

Studienarbeiten zur internetbasierten Kommunikation mit KobRA-Verfahren an der Universität Mannheim

Institutionelle Einbindung

Projektmodul im Masterstudiengang *Sprache und Kommunikation*

- Studiengang der Philosophischen Fakultät unter Beteiligung der Lehrstühle Anglistik, Germanistik und Romanistik
- Eigenständige Durchführung einer zweisemestrigen Studienarbeit

Zwei Studienarbeiten im Themenfeld der internetbasierten Kommunikation (IBK)

- Laufzeit: HWS 2014 bis FSS 2015
- Betreuung: Harald Lungen luengen@ids-mannheim.de
Christian Pölit christian.poelitz@cs.uni-dortmund.de
Angelika Storrer astorrer@mail.uni-mannheim.de
Tassja Weber tasweber@mail.uni-mannheim.de

Datengrundlage beider Studienarbeiten

- Diskussionsseiten der deutschsprachigen Wikipedia aus dem Jahr 2013 (264 Mio. Tokens/550.000 Diskussionsseiten), verfügbar über das IDS Mannheim: <http://www1.ids-mannheim.de/kl/projekte/korpora/verfuegbarkeit.html>
- Werkzeug zur Datenerhebung: RapidMiner mit KobRA-Plugin
- Auswertung mit Excel

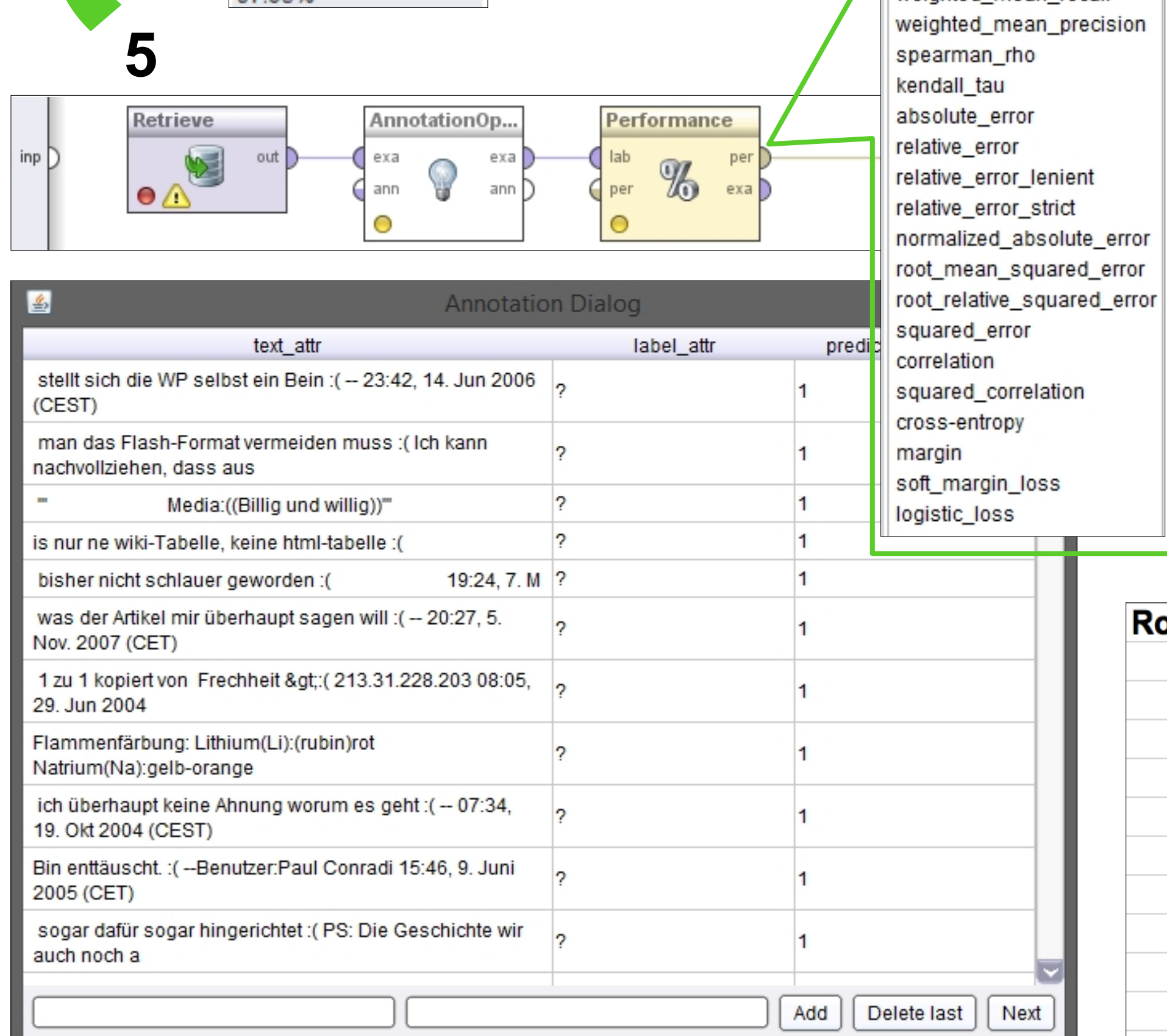
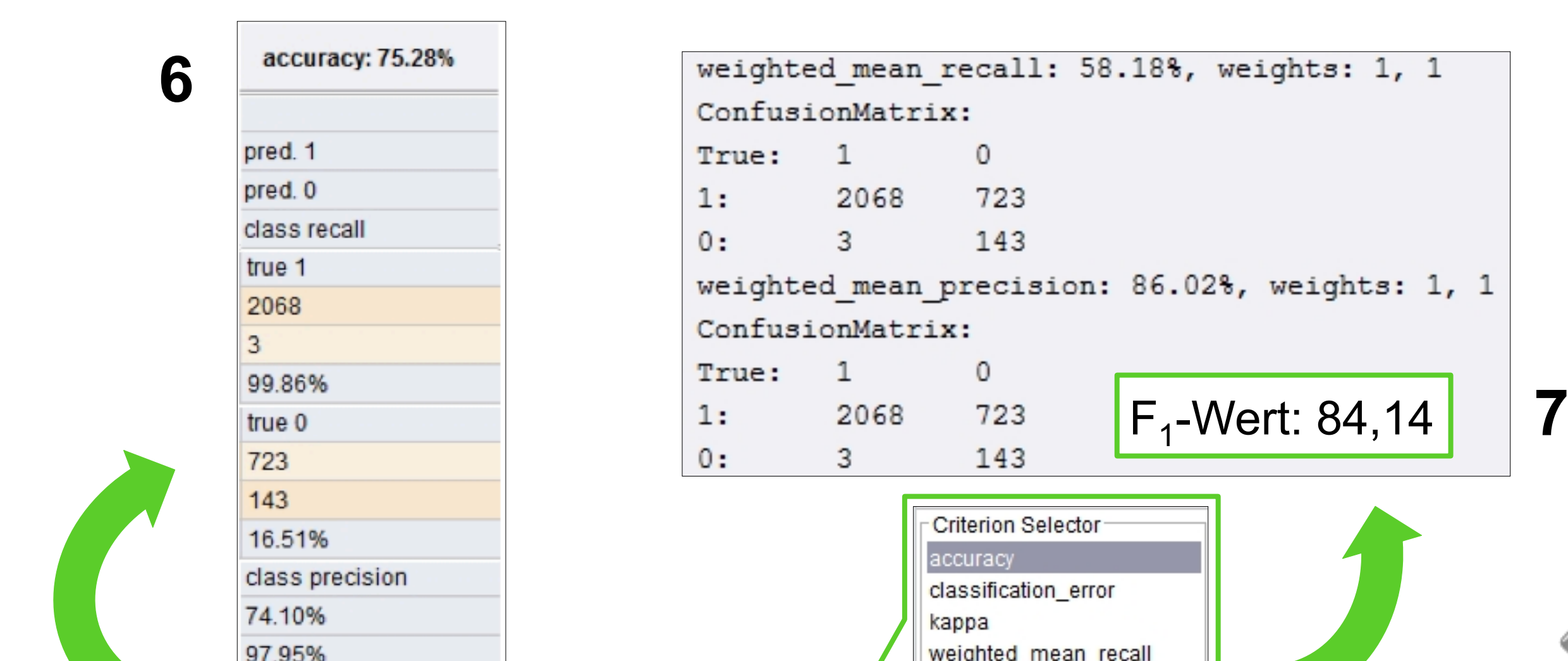
Disambiguierung mit RapidMiner und KobRA-Plugin: Fallstudie zum Active Learning

