Identifying Citing Sentences in Research Papers Using Supervised Learning

Kazunari Sugiyama* Tarun Kumar** Min-Yen Kan* Ramesh C. Tripathi**

* National University of Singapore
** Indian Institute of Information Technology, Allahabad, India

Introduction

Roles of Citing Sentences

"Shahabi and Chen [37] have pointed out that the item association generated from Web server logs might be wrong because Web usage data from server side are not reliable."

Acknowledgement of other's work

"Compared with our prior works [41,44], we scrutinize user's browsing history in one day closely and it allows each user to perform more fine-grained search by capturing changes of each user's preferences without any user effort."

Evidence for claims

Aim: "Identify citing sentences"

Application of citing sentences

Abstract

Users find it hard to delete unimportant personal Information which often results in cluttered workspaces. We present a full design cycle for GrayArea,

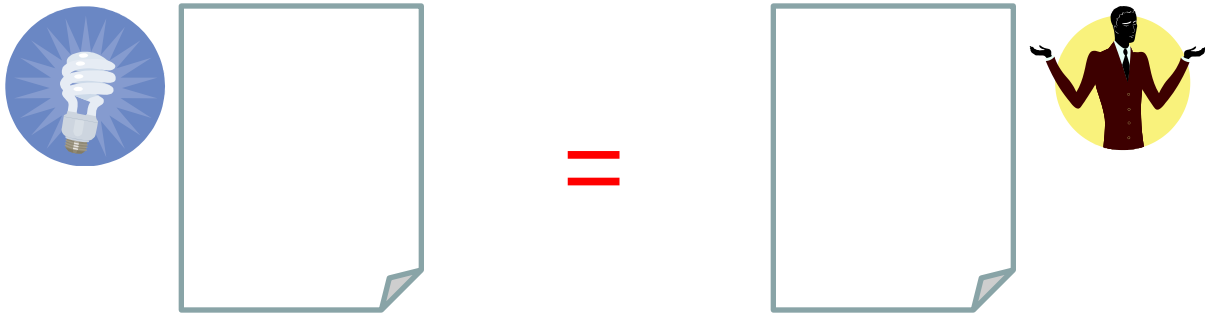
Introduction

Jones **[1]** claims that the decision whether “to keep or not to keep” information for future usage is prone to two types of costly mistakes: First information not kept is unavailable when it is needed later. Thus keeping irrelevant information not only causes guilt about being disorganized **[ref needed]** it also increases retrieval time.

Related Work

Citation Count

ISI impact factor



Recent works introduce PageRank to weight and control for the impact of papers

[Krapivin and Marchese, ICADL'08],
[Ma and Zhao, Information Processing and Management '08]
[Sayyadi and Getoor, SIAM Data Mining, '09]

Related Work

Citation Information

- **Bibliographic coupling**
  [Kessler, American Documantation '63]
- **Co-citation analysis**
  [Small, American Society of Information Science, '73]
- **Text summarization**
  [Teufel et al., EMNLP'06], [Qazvinian and Radev, Coling'08]
- **Thesaurus construction**
  [Schneider, PhD thesis, '94]
- **Information retrieval**
  [Ritchie et al., ECIR'07], [Ritchie et al., CIKM'08]

Related Work

Citation Information

Bibliographic coupling [Kessler, American Documantation '63]

Paper 1

References

[1] …

[2] …

[3] L. Page, et al. "The PageRank Citation Ranking Bringing Order to the Web" Technical Report …

[4] …

[5] G. Salton and M. J. McGill. "Introduction to Modern Information Retrieval" McGraw-Hill, 1983.

[6] …

[7] …

Paper 2

References

[1] …

[2] L. Page, et al. "The PageRank Citation Ranking Bringing Order to the Web" Technical Report …

