Identification of Tweets that Mention Books: An Experimental Comparison of Machine Learning Methods

Shuntaro Yada and Kyo Kageura
The University of Tokyo, Japan

ICADL2015, 10th December @South Korea

In this research...

- We examined four machine learning (ML) methods with different combination of features in order to identify Japanese **Tweets** that Mention Books (**TMBs**)

Agenda

1. Introduction
2. Methodology
3. Experiments
4. Conclusion

1 Introduction

Why do we identify TMBs?

- We are developing **a book recommendation system** named ***Serendy***
  (Shuntaro Yada, Development of a Book Recommendation System to Inspire “Infrequent Readers”, ICADL2014)

- TMB identifier is the core technical module of Serendy

What is Serendy?

Simulates the following situation online

a person discovers books by chance through *daily informal conversation* with friends or acquaintances

Twitter (SNS)

Background

The environment of reading or discovering books is changing

- Personalisation of information consumption
- Increasing popularity of e-Books

➡ We have fewer opportunities to be exposed to books by chance and/or unconsciously

Possible types of TMBs

TMBs could be tweets…

- Containing bibliographic information **explicitly**
    - Including *accurate* bibliographic information
    - Mentioning books *incorrectly/casually*
- Mentioning a book **implicitly**

Target type of TMBs

We focus on TMBs that **contain full book titles**

- They would be the majority of explicit TMBs
- We have comprehensive book titles database
- This task allows us to gain insight to other TMB types

Task definition (1)

- Identifying tweets containing full book titles can be defined as a kind of **named entity (NE) recognition task**

  ← A comprehensive book title list is available

- But identifying book title is not enough

  - **Noise**: Book titles consist of ordinary expressions (e.g., “Kidnapping”, “The Circle”)

  - **Spam**: Tweets containing full book titles are often spam

Task definition (2)

Therefore we solve the task of identifying TMBs as text classification:

- ➡ Classifying tweets that contain strings which are the same as book titles (NE) into TMB/non-TMB (Noise + Spam)

2 Methodology

Data set

- We gathered 70,844 tweets that contain only one book title string

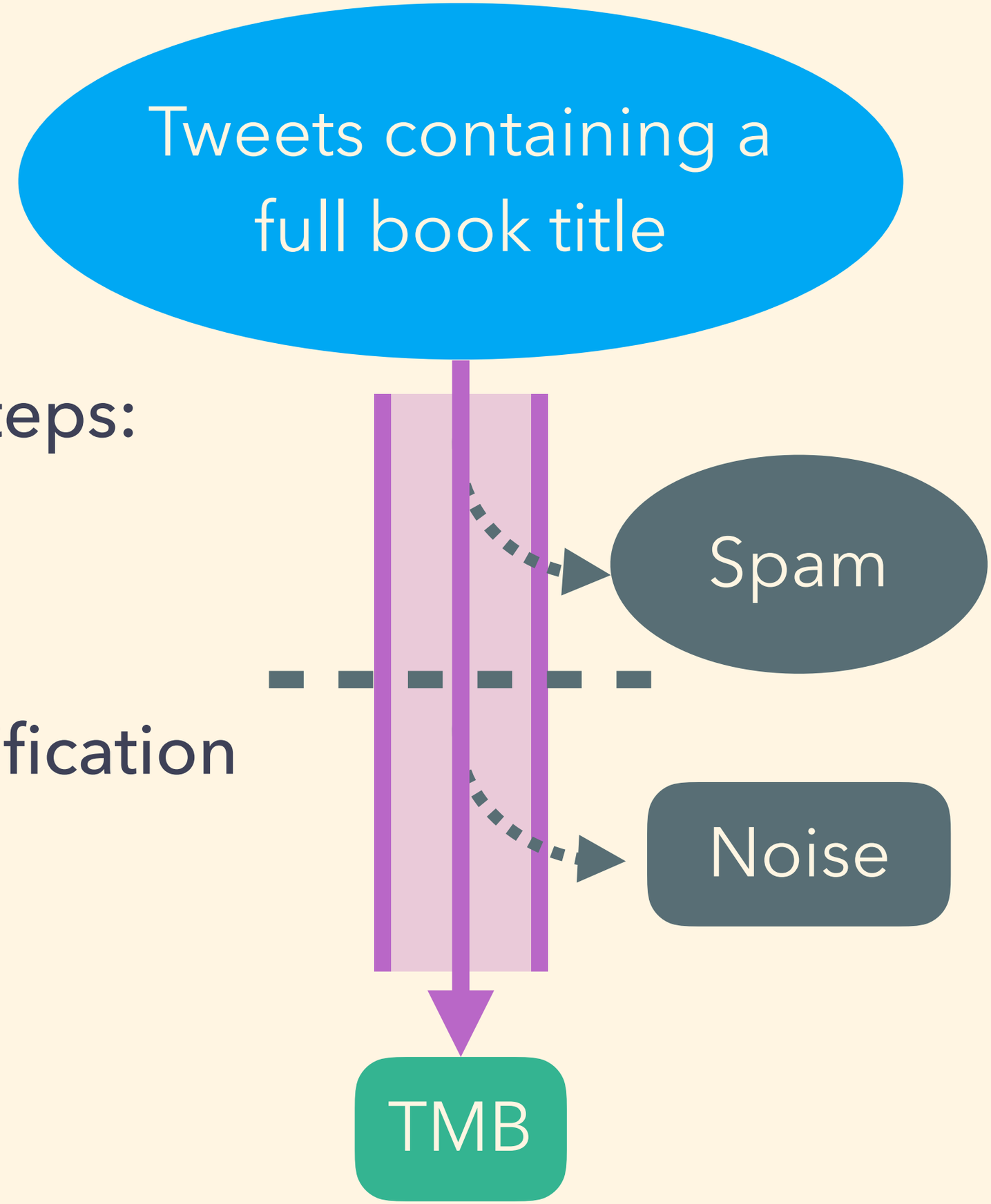
- Then we selected and manually annotated tweets containing book titles appearing less than three times in total: **8,528 tweets**

