Results on Tracks 1 and 2 of KDD Cup 2013

Chih-Jen Lin

Department of Computer Science
National Taiwan University

Joint work with members of the team "Algorithm" from National Taiwan University

Outline

1. Introduction
2. Track 1: paper-author identification
   - Feature generation
   - Classification
   - Ensemble and post-processing
   - Results
3. Track 2: author disambiguation
   - Strategies and architecture
   - Implementation
   - Ensemble and typo handling
   - Analysis
4. Conclusions

Outline

1. Introduction
2. Track 1: paper-author identification
   - Feature generation
   - Classification
   - Ensemble and post-processing
   - Results
3. Track 2: author disambiguation
   - Strategies and architecture
   - Implementation
   - Ensemble and typo handling
   - Analysis
4. Conclusions

Team Members

- At National Taiwan University, we organized a course for KDD Cup 2013
- Three instructors, three TAs, and 18 students
- 18 students split to six sub-teams named by algorithms

  A*, Binary-Search, Dijkstra, K-means, Quick-Sort, Simplex
- Submission quotas are equally divided to six sub-teams

Outline

1. Introduction
2. Track 1: paper-author identification
   - Feature generation
   - Classification
   - Ensemble and post-processing
   - Results
3. Track 2: author disambiguation
   - Strategies and architecture
   - Implementation
   - Ensemble and typo handling
   - Analysis
4. Conclusions

Paper-author Identification

- Given an (author, paper) pair, did the author write the paper?
- What information do we have?
  - Author and paper profiles
  - Labeled (author, paper) pairs
    - Confirmed: author wrote paper
    - Deleted: author didn't write paper
- Under a given (author, paper), we use target author and target paper to distinguish them from other authors/papers

Paper-author Identification (Cont'd)

- Submission: ranking query papers for each query author

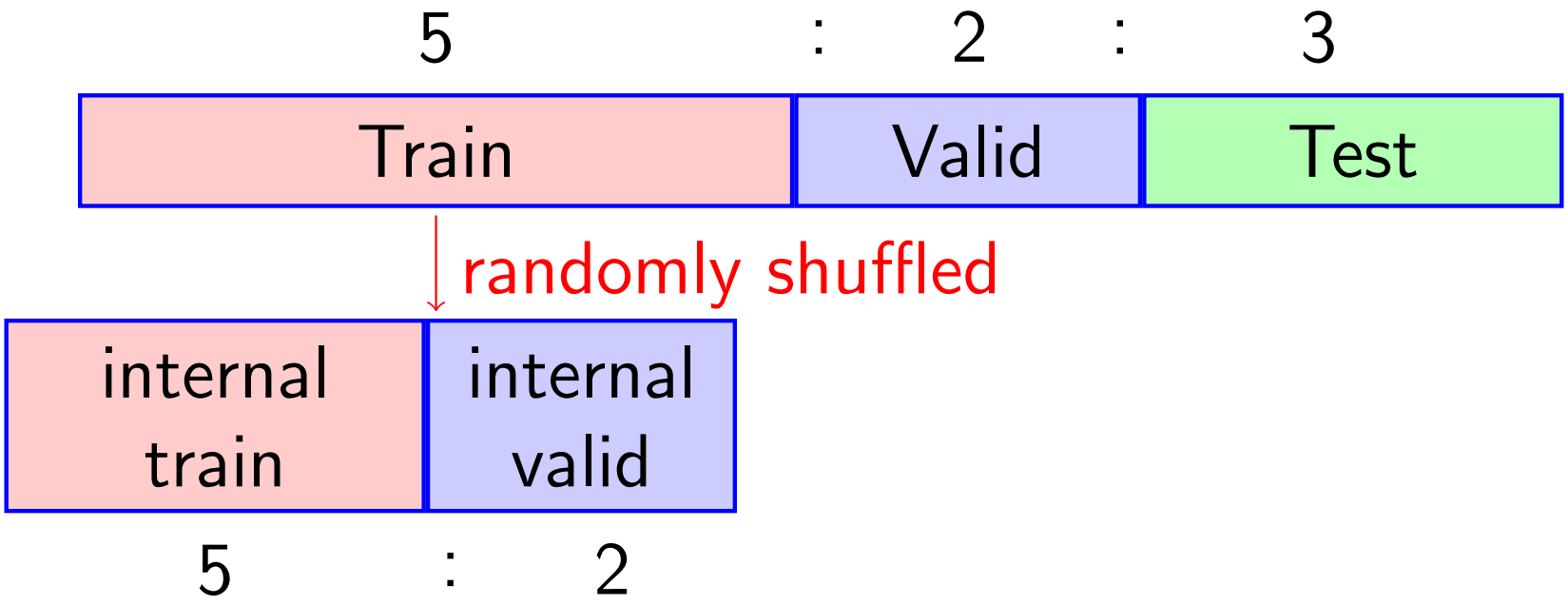
  Example: author 9417 has query papers 1, 2, 3, 6, and 9.

