Department of Computer Science and Engineering
Jahangirnagar University

Retrieving Top *K*% Relevant Patterns for Distant Supervision-Based Relation Extraction for Bangla Sentences

Presented by

Nishat Tasnim
Exam. Roll: 220545
MSc. Session: 2021-2022

Supervised by

Dr. Md. Musfique Anwar
Professor
Department of CSE

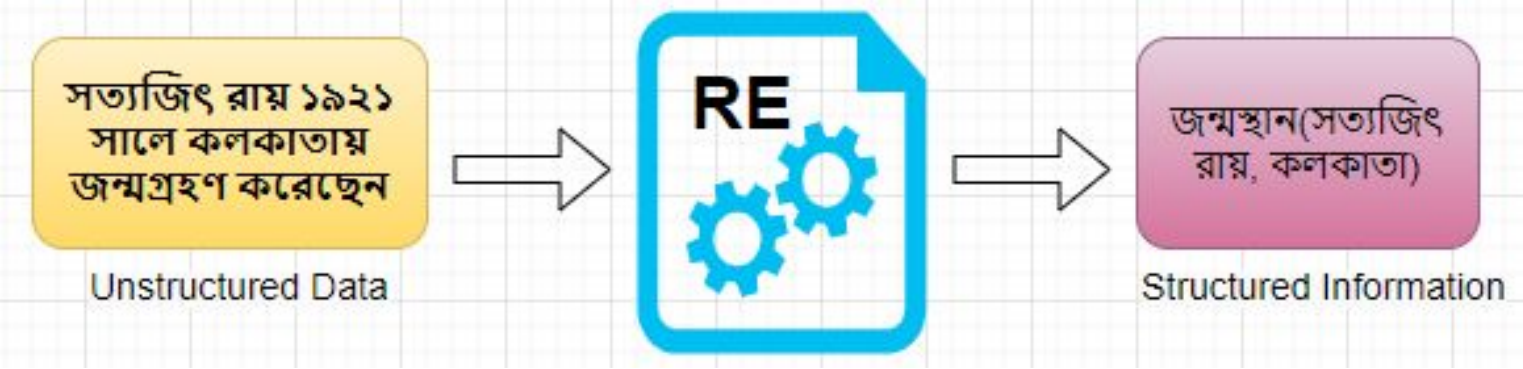
Tuesday, 17 September, 2024

Slide Outline

- Introduction
- Motivation and Objective
- Literature Review
- Contribution
- Methodology
- Proposed Algorithm
- Experimental Results
- Conclusion & Future Scope

Introduction

Relation extraction (RE) is a natural language processing (NLP) task that involves identifying relationships between entities in a text.



Motivation and Objective

The rapid growth of unstructured online content necessitates converting it into structured data through semantic annotations.

Manual annotation is impractical due to volume and diversity, particularly for low-resource languages like Bangla.

Distant supervision using knowledge bases (e.g., Freebase, DBpedia) automates dataset generation but **introduces noise due to oversimplified assumptions about relations.**

Motivation and Objective

A Sample **KB**,

Entity 1	Entity 2	Relation Name
হুমায়ুন আহেমদ	নেত্রকোনায়	জন্মস্থান

There are some sentences in the corpus, labeled based on DS assumption:

Sentence	Relation Name
হুমায়ুন আহেমদ **জন্মগ্রহণ করেছেন** নেত্রকোনায়	জন্মস্থান
হুমায়ুন আহেমদ বেড়াতে গিয়েছিলেন নেত্রকোনায়	জন্মস্থান
হুমায়ুন আহেমদ নেত্রকোনায় শুটিং এর কাজে গিয়েছিলেন	জন্মস্থান

The trained model may retrieve some wrong instances like (খান আতাউর রহমান, মুম্বই) for place of birth (জন্মস্থান) relation for this sentence: “খান আতাউর রহমান মুম্বই বেড়াতে গিয়েছিলেন”.

To make an accurate annotation. It is important to extract valid patterns for any relation from the text corpus.

Motivation and Objective

The objectives of my work are:

- **Create a structural knowledge base and annotated corpora** for Bangla.
- **Extract relevant patterns of relations** from Bangla sentences to improve the performance of DS-based RE.
- Improving information retrieval and enhancing semantic understanding

Literature Review

Table 1. Related Works

Reference	Problem Domain	Approach	Key Findings	Limitations
Castelli et al., 2014 [1]	Relation Extraction (RE) from Arabic Text	Supervised feature-based classifiers, supervised kernel-based classifiers and using semi-supervised methods	The paper addresses the language-specific difficulties that arise when extracting relations from Semitic languages, such as the lack of diacritics and the challenges posed by complex morphology	Need labeled data, which is very costly and time consuming.
Manzoor et al., 2021 [2]	Relation Extraction (RE) from English Text	Unsupervised RE approach using SBERT-based sentence encoding.	Used clustering to group similar sentences for relation extraction, with a confidence threshold to improve accuracy and prevent semantic drift.	Unsupervised models tend to extract many irrelevant relations, resulting in noisy outputs

Literature Review

Table 1. Related Works (Cont.)

