université de TOURS

Can models learned from a dataset reflect acquisition of procedural knowledge?

An experiment with automatic measurement of online review quality

Martina Megasari, Pandu Wicaksono, Chiao Yun Li, Clément Chaussade,
Shibo Cheng, Nicolas Labroche, *Patrick Marcel*, Verónika Peralta

DOLAP 2018

Contributions

- (Yet another) model of reviews helpfulness
- A first assessment of the skill of writing helpful reviews
- Showing that skill assessment makes sense even for models learned automatically

Principle

Reviewer

Review 1

★★★★★

Helpfulness

…

Review n

★★★★★

Helpfulness

Principle

Reviewer

Review 1

★★★★★

Helpfulness

…

Review n

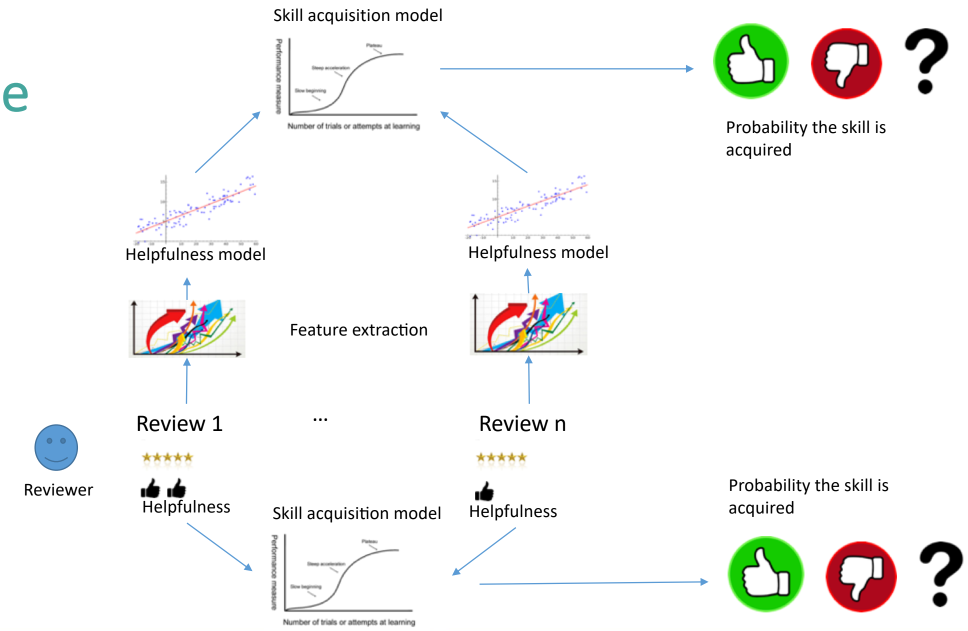
★★★★★

Helpfulness

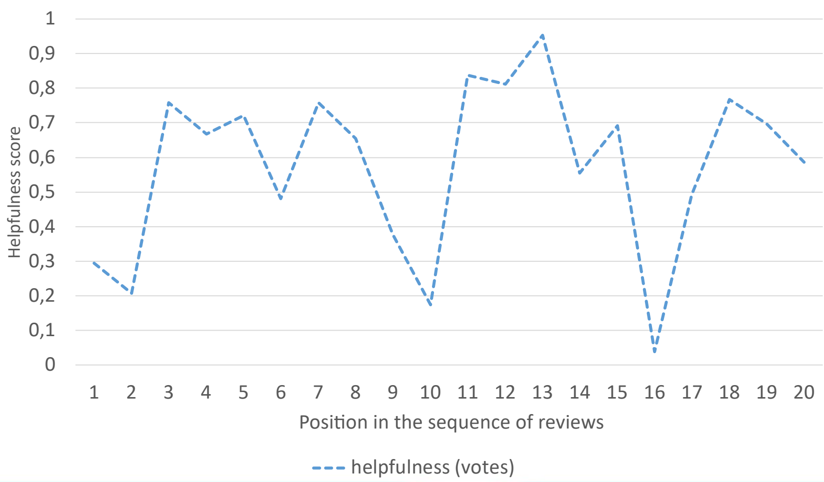
Skill acquisition model

Probability the skill is acquired

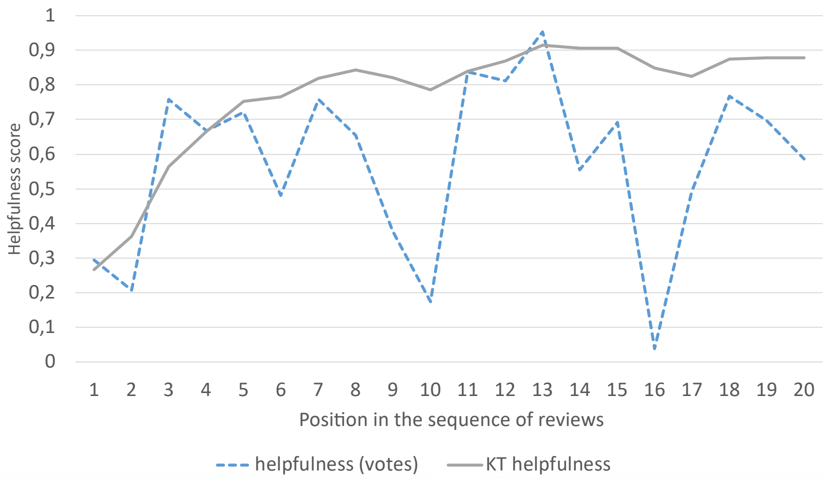
Principle



Tracing helpfulness for a reviewer

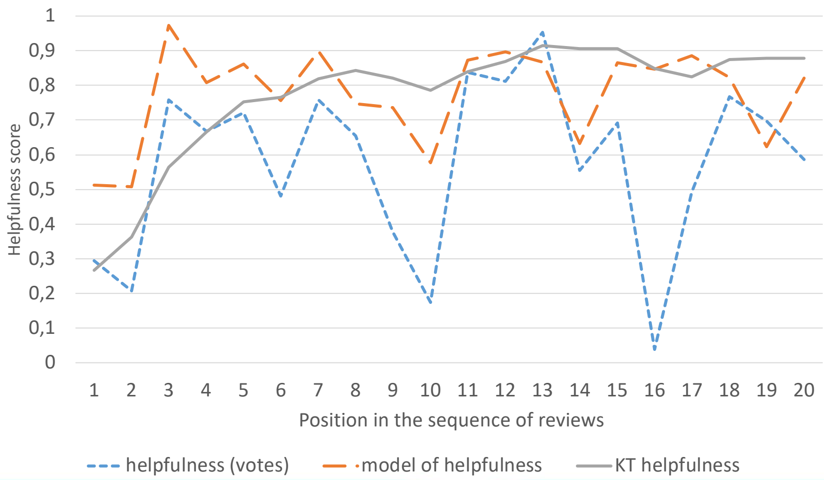


Tracing skill of the reviewer



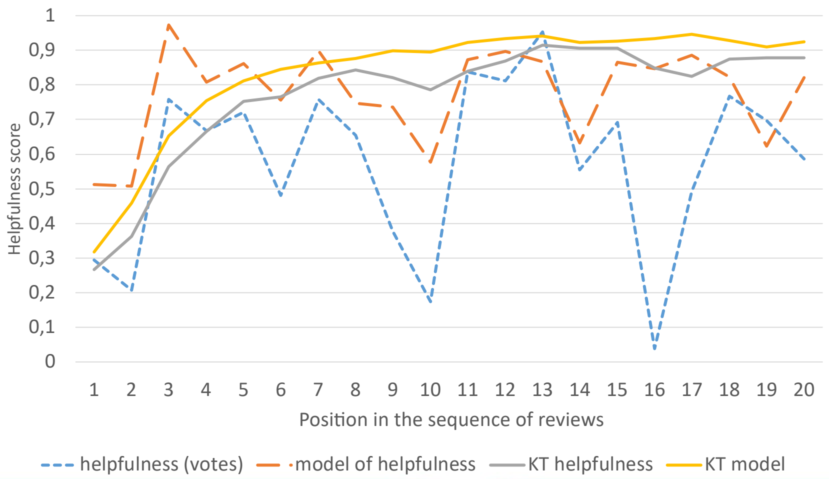
7

Tracing helpfulness of the model learned



8

Tracing skill of the model learned



What do we need?