[3] …

[4] …

[5] …

[6] …

[7] …

[8] …

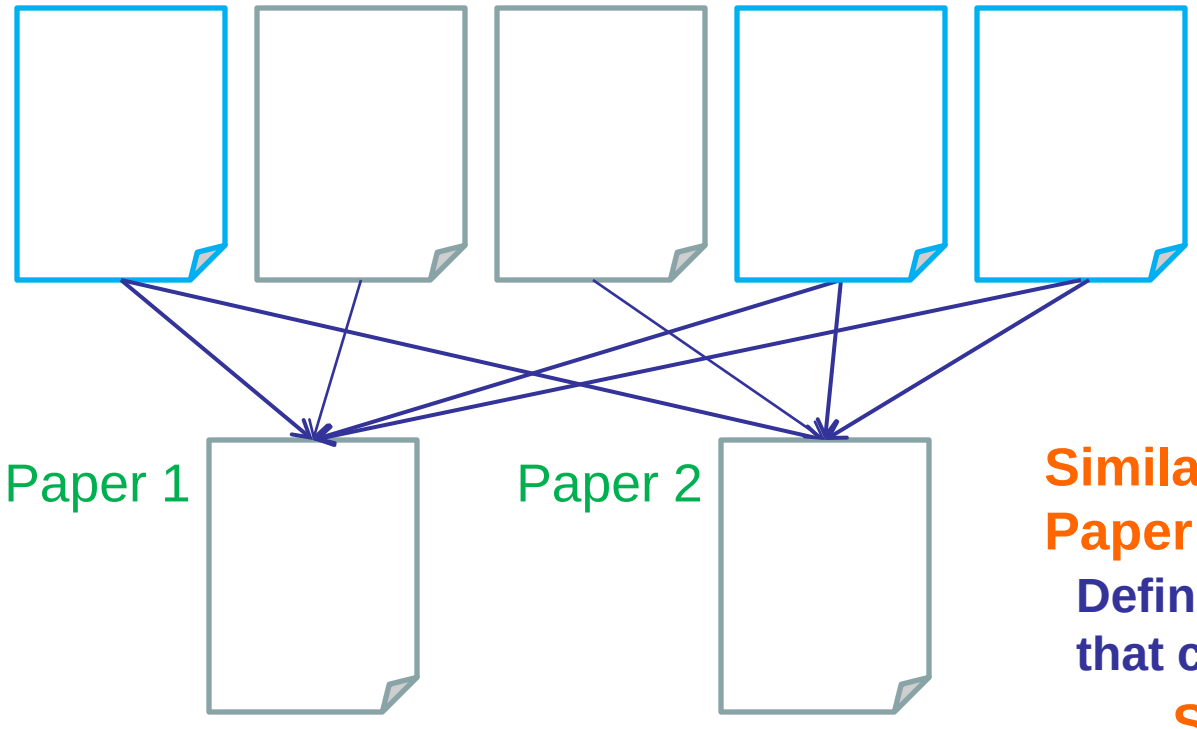
[9] G. Salton and M. J. McGill. "Introduction to Modern Information Retrieval" McGraw-Hill, 1983.

These two papers are coupled.

Related Work

Citation Information

Co-citation analysis [Small, American Society of Information Science, '73]

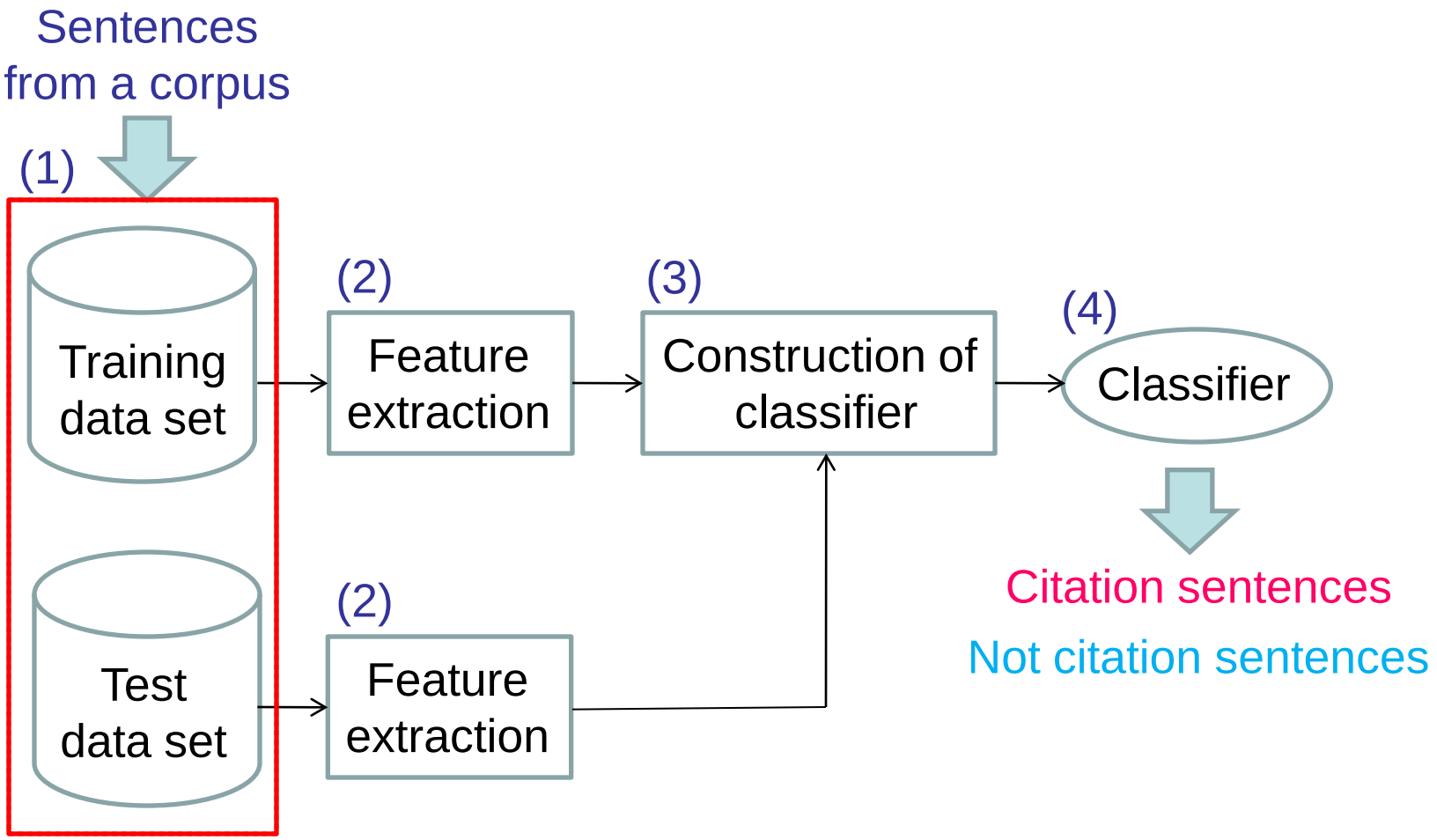


Similarity between Paper 1 and 2, Sim(P1, P2):