  If 3, 6 are confirmed and 1, 2, 9 are deleted, we should submit "9417, 3 6 1 2 9"
- MAP (Mean Average Precision) is the evaluation measure

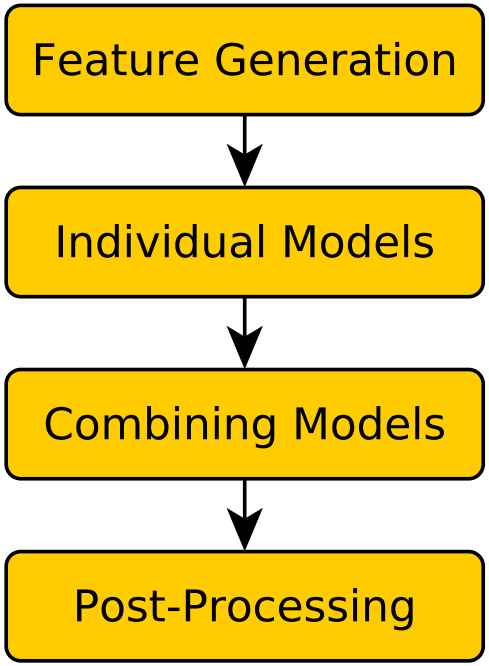
Internal Validation Set

We split `Train.csv` to internal training/validation sets due to the limited number of submissions per day.

This also avoids overfitting the leader board



System Overview



- List of 97 features can be seen in the paper

Outline

1. Introduction
2. Track 1: paper-author identification
   - Feature generation
   - Classification
   - Ensemble and post-processing
   - Results
3. Track 2: author disambiguation
   - Strategies and architecture
   - Implementation
   - Ensemble and typo handling
   - Analysis
4. Conclusions

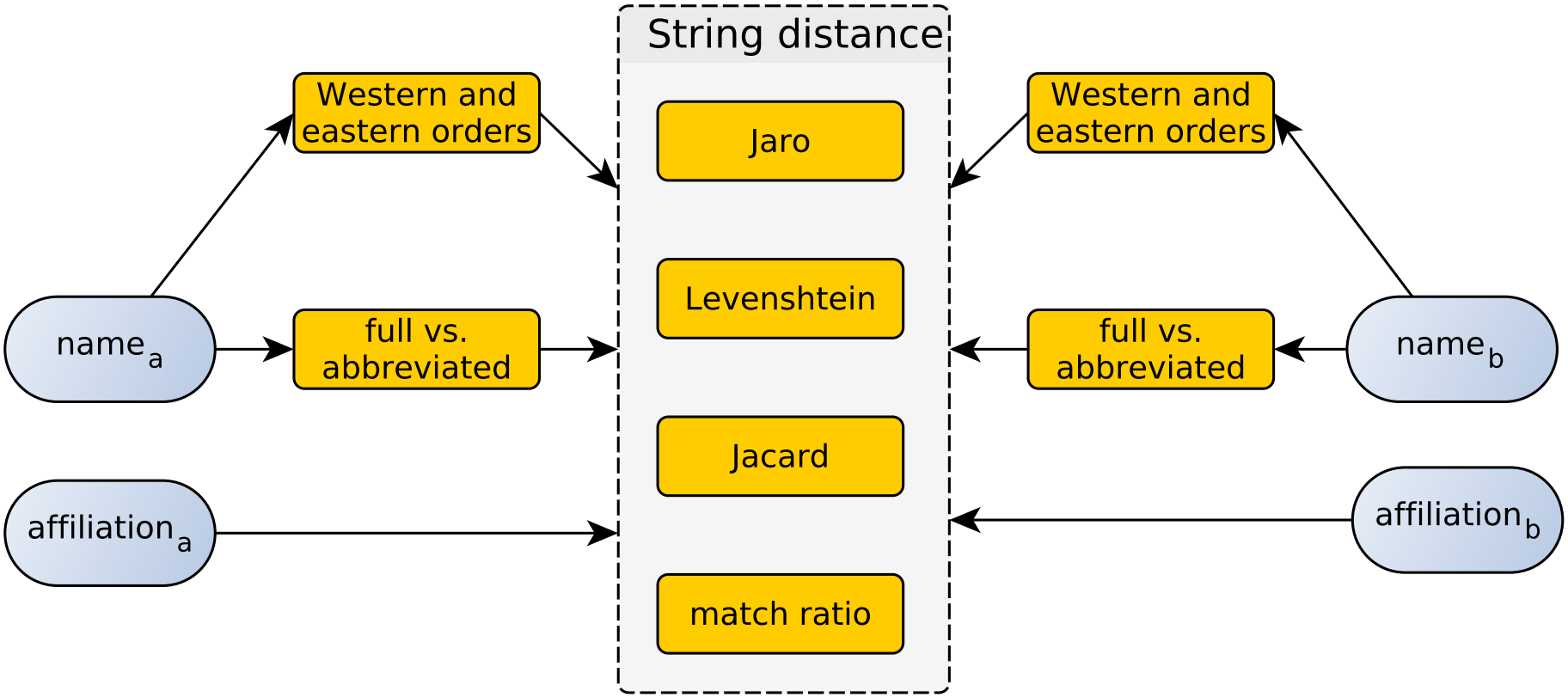
Features from Author Profiles

- Given a query ($\text{author}_{1360414}$, $\text{paper}_{1841516}$) .
  What information do we have about the author?
- `Author.csv`: 1360414,Chih-Jen Lin,National Taiwan University
- `PaperAuthor.csv`: 1841516,1360414,Chih-Jen Lin,"National Taiwan University, Taipei"
- Distance between target author's names, affiliations, etc. in two csv files $\Rightarrow$ features to indicate consistency

Features from Author Profiles (Cont'd)

- We need to address two issues
  - Distance between full and abbreviated names
  - Western and eastern order of names
    Example: "Chih Jen Lin" and "Lin Chih Jen"
- See paper for details

Features from Author Profiles (Cont'd)



Features from Coauthors Names

- Example: deleted paper 5633 of Li Zhang has two authors with the same name
- Relation between target author and authors of target paper can be features
- Examples
  1. Minimum name distance between the target author and authors of the target paper
  2. Same as 1, but check abbreviated names