- **Skill acquisition model**
  - Bayesian Knowledge Tracing
- **Data**
  - Amazon.com book reviews
- **Model**
  - Linear combination of features that participate in helpfulness

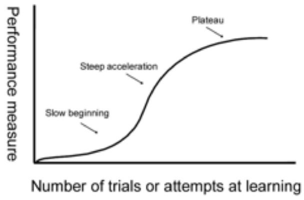
Skill acquisition: Bayesian Knowledge Tracing (KT)

User-centric paradigm for evaluating procedural knowledge
[Corbett & Anderson, UMAI 1995]

Skill S

OUTPUT: P(Ln)

Probability that skill S is mastered after exercice n



Hypotheses behind KT

- **It targets procedural knowledge**
  - Knowledge about how to do something
  - Application of procedural knowledge may not be easily explained
  - Different from declarative knowledge, that is often verbalized
- **Problem resolution is binary**
  - Pass/fail scheme
- **No forgetting**
- **4 parameters usually set empirically**

The 4 parameters of KT

- **P(L0): initial knowledge**
  - Probability the skill is already mastered before the first problem
- **P(T): transition from not mastered to mastered**
  - Probability the skill will be learned at each new opportunity
- **P(g): Guess**
  - Probability the learner will guess correctly while the skill is not mastered
- **P(s): Slip**
  - Probability the learner will make a mistake while the skill is mastered

Definition

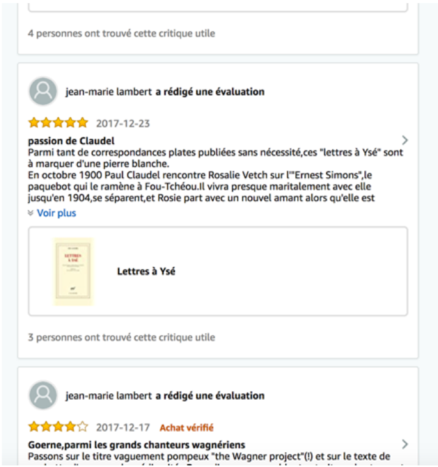
- **Probability the skill is mastered at step n**
  - $P(L_n|X_n = x_n) = P(L_{n-1}|X_n = x_n) + (1 - P(L_{n-1}|X_n = x_n))P(T)$
  - Intuition : probability the skill is learned at step n-1 or not learned at step n-1 but learned at this step n
- **With**
  - $X_n = 1$ means problem n resolved sucessfully, $X_n=0$ means not resolved
  - $P(L_{n-1}|X_n = 1) = P(L_{n-1})(1 - P(s)) / (P(L_{n-1})(1 - P(s)) + (1 - P(L_{n-1}))P(g))$
    - intuition: skill has been learned and used correctly / all cases of correct resolution
  - $P(L_{n-1}|X_n = 0) = P(L_{n-1})P(s) / (P(L_{n-1})P(s) + (1 - P(L_{n-1}))(1 - P(g)))$

KT extensions we implemented to fit our context

- **Non-binary problem resolution**
  - KT with partial credits [Wang & Heffernan, AIED 2013]
- **Parameter learning avoiding local minimum, degenerate parameters and computational costs**
  - Estimating the most likely opportunity at which each individual learned the skill [Hawkins & al., ITS 2014]
- **Github link**
  - https://github.com/Cubiccl/Continuous-Knowledge-Tracing/releases/tag/1.0

Data: Amazon book reviews [He & McAuley, WWW 2016]

- **In our context**
  - the skill is that of writing helpful reviews
  - each written review is treated like an opportunity to exercise the skill
  - Actual helpfulness is the ratio of helpful votes received by the review
- **Preprocessing details in the paper**



Features & metrics for the model of helpfulness

- **16 features grouped in 3 categories**
  - Conformity
    - Rating, polarity, deviation to average rating
  - Understandability
    - Spelling error ratio, 5 classical readability measures
  - Extensiveness
    - Text and summary length
- **Consistent with other models in the literature [Korfiatis & al., ECRA 2012]**
  - More sophisticated models exist, but our point was to test a "simple" one