Defined by the number of papers that cites both Paper 1 and 2

Sim(P1, P2) = 3

Proposed Method



(1) Constructing training and test data sets

- **Remove stop words and reference section from each paper**
- **Define the classes of each instance:**

Positive instance

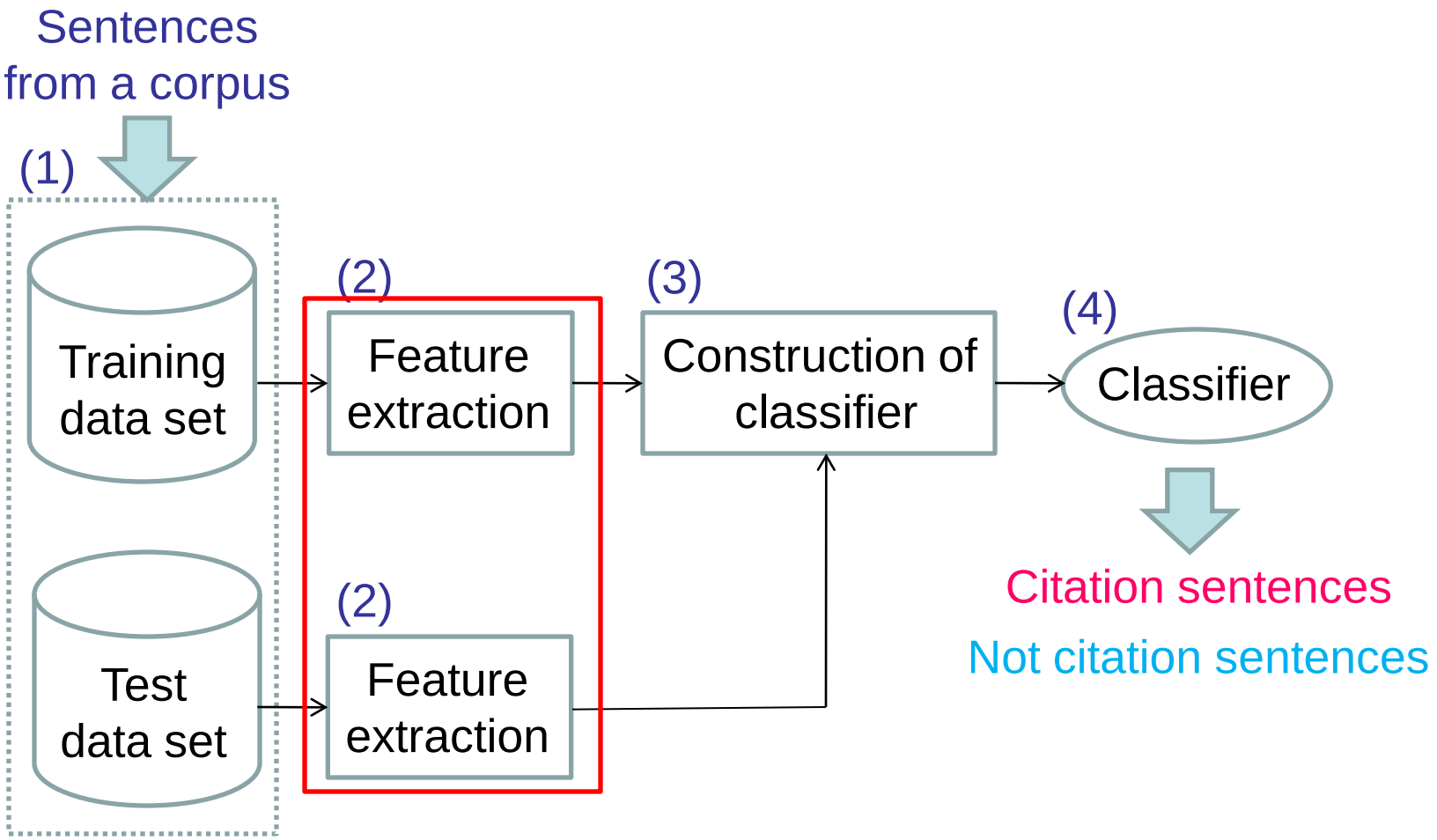
"In order to overcome the above shortcomings of impact factor, Sun and Giles [37] noted that popular venues are more influential in impact factors."

Citation is required !

Negative instance

"As we perform 10-fold cross validation, we divided the whole dataset into 90% of training data and 10% of test data."

Proposed Method



(2) Extracting features

- Unigram
- Bigram
- Proper nouns
- Previous and next sentence
- Position
- Orthographic

- All features above

(2) Extracting features: unigram, bigram

[Example sentence]

"In order to overcome the above shortcomings of impact factor, Sun and Giles [34] noted that popular venues are more influential in impact factors."

- **Unigram**

  "order," "overcome," "shortcoming," "impact,"
  "factor," "Sun," "Giles," …

- **Bigram**

  "order overcome," "overcome shortcoming," "shortcoming impact,"
  "factor Sun," "Sun Giles," …

(2) Extracting features: proper nouns

[Example sentence]

"In order to overcome the above shortcomings of impact factor, Sun and Giles [34] noted that popular venues are more influential in impact factors."

