Evaluating Evaluation Metrics for Ontology-Based Applications

Diana Maynard, Wim Peters, Yaoyong Li

University of Sheffield, UK

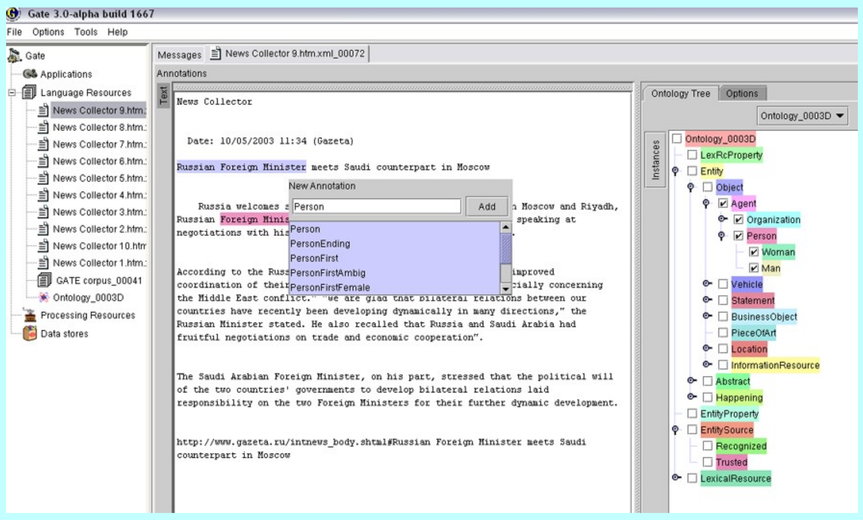
Motivation

Traditionally, Precision and Recall are used to evaluate IE systems, which gives a binary score for each entity found.

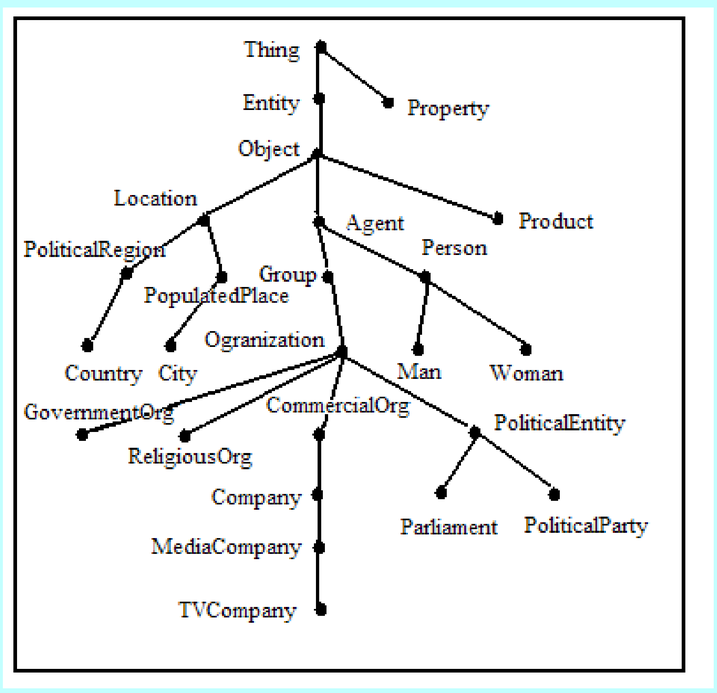
For ontology-based applications, this is insufficient because we want a more flexible measure that takes into account the degree of correctness of the result. Classifying a Man as a Person is less wrong than classifying a Man as a Location (see Figure).

We adopt an approach based on similarity between Key and Response, known as **BDM** (Balanced Distance Metric).

Aim to evaluate how useful the BDM is as a metric for ontology-based IE.



Text annotated in GATE according to KIMO ontology



Subset of Proton ontology

Guidelines for evaluation metrics

A metric should:

- Reach its highest value for perfect quality
- Reach its lowest value for worst quality
- Be monotonic
- Be clear and intuitive
- Correlate well with human judgement
- Be reliable and exhibit as little variance as possible
- Be cheap to set up and apply
- Be automatic

BDM measure

$$BDM = \frac{BR(CP/n1)}{BR(CP/n1) + (DPK/n2) + (DPR/n3)}$$

CP = shortest length from root to MSCA

DPK = shortest length from MSCA to Key

DPR = shortest length from MSCA to Response

n1 = av. chain length of all chains containing K and R

n2 = av. chain length of all chains containing K

n3 = av. chain length of all chains containing R

BR = branching factor of each relevant concept, divided by av. branching factor of all nodes excluding leaf nodes

Results

No.	Entity	Key	Response	BDM	LA
1	Sochi	City	Location	0.724	1.0
2	Federal Bureau of Investigation	Government Organisation	Organisation	0.959	1.0
3	Al-Jazeera	TV Company	Organisation	0.783	1.0
4	Islamic Jihad	Religious Organisation	Company	0.816	0.556
5	Brazil	Country	Object	0.587	1.0
6	Senate	Political Entity	Company	0.826	0.556
7	Kelly Ripa	Person	Man	0.690	0.667

Findings

Binary decisions are not sufficient for ontology evaluation involving hierarchical information

Both BDM and Learning Accuracy (LA) perform better than traditional metrics

BDM gives a better error analysis than LA in some situations

BDM is robust when dealing with different ontology sizes and densities

BDM enables better distinctions between some kinds of IE system (minor misclassifications less heavily penalised)