ICADL 2016 @ University of Tsukuba, Dec 8

Improved Identification of Tweets that Mention Books: Selection of Effective Features

Shuntaro Yada and Kyo Kageura
Laboratory of Library and Information Science
Graduate School of Education
The University of Tokyo

Task

- To classify tweets that contain full book title strings into Tweets that Mention Books (**TMB**s) or Noise
- Currently focusing on Japanese Twitter and Japanese tweets

Method

TMB	Noise
When Breath Becomes Air is the most profound, life changing book I have ever come across. It will stick with me through my whole life	To **the girl on the train** who is currently drawing her eyebrows on. No.

Both tweets can be distinguished utilising information of contextual words

- Solve this task as classification problem using supervised machine learning technique with Bag-of-Words based features

Purpose of this research

To propose additional effective features for our TMB identifier

- In our previous research, we tackled with this task and achieved a promising performance, but pursue its improvement for a practical level

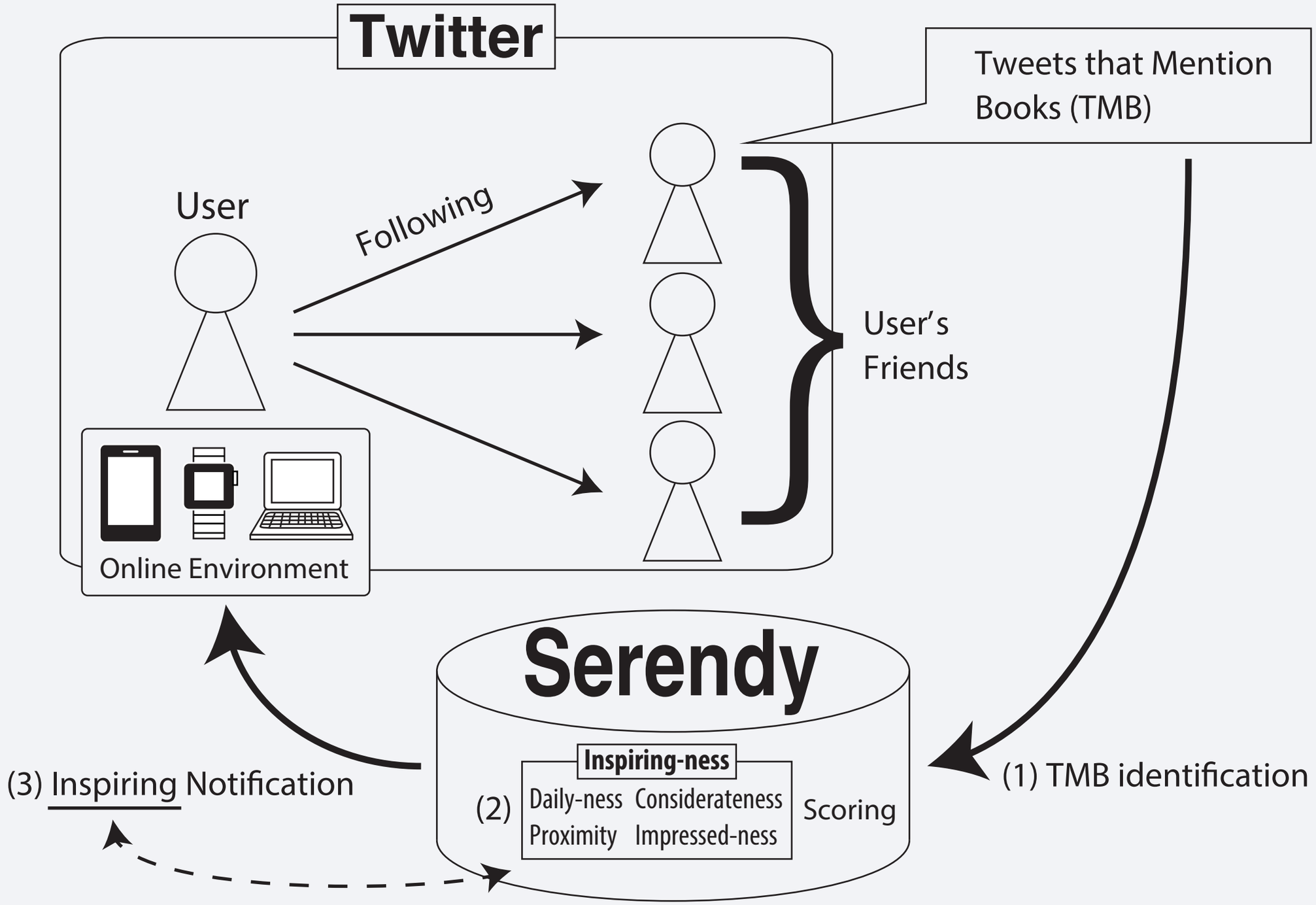
Yada, S., & Kageura, K. (2015). Identification of Tweets that Mention Books: An Experimental Comparison of Machine Learning Methods. ICADL2015. Seoul, Korea.