- **Proper nouns**

  "Sun," "Giles"

  Binary feature (unlike unigram and bigram)

  The feature for the example sentence: "hasProperNoun"

(2) Extracting features: previous and next sentence

[Example sentence]

"This notion of citation count is widely used in evaluating the importance of a paper because it has been shown to strongly correlate with academic document impact [23]. The Thomson Scientific impact factor (ISI IF) is the representative approach using citation count [10]. The advantages of citation count are (1) its simplicity of computation; and (2) that it is proven method which has been used for many years."

- **Previous and next sentence**

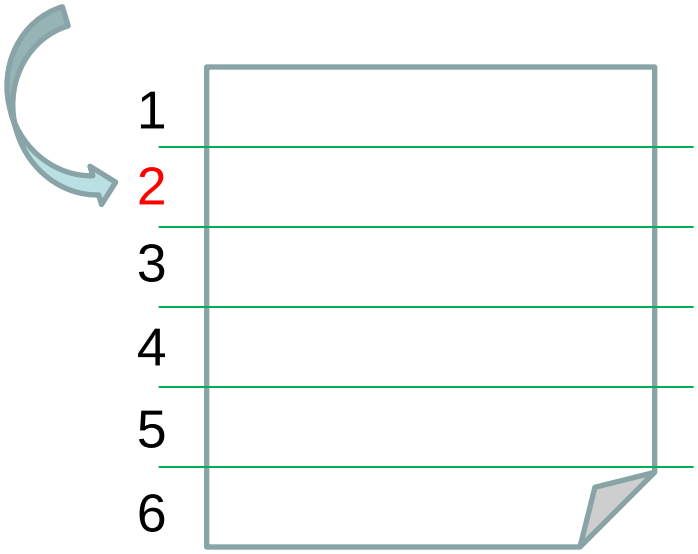
  The feature for the example sentence:

  "prevSentHasCite_NextSentHasNoCite"

(2) Extracting features: Position

[Example sentence]

"In order to overcome the above shortcomings of impact factor, Sun and Giles noted that popular venues are more influential in impact factors."



Divide a paper into
6 equal parts

The position for the example sentence:
"Position_2"

(2) Extracting features: Orthographic

[Example sentence]

"In order to overcome the above shortcomings of impact factor, Sun and Giles noted that popular venues are more influential in impact factors."