Reference	Problem Domain	Approach	Key Findings	Limitations
Mintz et al., 2009 [3]	Relation Extraction from English Text	Distant Supervision-based RE combining lexical & syntactic features	The distant supervision algorithm extracts high-precision patterns, and combining syntactic and lexical features enhances performance.	DS introduces noisy patterns, which negatively impact overall performance.
Augenstein et al., 2014 [4]	Relation Extraction (RE) from English Text	Distant supervision-based RE using statistical methods for targeted training data selection.	This paper focused on enhancing entity recognition across domains, extracts relations across sentence boundaries, and reduces noise.	Only focused on NER, not on noisy patterns.

Literature Review

Table 1. Related Works (Cont.)

Reference	Problem Domain	Approach	Key Findings	Limitations
Mahfuz et al., 2020 [4]	Relation Extraction from Bangla Text	Distant Supervision-based RE based on lexical features	Introduce a strategy for removing noisy patterns for DS-based RE based on conflict scores for any relation	A noisy pattern matching an entity pair in the knowledge base cannot be filtered out using this approach.

Contribution

Improve accuracy and reliability of relation extraction from Bangla text using distant supervision.

Approach:

- **Pattern Selection**: Retrieve top K% valid patterns for each relation based on conflict scores.
- **Filtering**: Further refine patterns using probability scores; relabel noisy patterns as "NONE."
- **Model Training**: Use an ensemble method to train the model for enhanced performance.

Methodology

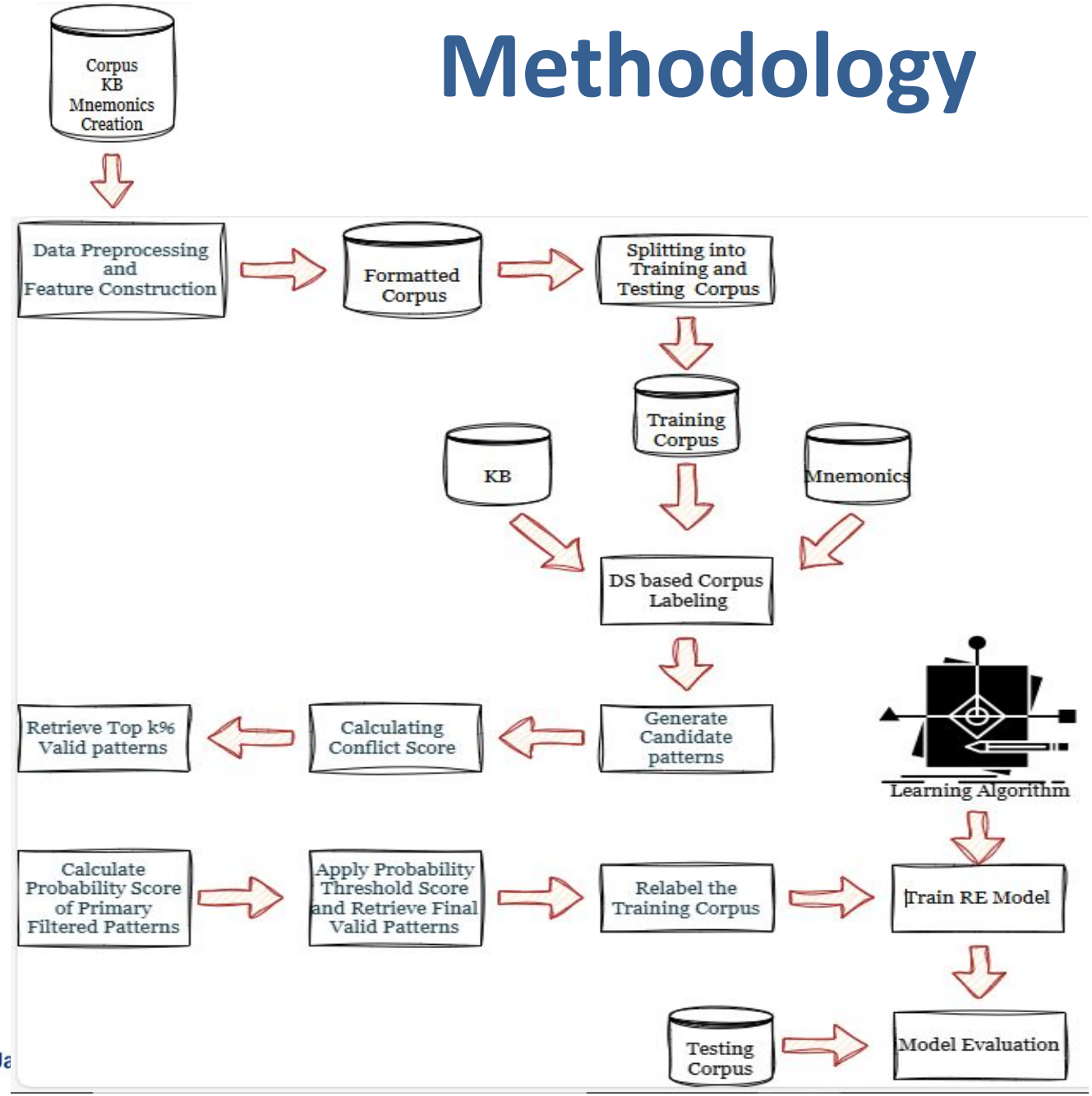


Fig 1: Proposed Methodology

Dataset Collection

Knowledge Base (KB) creation:

The KB data is sourced from Bangla **Wikidata**, offering a rich, structured dataset of entities and relationships. We extract relevant data using **SPARQL** queries.

