**Entity Resolution with Markov Logic**

**Summary: -** Entity Resolution (ER) refers to the task of identifying whether two records are same or not. Designing robust algorithm for entity resolution is one of the most challenging tasks in text analytics. In this paper [1], the authors propose an approach to match entities (e.g. paper, author, and venue) using First Order Logic (FOL) & Markov Random Fields (MRF) which is a generalization of Felligi-Sunter model [2].

**Problem Statement: -** Each entity (e.g. paper, author, and venue) of the real world (considering citation database here) is represented by one or more strings. For a given pair of entities (x1, x2), the goal is to identify whether they are identical or not? (Is x1=x2?)

**Existing Models: -**

**Field Comparison:** Basically, fields that have common word are more likely to be same *HasWord(field, word)* which is true iff field contains word;*∀x1, x2, y1, y2 HasWord(x1, y1) ∧ HasWord(x2, y2) ∧ y1 = y2 ⇒ x1 = x2* [*Disadvantage:* Treats misspellings, variant spellings, abbreviations of a word as totally dissimilar words]

**Felligi-Sunter model** [2] (FSM): ER is treated as a classification problem (match/non-match) based on vectors of similarity scores between features of entities; It uses Naïve Bayes (NB) to predict whether two entities are same using filed comparison as attributes [*Disadvantage:* Treats entities as independently distributed, ignoring interdependencies]

**Relational model**: Combination of FSM and transitivity with field matches are treated as features ∀x, y1, y2 HasAuthor(x, y1) ∧ HasAuthor(x, y2) ⇒ Coauthor(y1, y2)

**Experiment:-**

**Datasets:** Cora [3] and BibServ.org

**Proposed Approach/Model:** FOL sets probability of a word to zero if it violates any one formula of the Knowledge Base. MRF softens this restriction by reducing it without setting it to zero. So, a combination of FOL & MRF is used. Various approaches like NB, Markov Logic Network (MLN) and its variant are tried.

**Results & Discussions: -**  The conditional loglikelihood (CLL) and area under the precision-recall curve (AUC) for the match predicates are measured for each approach. In case of Venues of Cora [3], the AUC seem to increase monotonically with increase in features. For BibServ.org, MLN (B+N+C+T) is the best performing model for citations & authors. The best two performers as per AUC in venues are: MLN (G+C+T) & MLN (B). Overall, it is concluded that for both the datasets the AUC increases on increasing number of features except for Venue.

**How can this be applied by The Times Group practically?**

1) Identifying Companies with variations in their names

2) Identifying whether if a user has applied for a job multiple times using slightly different names each time

3) Cluster jobs of the same class together

4) … and many others!

**References**

[1] P Singla, P Domingos, *Entity Resolution with Markov Logic*, ICDM 2006

[2] I. Fellegi and A. Sunter. A theory for record linkage. J. American Statistical Association, 64:1183–1210, 1969.

[3] www.cs.umass.edu/∼mccallum/data/cora-refs.tar.gz