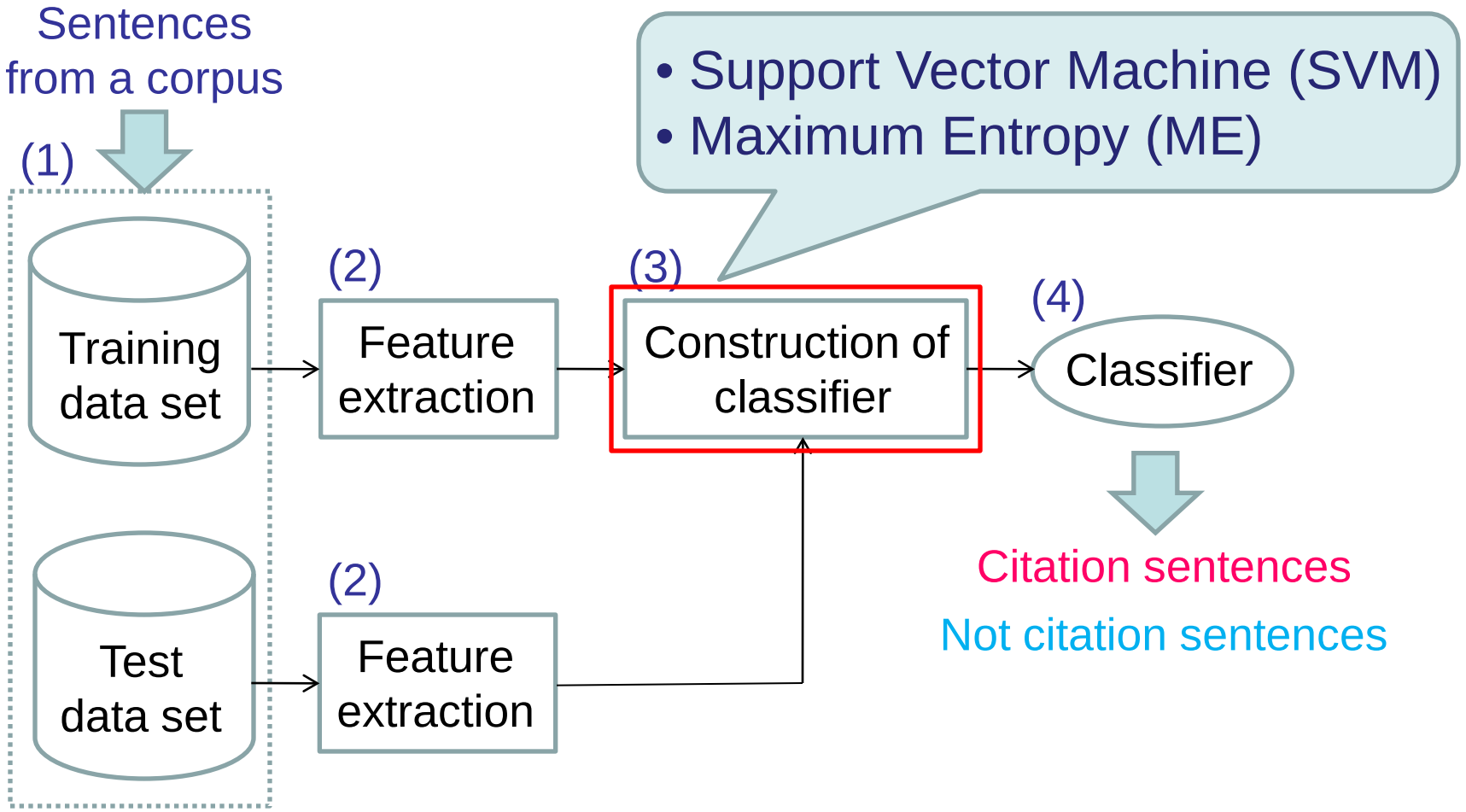
- **Orthographic**

  Numbers

  Capitalizations on any word.

  The orthographic for the example sentence:
  "SentHasNoNum_SentHasCap"

Proposed Method



Classification Approaches

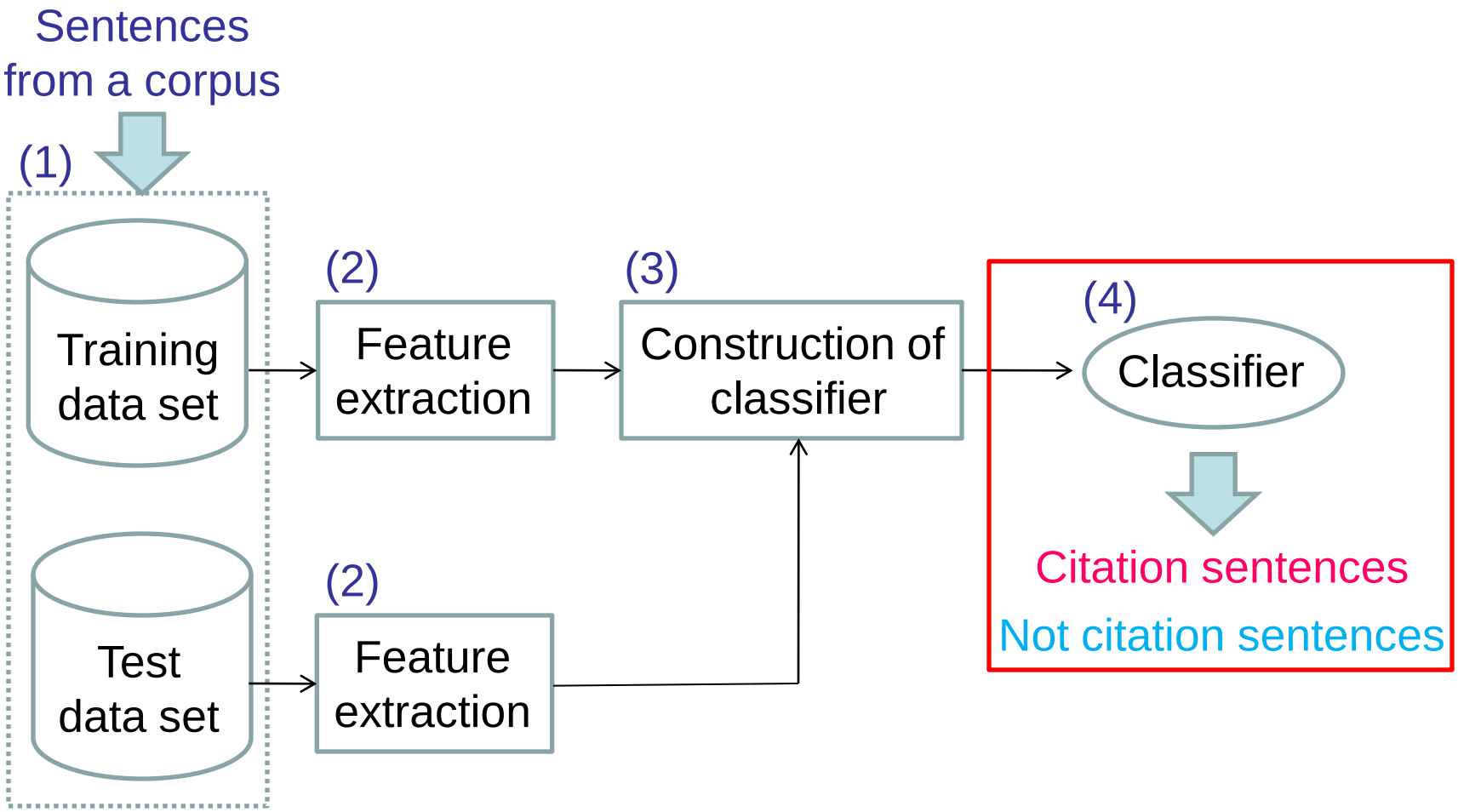
Support Vector Machine (SVM)

- Often brings superior results in binary classification problem

Maximum Entropy (ME)