Background

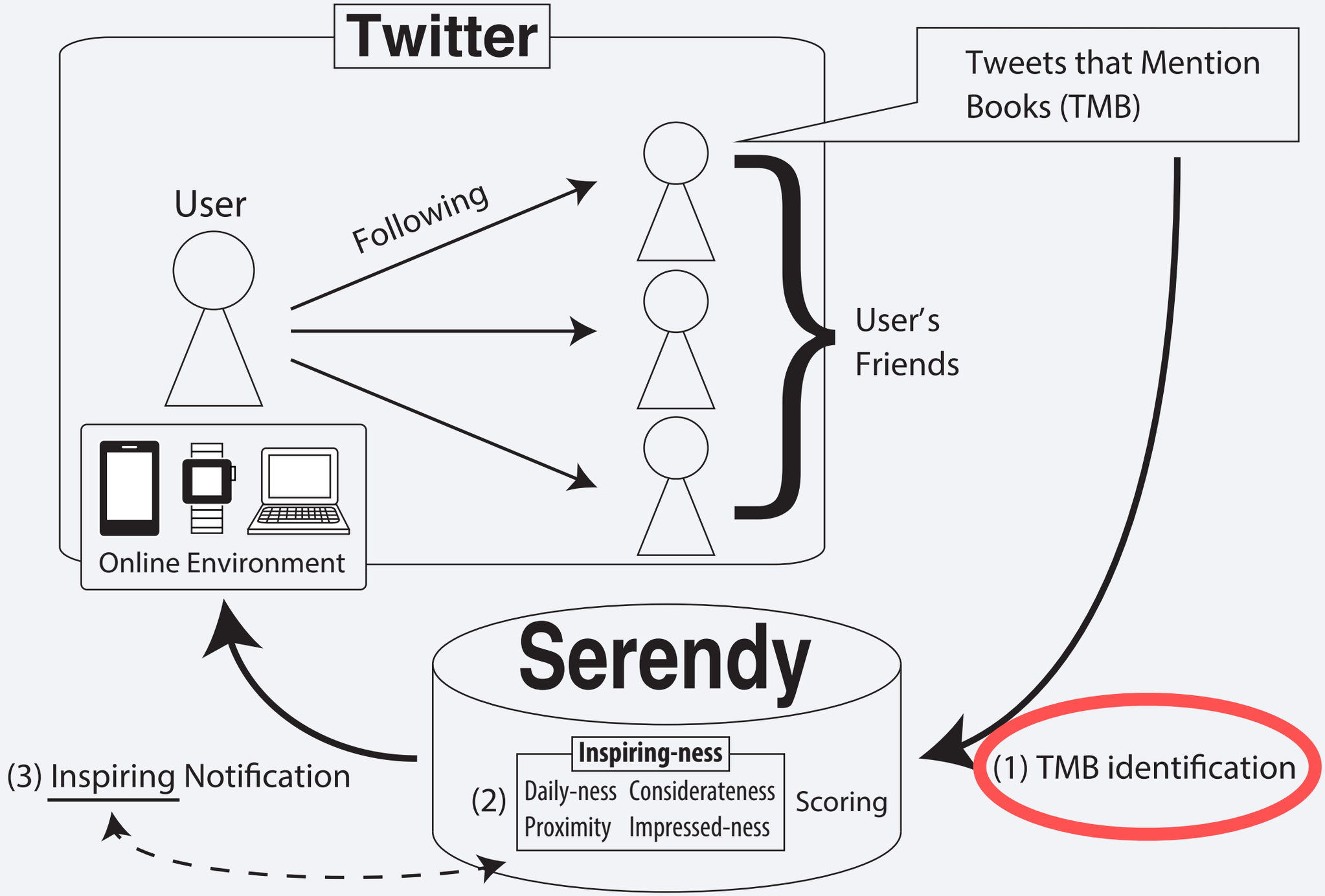
Development and evaluation of a book recommendation system named ***Serendy***