Features for Author/Paper Consistency

- Information should be consistent across papers and authors
- Examples
  1. Maximum distance between target author's affiliation and affiliations of co-authors in target paper
  2. Maximum distance between target paper's title and target author's papers

Missing Value Handling

- Two empty strings have zero distance

$$d(\text{'Chih J Lin'}, \text{'C Jen Lin'}) \geq d(\text{''}, \text{''})$$

- Replace distance between empty strings with non-zero value

Distance	value
Jaro	0.5
Jacard	0.5
Levenshtein	average length of all entries

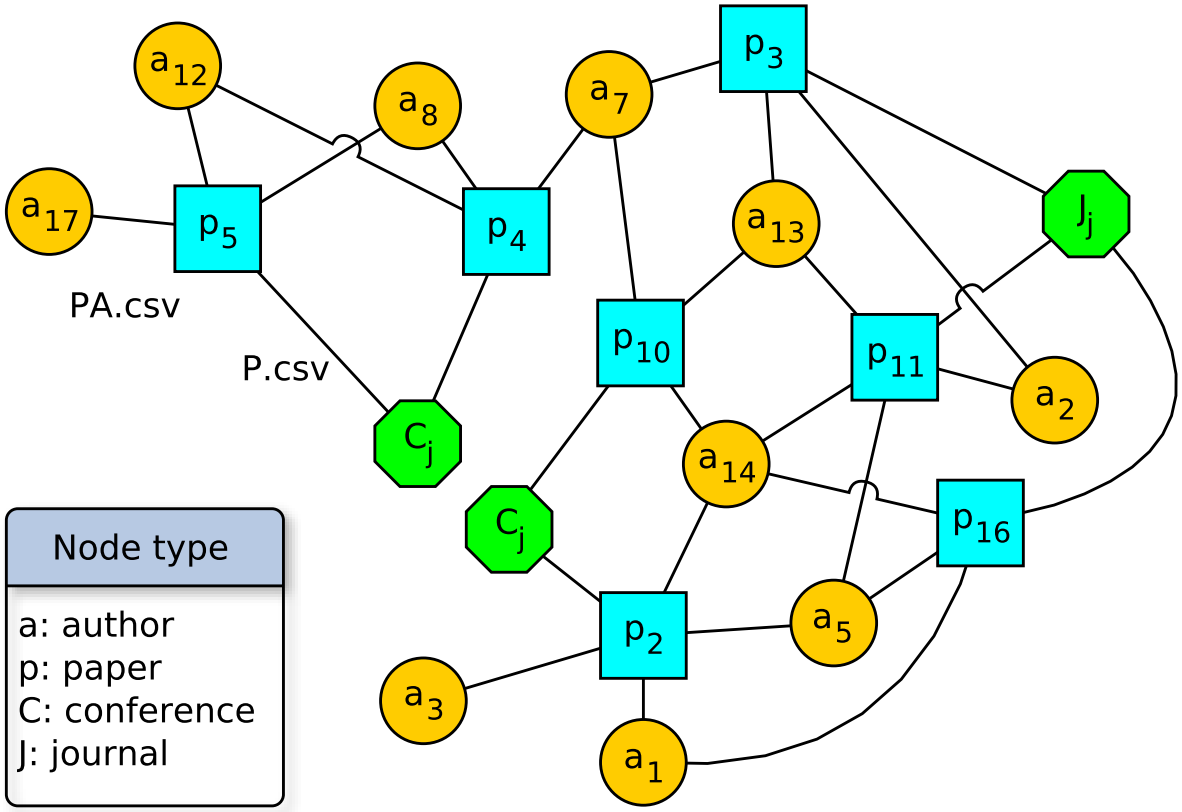
- Missing value indicators. Example: number of coauthors without affiliation information

Features Using Publication Time

- Examples
  1. Earliest/latest publication year of target author
  2. Publication year of target paper
- Data cleaning:
  - Years outside [1800, 2013] are removed
  - Then we must handle missing values

Features Based on A Network

- We construct a network of authors, papers, journals, and conferences



Features Based on A Network (Cont'd)

- From the network we can extract features to describe node relationships
- Examples
  1. # of publications of the author
  2. # of coauthored papers of the target author with all the coauthors of the target paper

Outline

1. Introduction
2. Track 1: paper-author identification
   - Feature generation
   - Classification
   - Ensemble and post-processing
   - Results
3. Track 2: author disambiguation
   - Strategies and architecture
   - Implementation
   - Ensemble and typo handling
   - Analysis
4. Conclusions

Classification

- Tree-based classifiers
  - Random forests (RF)
  - Gradient boosting decision tree (GBDT)
  - LambdaMART (LM)

classifier	tree ensemble	# of trees	parallel	MAP on `Valid.csv`
RF	bagging	12,000	yes	0.983340
GBDT	boosting	300	no	0.983046
LM	boosting	300	no	0.983047

- RF is sensitive to the initial random seed. Using 12,000 trees stabilizes the results

Outline

1. Introduction
2. Track 1: paper-author identification
   - Feature generation
   - Classification
   - Ensemble and post-processing
   - Results
3. Track 2: author disambiguation
   - Strategies and architecture
   - Implementation
   - Ensemble and typo handling
   - Analysis
4. Conclusions

Ensemble

- Weighted average over RF, GBDT, and LambdaMart
- Didn't use more complicated settings like regression because we have only three models
- A simple grid search on weights
- Final weights

  RF: 5, GBDT: 1, and LambdaMart: 1.