- Successfully applied to various fields in natural language processing task

Proposed Method



Experiments

Experimental Data

- **ACL Anthology Reference Corpus (ACL ARC)**
  - 10,921 papers
  - 955,755 sentences
    - 112,533 sentences contain citing information
    - 843,242 sentences do not contain citing information

Evaluation Measure

- **Accuracy**

$$\text{Accuracy} = \frac{\text{(Number of correct classifications)}}{\text{(Total number of test cases)}}$$

Classification Accuracy by 10-fold Cross Validation

Feature	Accuracy (ME)	Accuracy (SVM)
(1) Unigram	0.876	0.879
(2) Bigram	0.827	0.851
(3) Proper Noun	0.882	0.882
(4) Previous and Next Sentence	0.882	0.882
(5) Position	0.875	0.877
(6) Orthographic	0.878	0.880
(7) All [(1) - (6)]	0.876	0.878

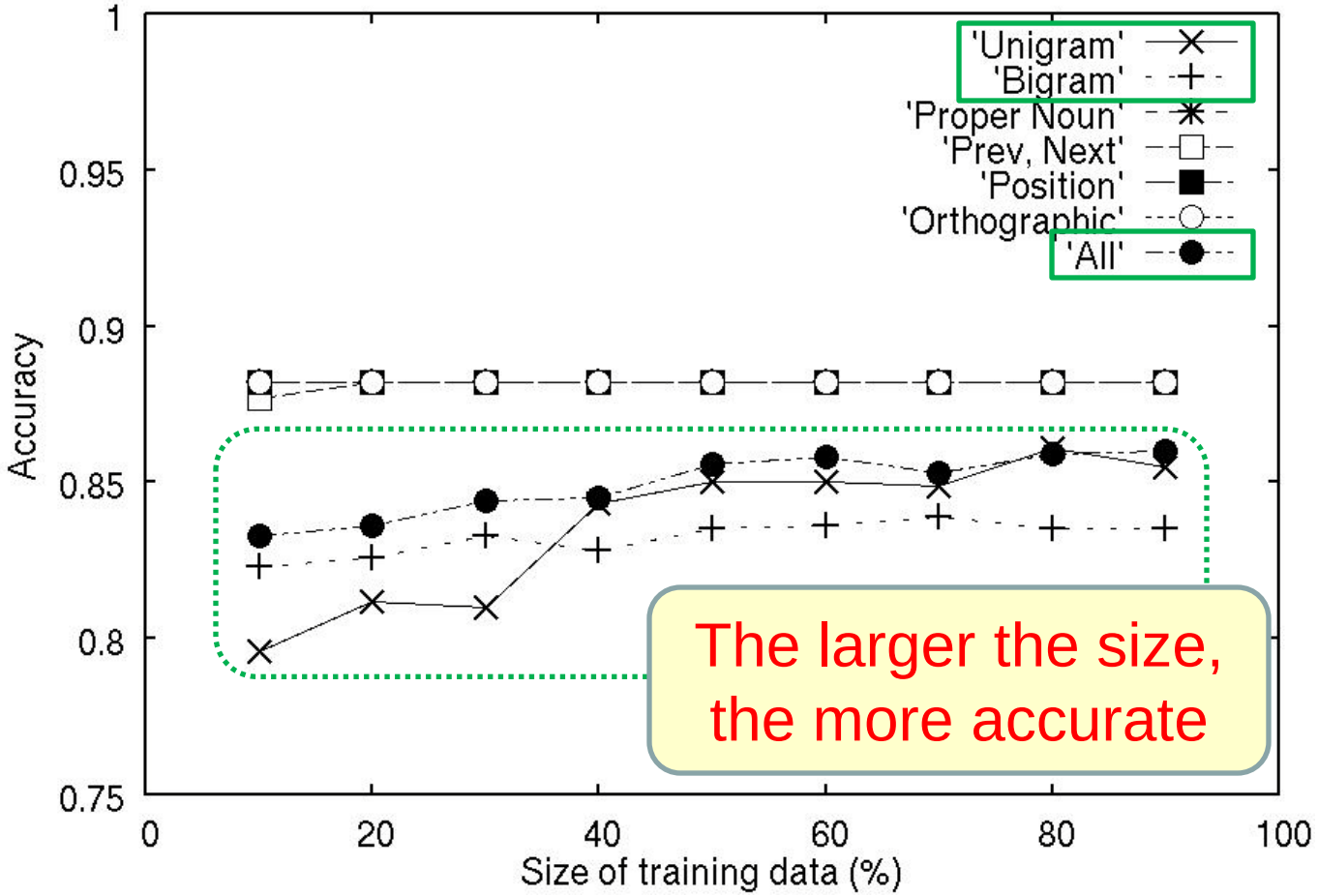
The best

Classification Accuracy by 10-fold Cross Validation

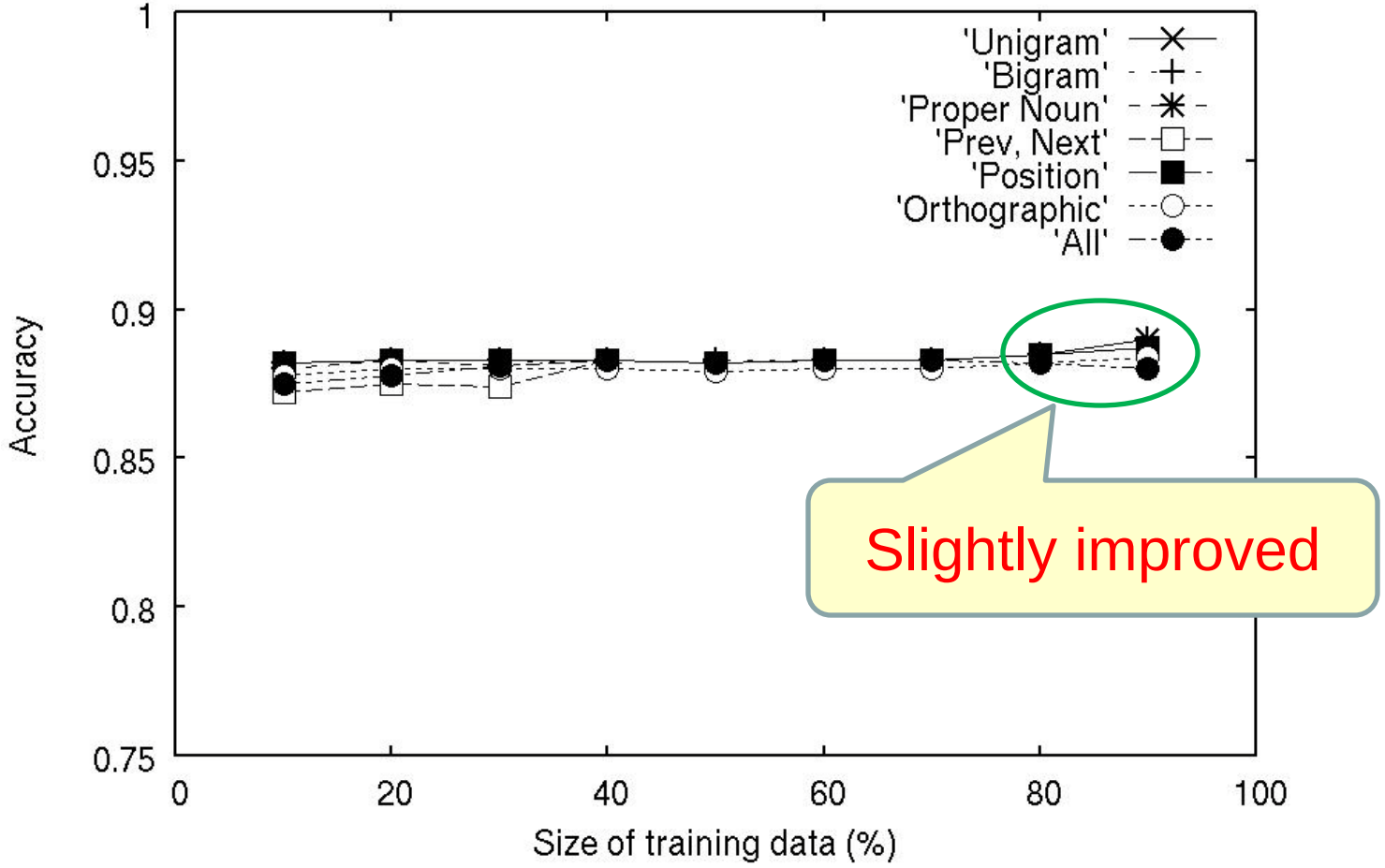
Feature	Accuracy (ME)	Accuracy (SVM)
(1) Unigram	0.876	0.879
(2) Bigram	0.827	0.851
(3) Proper Noun	0.882	0.882
(4) Previous and Next Sentence	0.88	
(5) Position	0.8	
(6) Orthographic	0.878	0.880
(7) All [(1) - (6)]	0.876	0.878

Not effective due to sparse data

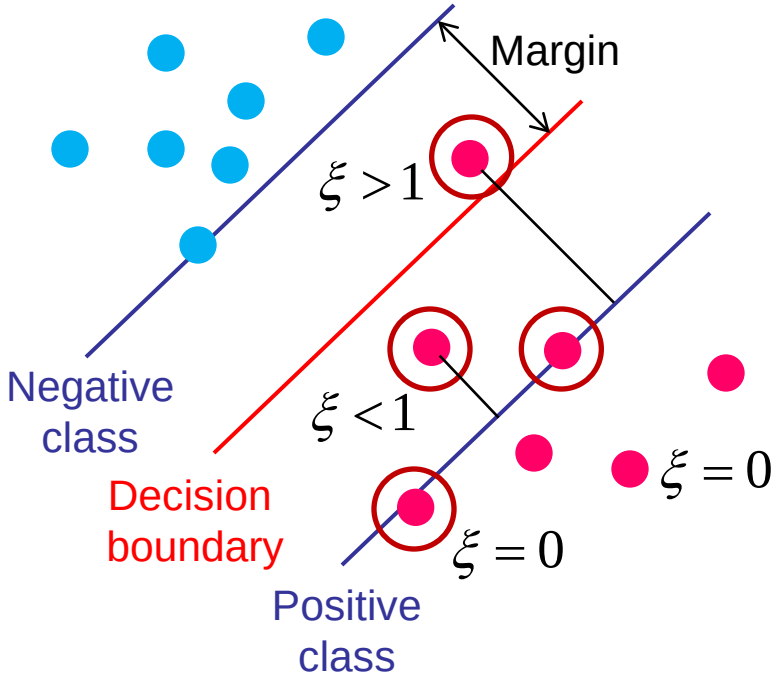
Classification Accuracy on Different Size of Training Data (ME)