- **User:** infrequent readers
- **Methods:** provide users with recommendations of some books that the users' friends mentioned or alluded to in SNSs

Background


Twitter
Tweets that Mention Books (TMB)
User
Following
User's Friends
Online Environment
Serendy
Inspiring-ness
(2)
Daily-ness
Considerateness
Proximity
Impressed-ness
Scoring
(1) TMB identification
(3) Inspiring Notification

Background


Twitter
Tweets that Mention Books (TMB)
User
Following
User's Friends
Online Environment
Serendy
Inspiring-ness
(2)
Daily-ness
Considerateness
Proximity
Impressed-ness
Scoring
(1) TMB identification
(3) Inspiring Notification

Dataset

Initial data set (Yada and Kageura, 2015):

- Using a comprehensive list of book titles, we searched for tweets containing the same strings as book titles
- We annotated a small sample of the tweets manually

TMB	Noise
436	5,563

Keyword augmentation

Keyword augmented data:

Searched for tweets containing one of the five keywords below as well as book titles

- Keywords were selected from a set of words appearing much more frequently in the TMBs than noise tweets

Keywords	読了 finished reading	読む read	再読 re-read	読破 read through a book	読み応え worth reading
#tweets	2,625	881	378	319	173

TMB	Noise
4,839	5,563

Design of TMB identifier (1)

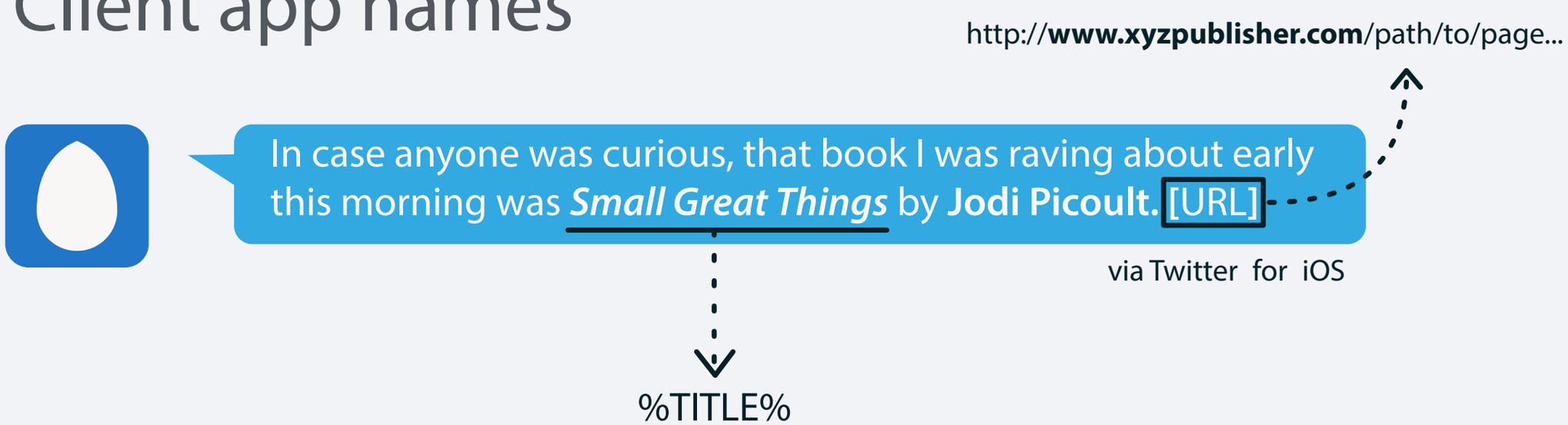
Classification algorithm: Maximum Entropy Modelling (MaxEnt)