Post-Processing

- Our post-processing procedure is simple, but one thing to note is duplicated paper IDs.
- If an author has confirmed papers 1,2,2,4 and deleted paper 3
- The evaluation code seems to consider the 2nd "2" as a deleted paper
- Thus, MAP of 1,2,4,3,2 $>$ MAP of 1,2,2,4,3
- We move duplicated ones to the end

Outline

1. Introduction
2. Track 1: paper-author identification
   - Feature generation
   - Classification
   - Ensemble and post-processing
   - Results
3. Track 2: author disambiguation
   - Strategies and architecture
   - Implementation
   - Ensemble and typo handling
   - Analysis
4. Conclusions

Results

	Public	Private
1st of public	0.98554	0.98100
12th of public (ours)	0.98235	0.98259 (1st)

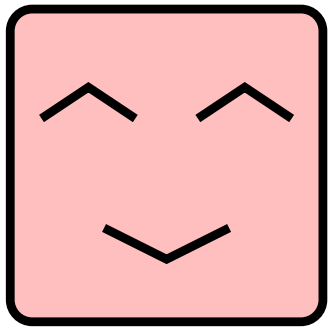
- Possible reasons of the best result in the end
  - Improvements after `Valid.csv` is released
    1. Data cleaning: unicode $\rightarrow$ ASCII
    2. Missing-value handling (0.98334)
    3. Ensemble (0.98390)

    We didn't give up even though we were the 12th!
  - We effectively use the internal validation set to avoid overfitting

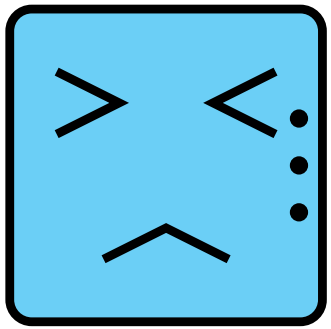
Outline

1. Introduction
2. Track 1: paper-author identification
   - Feature generation
   - Classification
   - Ensemble and post-processing
   - Results
3. Track 2: author disambiguation
   - Strategies and architecture
   - Implementation
   - Ensemble and typo handling
   - Analysis
4. Conclusions

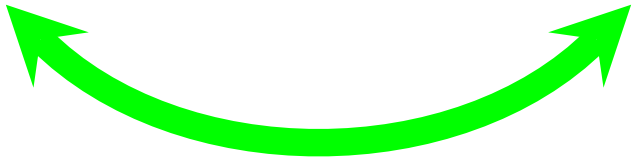
Author disambiguation



C. J. Smile Lin
National Taiwan Univ.
LIBSVM Guide



C. J. Cry Lin
Univ. of Michigan
LIBLINEAR Guide



Are they duplicates?