Classification Accuracy on Different Size of Training Data (SVM)



Value of "*C*" in SVM



Goal

Maximize the margin while softly penalizing points that lie on the wrong side of the margin boundary

Minimize: $\frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{N}\xi_i$

C (> 0)

Control the trade-off between the slack variable penalty and the margin

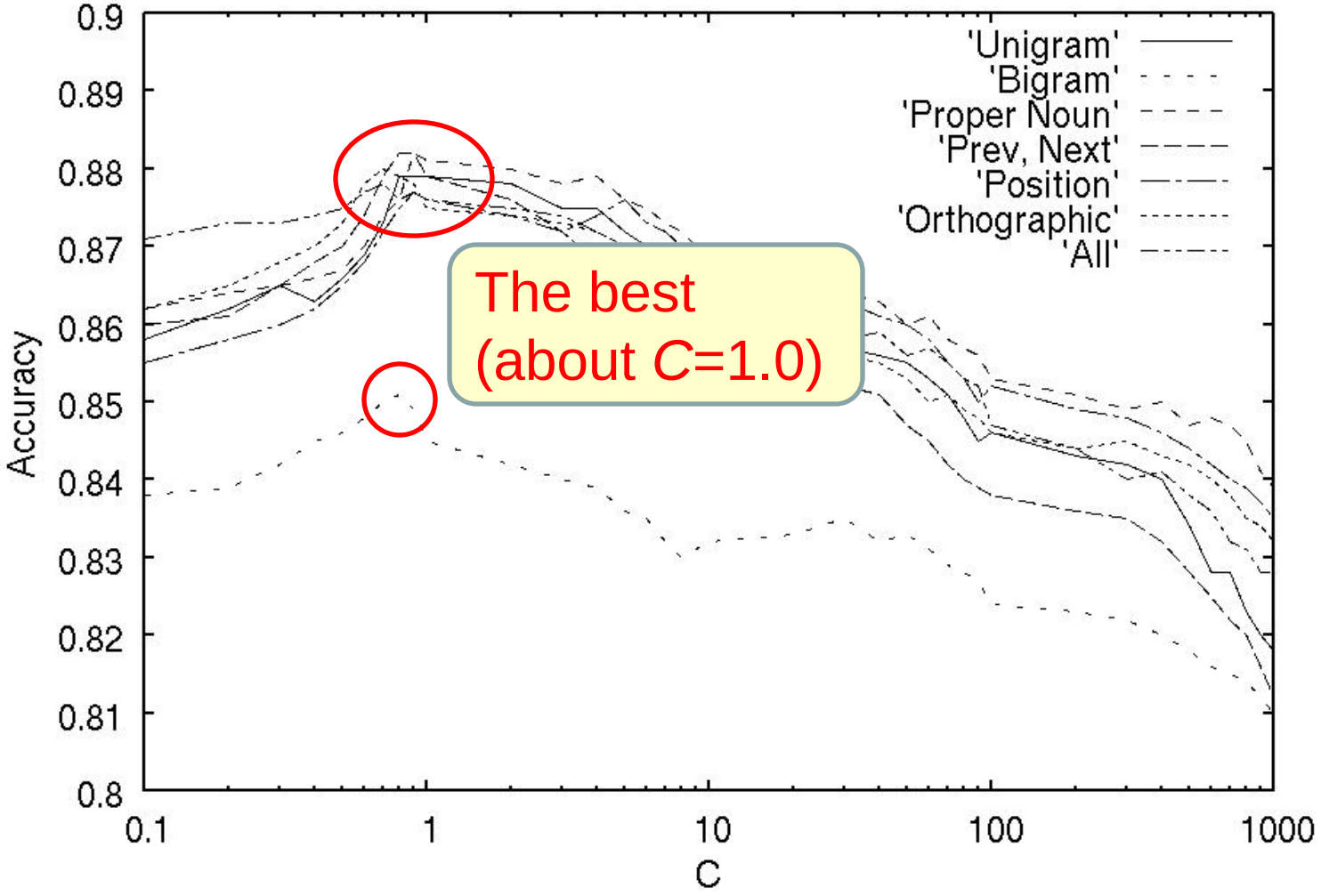
Data points

$\xi = 0$: Correctly classified

$0 < \xi \leq 1$: Lie inside the margin on the correct side of decision boundary

$\xi > 1$: Wrong side of the decision boundary

Classification Accuracy in Different Value of “C”



Classification Accuracy in Different Value of "*C*"

- **Optimal Value of "*C*" in Each Feature**

Feature	Optimal Value of "*C*"	Accuracy
(1) Unigram	0.8	0.879
(2) Bigram	0.8	0.851
(3) Proper Noun	0.9	0.882
(4) Previous and Next Sentence	0.9	0.882
(5) Position	0.9	0.877
(6) Orthographic	0.9	0.880
(7) All [(1) - (6)]	0.9	0.878

The best

Future Work

Technical point:

- **Analyze the detailed cause of slight improvement in varying the number of training data in SVM**

The larger picture:

- **Build an editor that will help authors write a research paper when the sentences in the paper need a citation or not.**

***Thank you very much*!**