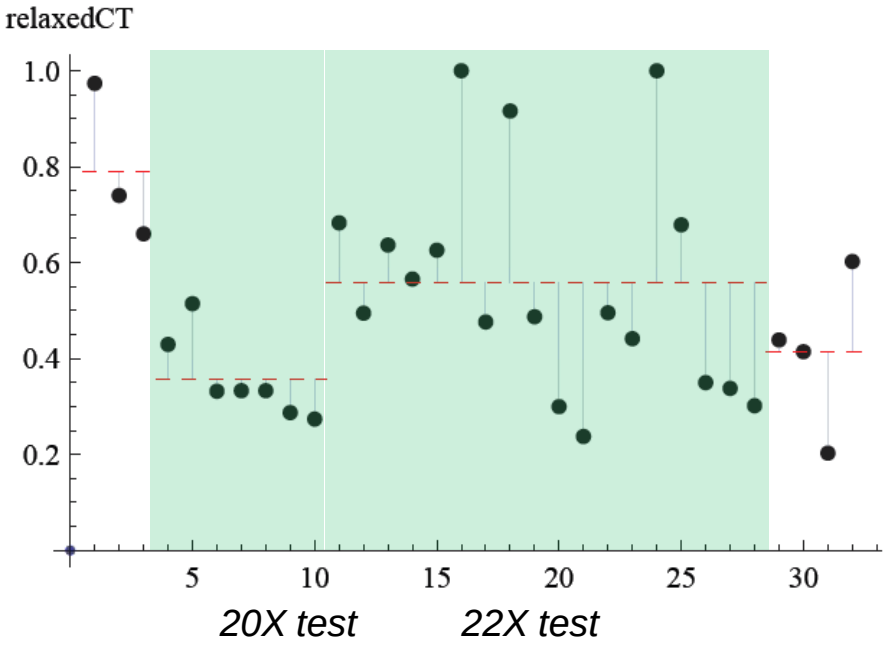


Note: Measuring the dispersion of relaxedFCTs can help us to estimate the difficulty in obtaining maxF-measure for a given matching method.