- This performed better in terms of the balance of scores and training speed than several other algorithms

Design of TMB identifier (2)-1

Baseline Features (Yada and Kageura, 2015):

- Bag-of-Words of:
  - Tweet texts (with book titles abstracted)
  - URL host names
  - Client app names

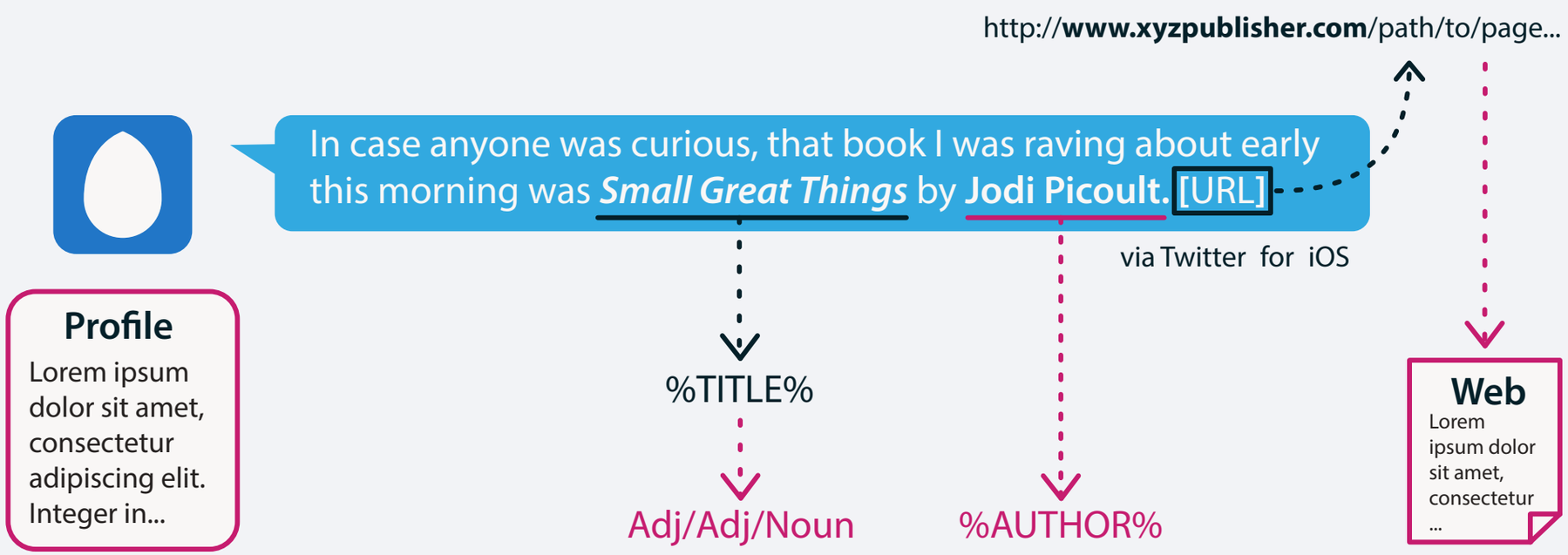


Design of TMB identifier (2)-2

Proposed Features:

- Bag-of-Words of:
  - Profile texts (**profile**)
  - Linked web pages' body texts (**link**)
  - POS tags of book titles (**title-ness**)
  - Abstracted bibliographic fields (**bib**)

"**diff**" **option**: whether or not to be differentiated from words derived from other features (such as tweet texts)



