
methodology for the comparison of human judgments with metrics for coreference resolution

Mariya Borovikova^{*1}, Loïc Grobol^{†‡1,2}, Anaïc Lefeuvre-Halftermeyer¹, and Sylvie Billot¹

¹Laboratoire d'Informatique Fondamentale d'Orléans – Université d'Orléans : EA4022, Institut National des Sciences Appliquées - Centre Val de Loire : EA4022, Institut National des Sciences Appliquées – France

²Modèles, Dynamiques, Corpus – Université Paris Nanterre : UMR7114, Centre National de la Recherche Scientifique : UMR7114 – France

Résumé

We propose a method for investigating the interpretability of metrics used for the coreference resolution task through comparisons with human judgments. We provide a corpus with annotations of different error types and human evaluations of their gravity. Our preliminary analysis shows that metrics considerably overlook several error types and overlook errors in general in comparison to humans. This study is conducted on French texts, but the methodology is language-independent.

Mots-Clés: coréférence, évaluation, interprétabilité

*Auteur correspondant: mariya.borovikova@sorbonne-nouvelle.fr

†Intervenant

‡Auteur correspondant: loic.grobol@gmail.com