TMB	Spam	Noise
350	4040	4138

Pipeline

The following two steps:

1. Spam filtering
2. TMB/Noise classification



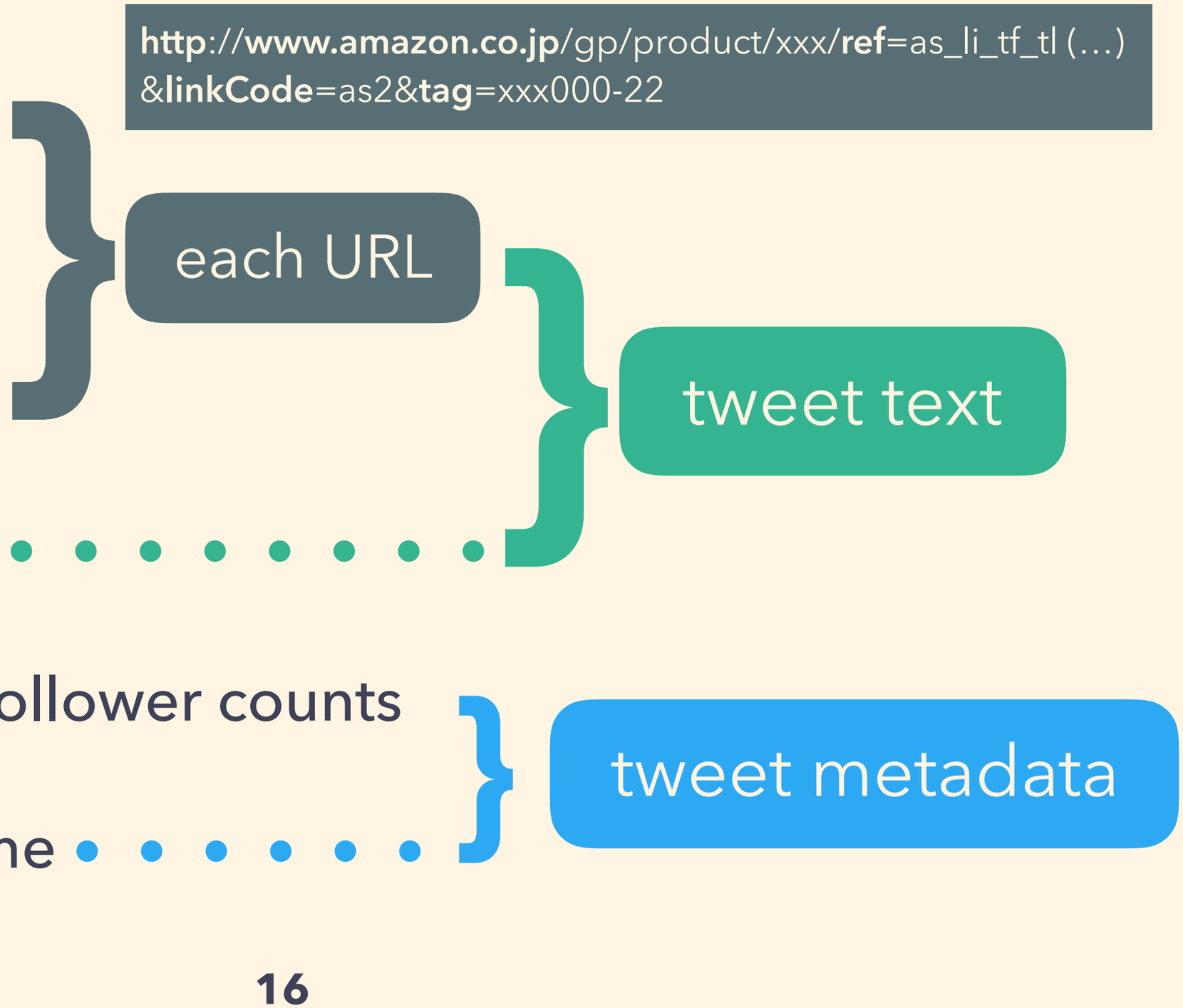
ML methods

Four candidates of algorithms have been frequently applied and have performed well in text classification

- Naïve Bayes
- MaxEnt (Maximum Entropy Modelling)
- SVM (Support Vector Machine)
- Random Forest

Features used for spam filtering

- URL domain
- URL query keys
- SSL/TLS usage
- Hashtag count
- Following and follower counts
- Application name



Features used for TMB/Noise classification

- URL domain
- Application name

} same as spam filtering

- **Tokens within tweets**
  - ❖ Part-Of-Speech tag
  - ❖ NE abstraction
  - ❖ Identifying tokens surrounding the title

} Token options

Experiments

Experimental work

Two phases:

Phase 1: Finds the best combination of a ML algorithm and its settings for each step of the proposed pipeline

Phase 2: Evaluates our proposed method by controlled experiments

Phase 1

Find the two best classifiers and their settings for each step of the proposed pipeline

Experiment (a): Grid-search for spam filtering

Experiment (b): Grid-search for TMB/Noise classification

Each best classifier will be used for the pipeline

Phase 1: Experiment (a)

- MaxEnt performed the best for spam filtering

ML methods	F-score
MaxEnt	0.978
Random Forest	0.937
Naïve Bayes	0.772
SVM	0.542

Phase 1: Experiment (b)

- SVM performed the best for TMB classification
- Tokens surrounding the title were effective

ML methods	Token options	F-score
SVM	(**Title**)	**0.699**
MaxEnt	(**Title**)	**0.656**