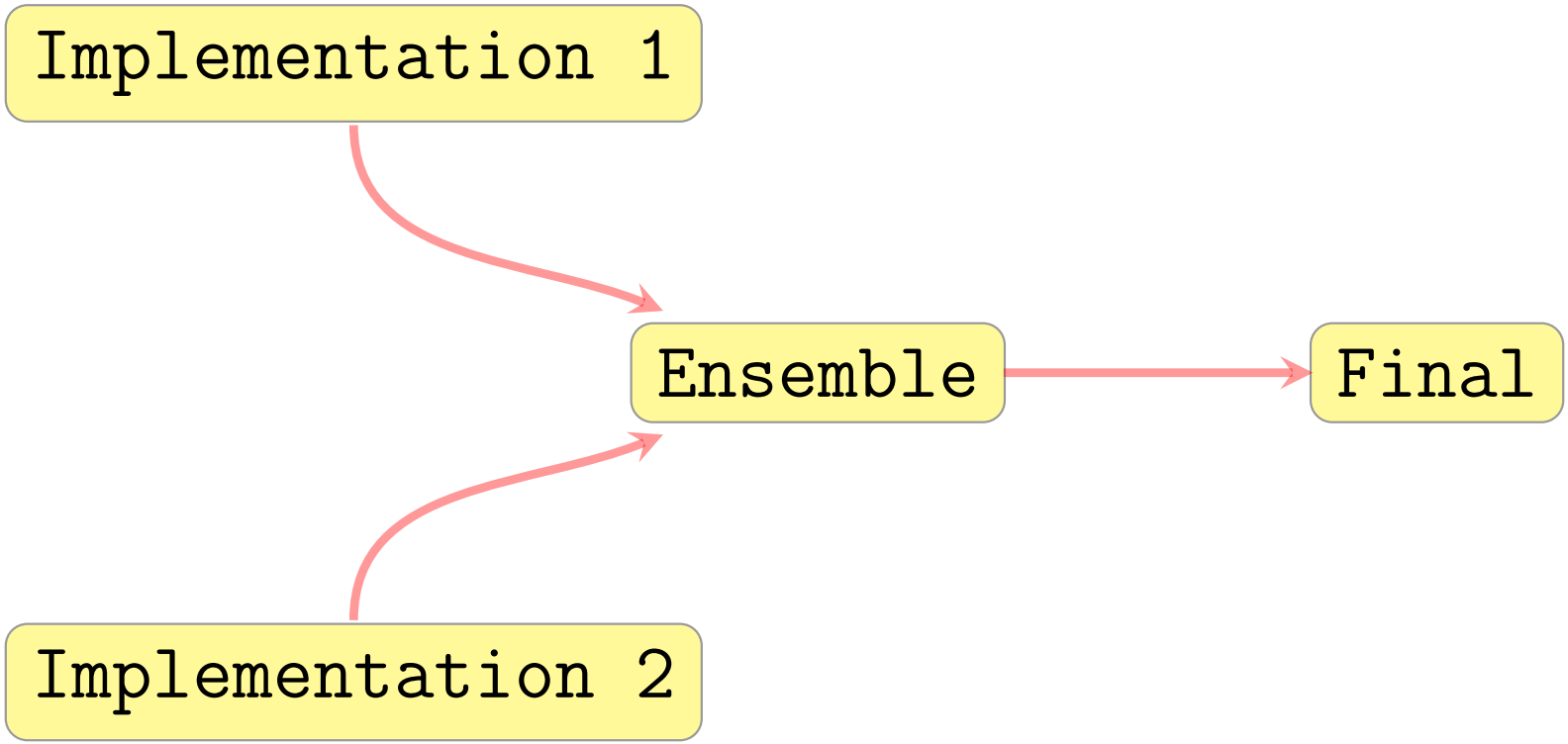
Outline

1. Introduction
2. Track 1: paper-author identification
   - Feature generation
   - Classification
   - Ensemble and post-processing
   - Results
3. Track 2: author disambiguation
   - Strategies and architecture
   - Implementation
   - Ensemble and typo handling
   - Analysis
4. Conclusions

Main Strategies

- Using string matching rather than other learning techniques
- An author without any papers is treated as a single group without duplicates
- Recognizing if an author is Chinese or not

Architecture



Results

Method	Public	Private
Implementation 1	0.99186	0.99198
Implementation 2	0.99071	0.99083
Final	0.99195	0.99202

Framework of the Two Implementations

1. **Cleaning**: remove redundant information
2. **Chinese-or-not**: classify each author as Chinese or non-Chinese
3. **Selection**: select a set of candidates of possible duplicates for each author
4. **Identification**: identify duplicates from the set of candidates for each author
5. **Splitting**: split incorrect cases (not discussed here)

Differences between Two Implementations

- The basic elements are different

  Implementation 1: author identifiers

  Implementation 2: author names

Author identifier	1001
Name in `Author.csv`	Chih Jen Lin
Names in `PaperAuthor.csv`	C. J. Lin Chih Jen Peter Lin C. J. P. Lin

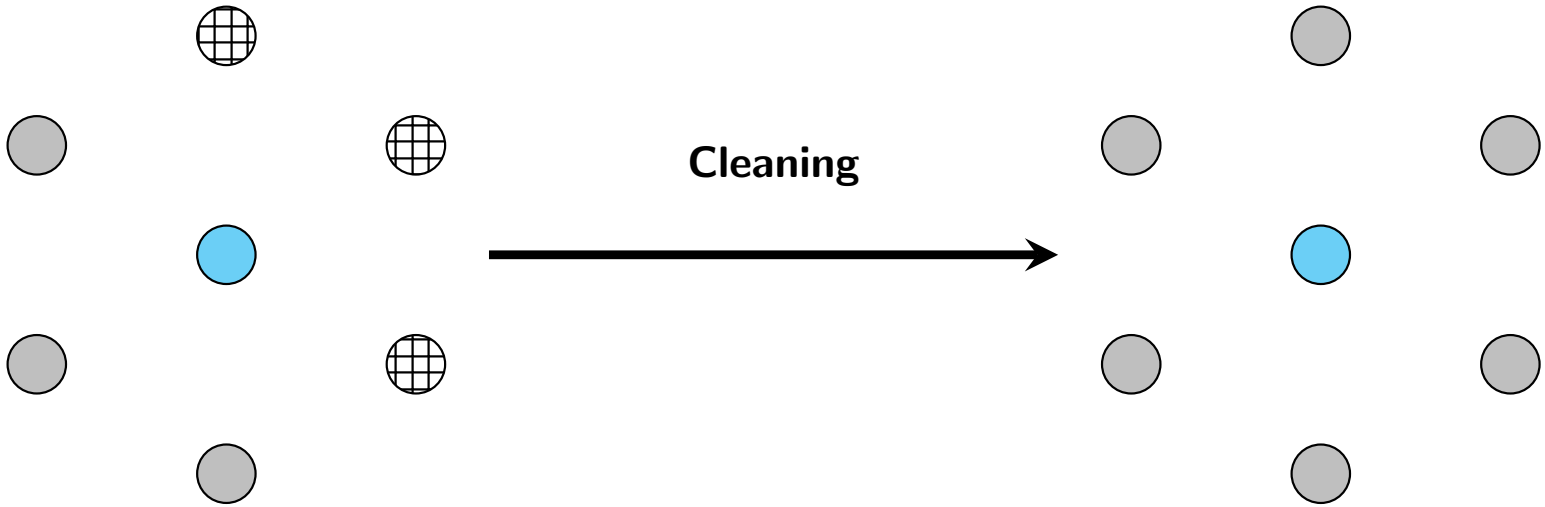
- More (complicated) rules in Implementation 1
- Focus on Implementation 1 because of time limitation

Outline

1. Introduction
2. Track 1: paper-author identification
   - Feature generation
   - Classification
   - Ensemble and post-processing
   - Results
3. Track 2: author disambiguation
   - Strategies and architecture
   - Implementation
   - Ensemble and typo handling
   - Analysis
4. Conclusions

Cleaning

- Clean redundant information.