- Untersuchungsgegenstand: Smiley :(in den Diskussionsseiten
- Motivation: IBK-spezifische Einheiten wie Smileys sind oftmals ambig, die Aussortierung von Falsch-Positiven (s.u.) ist notwendig:

- 1) *parboiled* > [pa:(r)bold, engl.]
- 2) *Aus: § 2 EBO:(Absatz 3) Die Vorschriften [...]*
- 3) *einfache Intervalle [...] wie 1:2, 2:3, 3:4, n:(n+1)*

- Analyse und Ergebnisermittlung mit RapidMiner und KobRA-Plugin
- Vorgehen:**

1. Abfrage im Korpus (ggf. Festlegung der Zufallsstichprobe (sample size) und Annotationsrunden)
2. Durchführung des Active Learning (AL)-Verfahrens
3. Manuelle Annotation (Treffer vs. Pseudotreffer) im ausgegebenen *Annotation Dialog*; Annotationsrunden gemäß Festlegung
4. Anwendung des Verfahrens auf die restliche Stichprobe
5. Bewertung der automatischen Annotation (durch *Performance Classification*) mit anschließender manueller Nachannotation
6. Berechnung der Güte des Verfahrens (*Accuracy*)
7. Evaluation des Verfahrens: Berechnung des F_1 -Wertes



4

Row No.	labeled	labeled 0	labeled 1	unlabeled	minconfdiff
1	100	38	62	2828	0.214
2	200	138	62	2728	0.200
3	300	141	159	2628	0.131

Row No.	text_attr	label_attr	prediction	confidence(0)	confidence(1)
1	wurde leider aus der	?	1	0,3983208	0,6016792
2	stellt sich die WP se	?	1	0,3983208	0,6016792
3	man das Flash-Form	?	1	0,3983208	0,6016792
4	" Media:((Billig und	?	1	0,3949579	0,6050421
5	is nur ne wiki-Tabelle	?	1	0,3983208	0,6016792
6	bisher nicht schlauer	?	1	0,3983208	0,6016792
7	was der Artikel mir ü	?	1	0,3983208	0,6016792
8	1 zu 1 kopiert von Fr	?	1	0,3983208	0,6016792
9	Flammenfärbung: Li	?	1	0,3983210	0,6016790
10	ich überhaupt keine	?	1	0,3983208	0,6016792
11	-- (leider nicht hier re	1	1	0,2690453	0,7309547
12	Bin enttäuscht. :(--B	?	1	0,3628460	0,6371540
13	sogar dafür sogar hi	?	1	0,3983208	0,6016792
14	und habe dafür gera	?	1	0,3983208	0,6016792
15	Jetzt hab ich ja nix z	?	1	0,3983208	0,6016792