17

The model

- **Linear combination of feature scores**
- **Learned with linear regression, perceptron, SVM**
  - Regression was the best compromise between time and effectiveness
  - Feature selection had no significant impact

rating	0.31117594
polarityReviewText	0.36708846
polaritySummary	0.05166703
deviation	-0.20847153
reviewTextSER	0
summarySER	-0.28603436
reviewTextFOG	-1.10263702
summaryFOG	0
reviewTextFK	4.37638627
summaryFK	0.12251469
reviewTextARI	5.01873535
summaryARI	-0.4099729
reviewTextCLI	0.31215745
summaryCLI	0.79694206
reviewTextLength	0.30807426
summaryLength	0
bias	-4.26391009

18

Tests

- **Implementation**
  - Java 8
  - Weka 3.8 for model learning
  - SentiWordNet for polarity extraction
  - Stanford POS tagging library for part-of-speech tagging
- **2 preprocessed datasets**
  - minVotes = 12: 41,681 reviews
  - minVotes = 23: 11,083 reviews
  - In each dataset, reviewers have between 30 to 50 reviews

Tests

- **Helpfulness model accuracy is similar to the recent proposals of the state-of-the-art**
  - RMSE: the error between the helpfulness *model* scores and the *actual* helpfulness scores

Correlation Coefficient	0.608
Efron's R^2	0.3697
Mean Absolute Error	0.1521
Root Mean Squared Error	0.2014

Tests

- **Using KTs**
  - **a-mKRMSE**: error between the KT of the *actual* helpfulness scores and the KT of the helpfulness as computed with the *model*
  - **a-AggKRMSE**: error between the KT of the *actual* helpfulness scores and the aggregation of the KT scores of *each feature taken independently* (sub-skill)

	Scores	minVotes = 12	minVotes = 23
Actual skill	mean(L_n)	0.968337	0.960511
	variation(L_n)	0.025213	0.033238
Model	$P(L_0)$	0.007504	0.033457
	$P(T)$	0.030262	0.077669
	mean($P(G)$)	0.349992	0.369982
	variation($P(G)$)	0.007067	0.0147
	mean($P(S)$)	0.412574	0.412882
	variation($P(S)$)	0.016212	0.025905
	mean(L_n)	0.783885	0.800687
	variation(L_n)	0.090915	0.090820
Aggre-gated	mean(L_n)	0.999943	0.999991
	variation(L_n)	0.000584	0.000122
	a-mKRMSE	0.164619	0.156373
	a-AggKRMSE	0.064818	0.081964

Lessons learned & perspectives

- **KT is optimistic and has an intrinsic smoothing behavior**
- **Finer skills works better than coarser ones**
- **Perspectives**
  - Short term
    - Testing with more helpfulness models and skill acquisition models
    - Understanding better the relationship between the linear coefficient learned for the helpfulness model and the KT parameters of the corresponding sub-skills
  - Longer term
    - Application to other datasets, contexts and skills
      - Eg, how to assess data exploration, or how to assess deep learning's productions

References

- [Corbett & Anderson, UMAI 1995]
  - Albert T. Corbett, John R. Anderson: Knowledge Tracing: Modelling the Acquisition of Procedural Knowledge. User Model. User-Adapt. Interact. 4(4): 253-278 (1995)
- [Wang & Heffernan, AIED 2013]
  - Yutao Wang, Neil T. Heffernan: Extending Knowledge Tracing to Allow Partial Credit: Using Continuous versus Binary Nodes. AIED 2013: 181-188
- [Hawkins & al., ITS 2014]
  - William J. Hawkins, Neil T. Heffernan, Ryan Shaun Joazeiro de Baker: Learning Bayesian Knowledge Tracing Parameters with a Knowledge Heuristic and Empirical Probabilities. Intelligent Tutoring Systems2014: 150-155
- [He & McAuley, WWW 2016]
  - Ruining He and Julian McAuley. 2016. Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering. In WWW. 507–517.
- [Korfiatis & al., ECRA 2012]
  - Nikolaos Korfiatis, Elena García-Bariocanal, and Salvador Sánchez-Alonso. 2012. Evaluating content quality and helpfulness of online product reviews: The interplay of review helpfulness vs. review content. Electronic Commerce Research and Applications 11, 3 (2012), 205–217.

Q&A