- Examples:

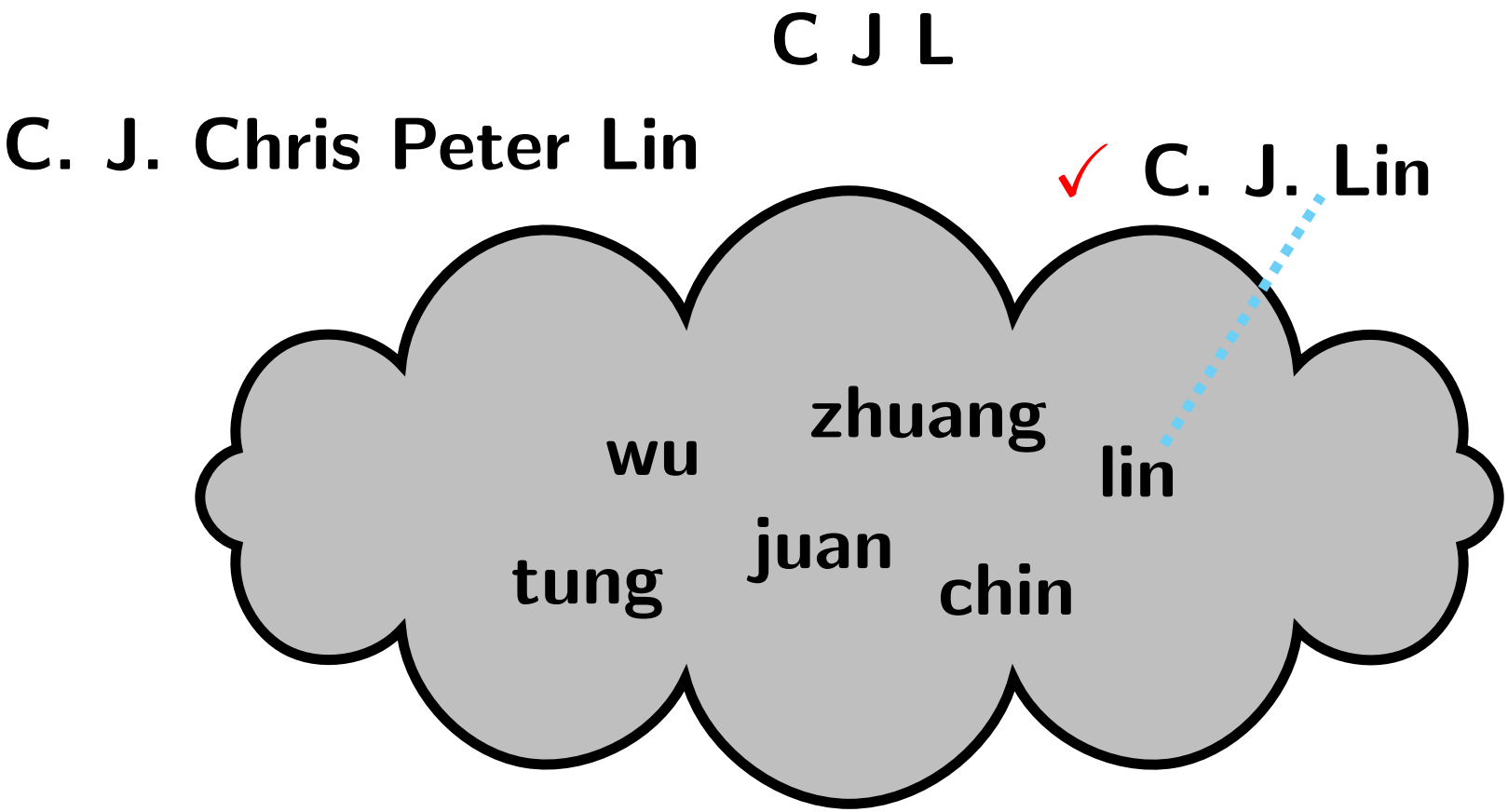
CHih JEN LIn	→	chih jen lin
Mr. Chih Jen Lin	→	chih jen lin
Chih Jén Lin	→	chih jen lin
Chih Jen Bill Lin	→	chih jen william lin

Chinese or not

Chinese and non-Chinese names are very different

- No middle name in Chinese. "Chih Lin" and "Chih J. Lin" are likely different
- Some Chinese last names like "Wang" are too common. Also, "林" and "藺" are romanized to "Lin"

Chinese or not (Cont'd)



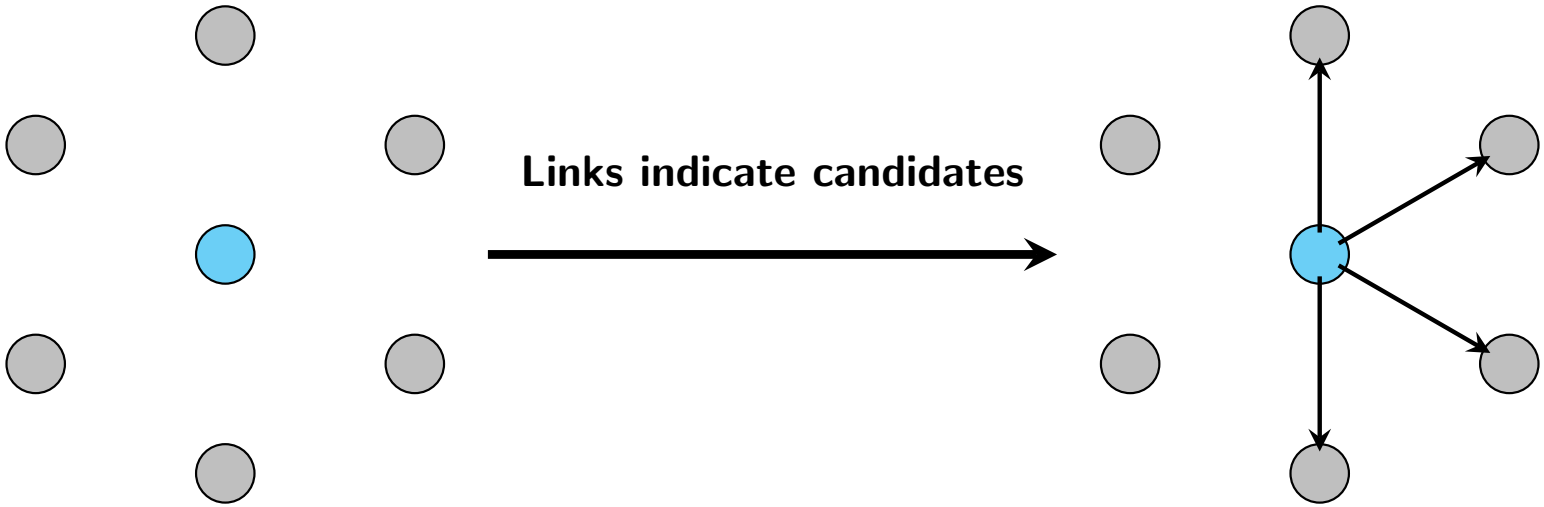
- Using common Chinese last names and words as a dictionary