Usage of STD Score

- Reflects the stability of a matching method





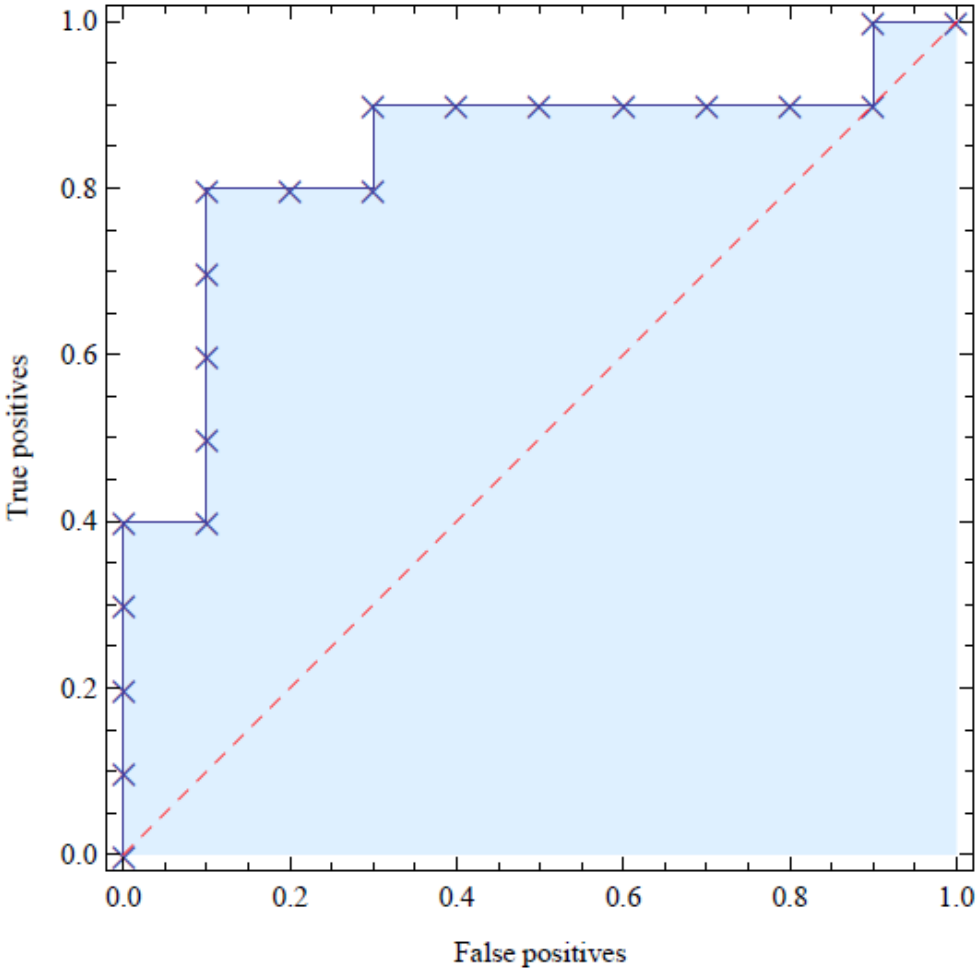
Falcon-AO & Lily in Benchmark Test

*lexical information, structural information

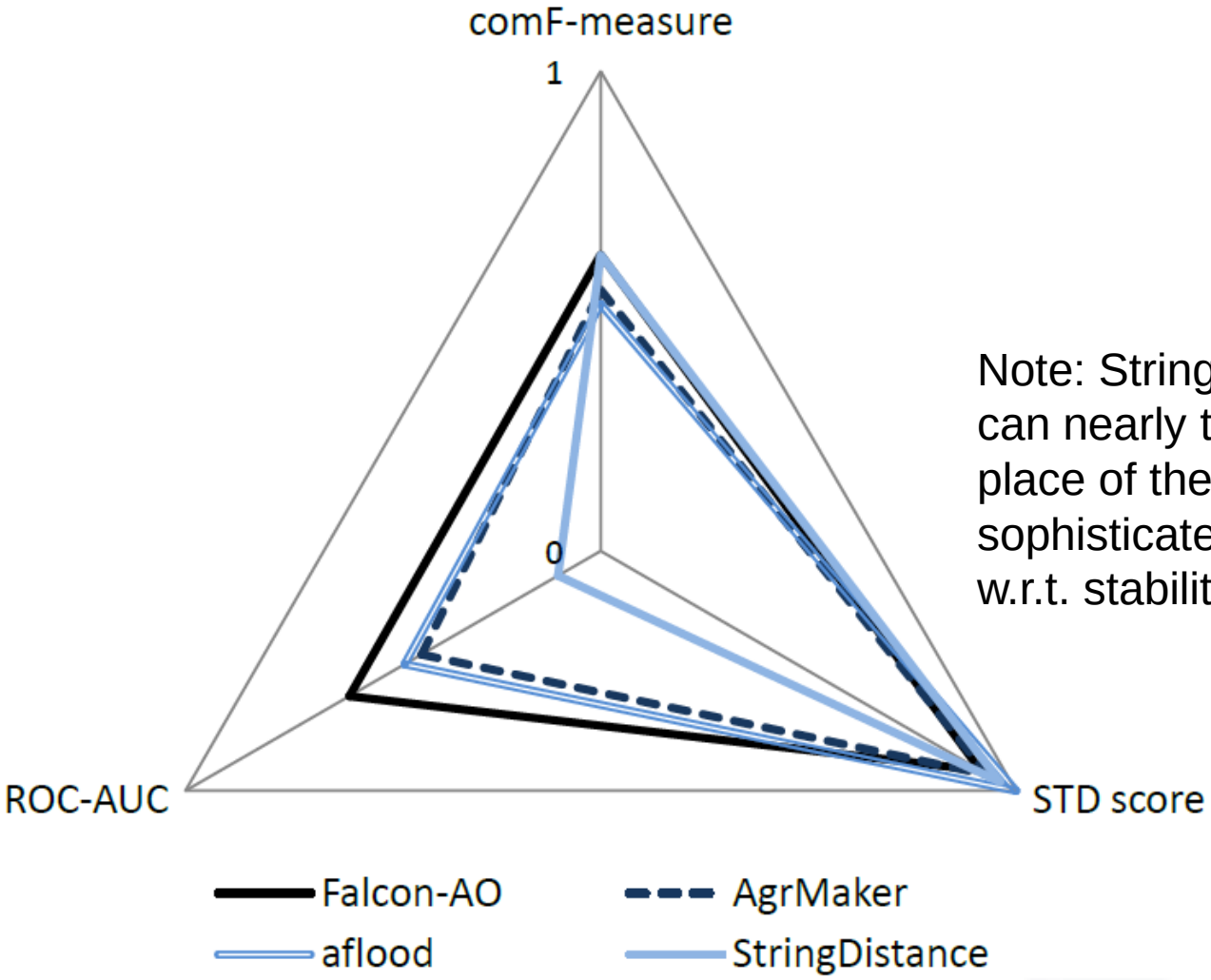
ROC-AUC score

ROC-AUC is the area under ROC curve.

- Reorder matches in descending order by their confidence values.
- Starting from the match with the highest confidence value, if it is true positive, the curve climbs up by one unit distance; otherwise, the curve move horizontally to the right by one unit distance.



Usage of ROC-AUC Score



Note: StringDistance can nearly take the place of the sophisticated ones w.r.t. stability.

Spider Chart for Conference Test Unit

Agenda

- *Introduction*
- *Basic Concepts*
  - *Confidence Threshold*
  - *Test Unit*
- *Evaluation Measures and Their Usages*
  - *Comprehensive F-measure*
  - *STD score*
  - *ROC-AUC score*
- Discussion and Future Work

Discussion and Future Work

- Discussion
  - New evaluation measures
    - *Comprehensive F-measure*
    - *STD score*
    - *ROC-AUC score*
  - Deep analysis
    - Find the potential weakness of these methods and help researchers to improve them to a certain extent.
- Future Work
  - Extend our evaluation measures to a comprehensive strategy
  - Test more matching methods under other datasets
  - Make both stability and credibility as standard evaluation measures for ontology matching

Thanks!