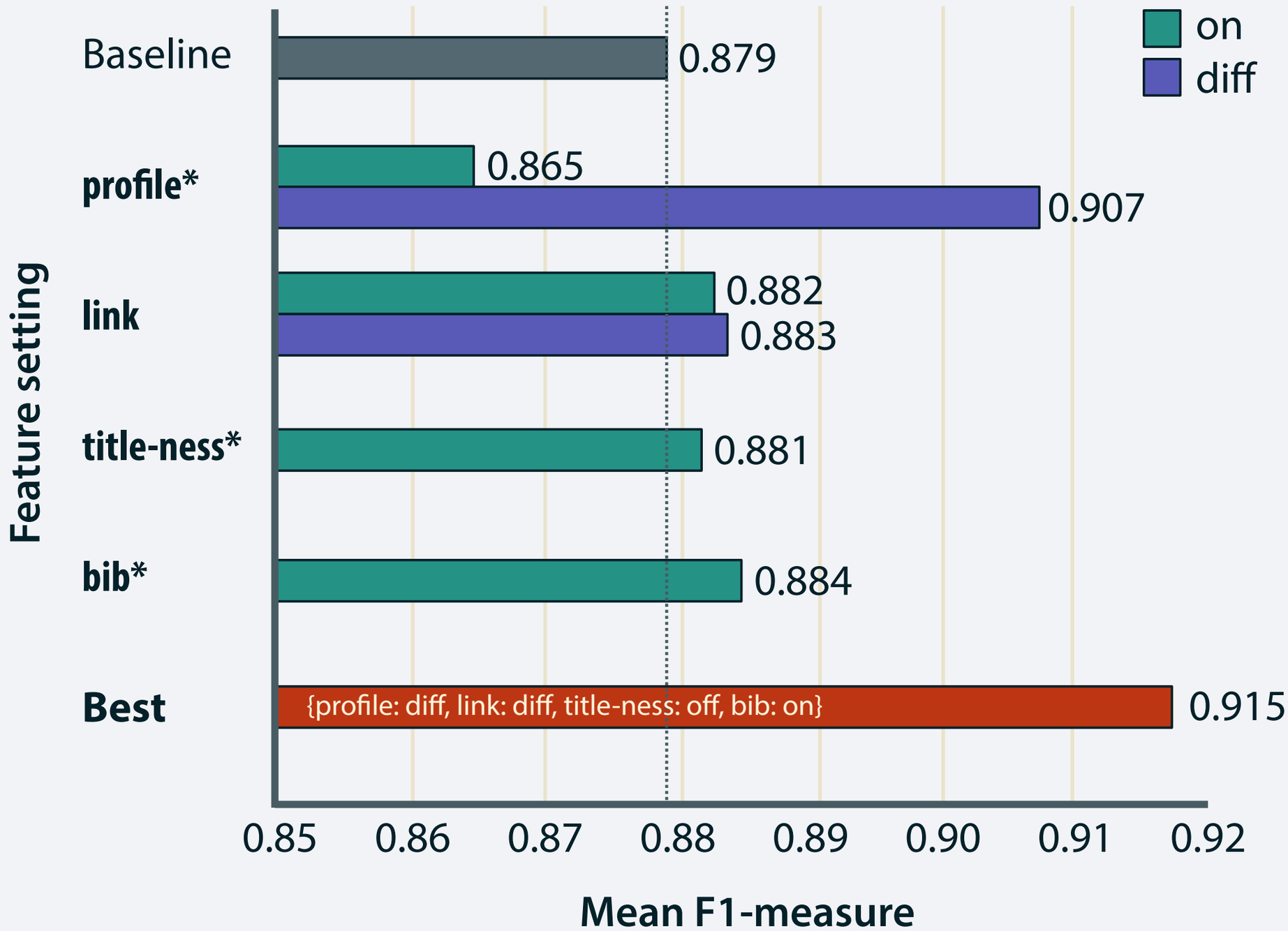
Experiment

- 3-fold cross validation for each combination of proposed four features
  - profile, link = {off, on, diff}
  - title-ness, bib = {off, on}

test for all combinations

- Use F1-measure (harmonic mean of precision and recall) for evaluation

Result



* statistically significant (t-test or ANOVA, $p < 0.5$)

Conclusion

- TMB identifier achieved **0.915** F1-measure
  - Ready for practical use
- **profile** contributed most
- More precise representation of **title-ness** seems to exist