Chinese or not (Cont'd)

- Check if the name contains words in our dictionary
- Examples:
  - Without full word → Non-Chinese; full word: a word without "." and longer than 1
    e.g., C J L
  - Only one full word and it is in Chinese dictionary → Chinese
    e.g., C. J. Lin
  - More than one full word not in Chinese dictionary → Non-Chinese
    e.g., C. J. Chris Peter Lin

Selection

- Find candidates of duplicates to reduce square complexity to linear in future comparison



- Each author generates several keys. "Chih Jen Lin" has:
  "Chih" "Jen" "Lin" "Chih Jen"
  "Jen Lin" "Chih Lin" "Chih Jen Lin"

Selection – Chih Jen Lin's candidates

- One is a candidate of another if two share the same key. Ignore common keys.

Selection – Chih Jen Lin's candidates

Peng Jen Chen

Chien Chih Wang

Shou De Lin

C. J. Lin

Felix Wu

Yu Chin Juan

Chih Jen Lin

Wei Sheng Chin

Yong Zhaung

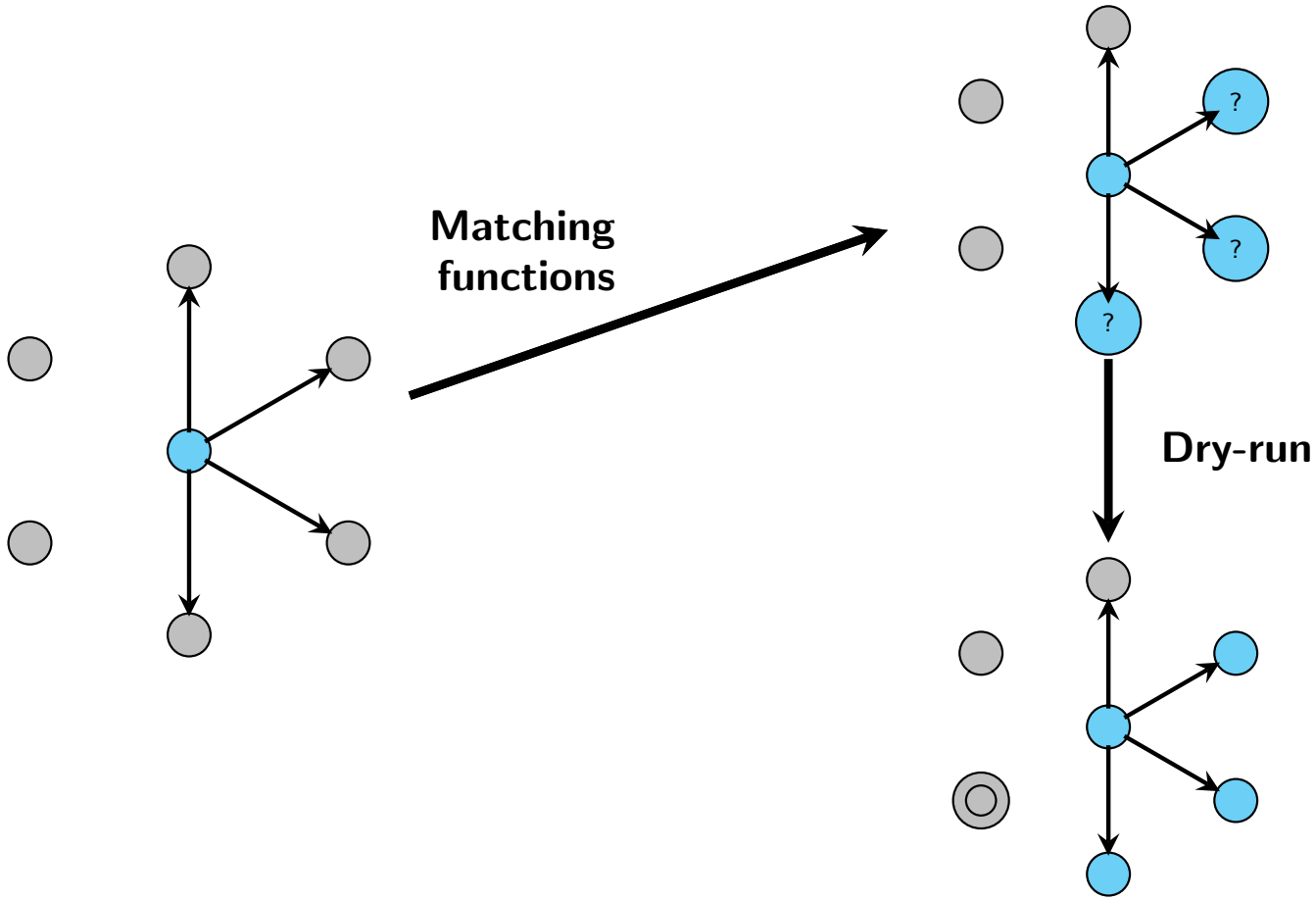
Lin Chih Jen

Hsiao Yu Tung

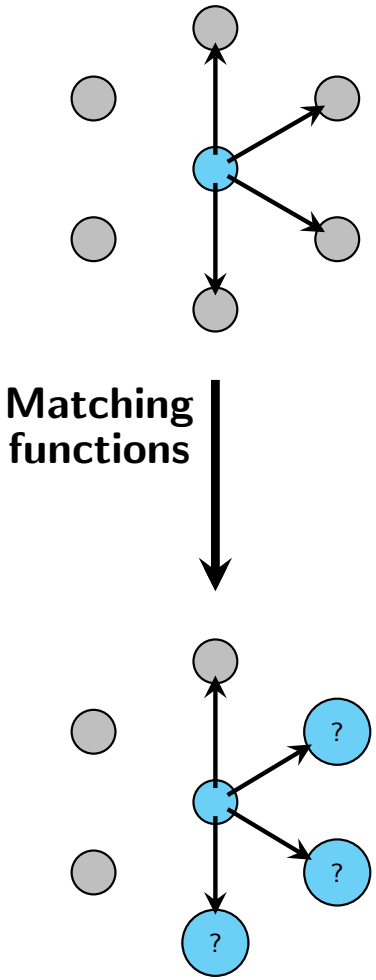
Hsuan Tien Lin

Identification

- Find duplicates from candidates



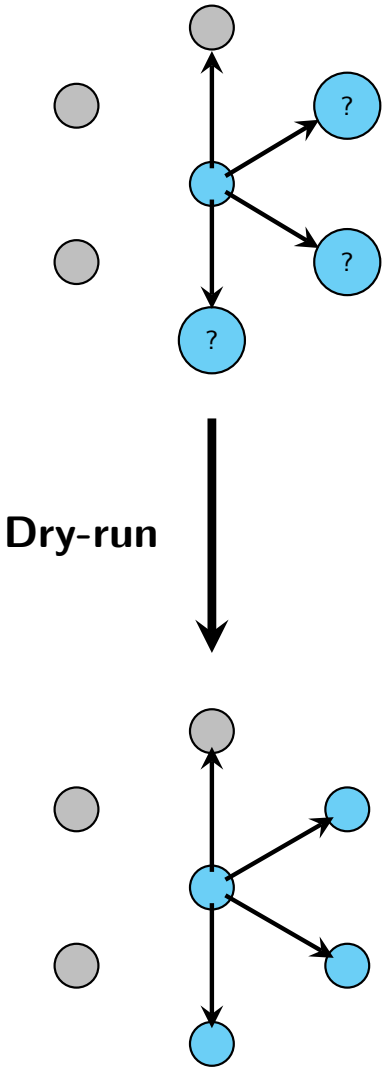
Identification – Matching Functions



- 13 matching functions
  1. Two authors have the same words
  2. (Non-Chinese only) Only one author has middle name and their last names differ in the last two characters
  3. ...
- Examples:

Two names	
Chih Jen Lin, Lin Chih Jen	Fun. 1
Michael I. Jordan, Michael Jordan	Fun. 2

Identification – Dry-run



- Making corrections as matching functions may wrongly identify duplicates
- Check if two names are "loosely identical"
- Examples:

Potential duplicates	Pass
C. J. Lin, Chih Jen Lin, Chih Lin, Chen Ju Lin	✗
C. J. Lin, Chih Jen Lin, Chih J. Lin	✓

Outline

1. Introduction
2. Track 1: paper-author identification
   - Feature generation
   - Classification