Naïve Bayes	(**Title**, NEabst, POS)	**0.577**
Random Forest	(**Title**, NEabst, POS)	**0.503**

Phase 2

Evaluate the proposed method against one-step approach to identify TMBs

Experiment (c): Two-step pipeline (proposed method)

Experiment (d): Direct classification of TMB (one-step approach)

Phase 2: Result

Experiment (c)

Method	Precision	Recall	F-score
Two-step pipeline	0.698	0.681	0.686

Experiment (d)

ML methods	Precision	Recall	F-score
Random Forest	0.892	0.343	0.490
SVM	0.331	0.666	0.439
MaxEnt	0.291	0.161	0.389
Naïve Bayes	0.287	0.426	0.342

4 Conclusion

Summary (1)

- We tackled the task of identifying Japanese tweets that mention books (TMBs)
- Starting from tweets that contain full book titles based on a book title list, we proposed a two-step pipeline:

  ➡ Spam filtering followed by TMB/Noise classification

Summary (2)

- For spam filtering, MaxEnt performed the best, while SVM performed the best for TMB/Noise classification
- Two-step pipeline performed better than direct (one-step) TMB classifiers

Outlooks

- Improve our pipeline by using tweet authors' profiles
- Handle TMBs containing abbreviated book titles

Applicability of this research

- This approach might be applicable to other difficult types of NE recognition similar to book title:
  - Song
  - Film
  - TV program…

Thanks for your listening

Reprint of summary (1)

- We tackled the task of identifying Japanese tweets that mention books (TMB)
- Starting from tweets that contain full book titles based on a book title list, we proposed a two-step pipeline:

➡ Spam filtering followed by TMB/Noise classification