   - Ensemble and post-processing
   - Results
3. Track 2: author disambiguation
   - Strategies and architecture
   - Implementation
   - Ensemble and typo handling
   - Analysis
4. Conclusions

Ensemble

Method	Author identifier	Duplicates
Implementation 1	10	10,11
Implementation 2	10	10,11,12,13,14
Ensembled	10	10,11,12,14

- Implementation 1 considered as major predictions
- $\{12, 13, 14\}$ become additional duplicates
- Check if each of (10, 12), (10, 13), (10, 14) has similar affiliations or fields

Outline

1. Introduction
2. Track 1: paper-author identification
   - Feature generation
   - Classification
   - Ensemble and post-processing
   - Results
3. Track 2: author disambiguation
   - Strategies and architecture
   - Implementation
   - Ensemble and typo handling
   - Analysis
4. Conclusions

Analysis

- We conduct some analyses after the competition. Thank Kaggle for re-opening the submission site

Method	Public	Private
Final	0.99195	0.99202
Implementation 1	0.99186	0.99198
Without Chinese-or-not	0.99109	0.99125
Without dry-run	0.99097	0.99112
Without both	0.98891	0.98934

- Splitting Chinese/non-Chinese and the dry-run function in the identification stage are useful

Outline

1. Introduction
2. Track 1: paper-author identification
   - Feature generation
   - Classification
   - Ensemble and post-processing
   - Results
3. Track 2: author disambiguation
   - Strategies and architecture
   - Implementation
   - Ensemble and typo handling
   - Analysis
4. Conclusions

Conclusions

- Our code is available at
  - `github.com/kdd-cup-2013-ntu/track1`
  - `github.com/kdd-cup-2013-ntu/track2`
- Papers are at `www.csie.ntu.edu.tw/~cjlin/papers/kddcup2013/kddcup2013track1.pdf` and kddcup2013track2.pdf
- We thank the organizers and the support from National Taiwan University