Table 2 presents the Knowledge Base we utilized, detailing its size along with an example of each relation.

Dataset Collection(Cont.)

Relation Name	Size	Example
চলচিত্র অভিনেতা (Movie Actor)	38064	টম কার্টিস, দ্য কেজ
জন্মস্থান (Place of Birth)	13298	আতুল চিল্রিস, বার্লিন
চলচিত্র পরিচালক (Movie Director)	6164	সিডনি লুমেট, টুয়েলভ অ্যাংরি মেন
লেখক (Writer)	2630	সত্যজিৎ রায়, জয় বাবা ফেলুনাথ
মৃত্যুস্থান (Place of Death	1808	সত্যজিৎ রায়, কলকাতা
প্রতিষ্ঠানের অবস্থান (Company Location)	1011	সপ্তম ফটো এজেন্সি, নিউ ইয়র্ক সিটি
প্রতিষ্ঠাতা (Company Founder)	281	সপ্তম ফটো এজেন্সি, ক্রিস্টোফার মরিস

Table 2: Knowledge Base with the size of each relations

Dataset Collection(Cont.)

Corpus Creation: We constructed a corpus of **90,441** text collecting from Wikipedia and other online resources.
Table 3 shows sample corpus:

Sample Text	Relation Name
হুমায়ুন আহমেদ হলুদ হিমু লিখেছেন।	লেখক
আতুল চিল্লিস এর জন্মস্থান বার্লিনে	জন্মস্থান
আহমদ ছফা ঢাকায় শেষ নিশ্বাস ত্যাগ করেছেন।	মৃত্যুস্থান
জেফ্রি হান্টার দ্য কেজ সিনেমায় অভিনয় করেছে	চলচিত্র অভিনেতা
হুমায়ুন আহমেদ দুই দুয়ারি মুভির পরিচালক	চলচিত্র পরিচালক
ক্রিস্টোফার মরিস সপ্তম ফটো এজেন্সি প্রতিষ্ঠানের প্রতিষ্ঠাতা	প্রতিষ্ঠাতা
রতন টাটা পত্তন করেছিলেন টাটা প্রতিষ্ঠানের	প্রতিষ্ঠানের অবস্থান

Table 3: Sample Corpus

Dataset Collection(Cont.)

Mnemonics Creation:

Mnemonics are powerful cognitive tools that aid in memory retention and recall by creating associations between information and easily memorable cues.

For instance, in my Knowledge Base, there is a seed instance (সত্যজিৎ রায়, কলকাতা). However, in my corpus, there is a sentence like " সত্যজিৎ রায় ভারতে জন্মগ্রহন করেছেন". According to the distant supervision method, the sentence is labeled with a 'None' relation. However, since কলকাতা (Kolkata) is a city in ভারত (India), the sentence should be labeled with the জন্মস্থান (place of birth) relation. This is where our mnemonics help.

We have developed mnemonics for **440** locations.

Table 4 shows sample mnemonics of location entity.

Dataset Collection(Cont.)

Small City	Big City	Country
সাগরদাড়ি	যশোর	বাংলাদেশ
স্ট্র্যাটফোর্ড-আপন-অ্যাভন	ওয়ারউইকশায়ার	ইংল্যান্ড
তাম্বুলখানা	ফরিদপুর	বাংলাদেশ
কলকাতা	পশ্চিম-বঙ্গ	ভারত
রায়পুরা	নরসিংদী	বাংলাদেশ
সান্টা ক্লারা	ক্যালিফোর্নিয়া	মার্কিন যুক্তরাষ্ট্র
টালাহাসি	ফ্লোরিডা	মার্কিন যুক্তরাষ্ট্র

Table 4: Sample Mnemonics

Data preprocessing and Feature Construction

- Preprocessed text documents and applied **Named Entity Recognition (NER)** to identify entities (e.g., person, location, organization).
- Constructed feature vectors using a window of **k words** between and around entities, along with **Part-of-Speech (POS)** tags.
- Extracted patterns from sentences based on these features.
- Split the dataset into **70:30** for training and testing.

Table 5 shows sample texts with **NER** and **POS** tagging, while Table 6 highlights the extracted lexical features.

Data preprocessing and Feature Construction (Cont.)

Formatted Sentence
পল্লীকবি/NP জসীম উদ্দীন/PER ফরিদপুর/LOC জেলায়/NC জন্মগ্রহণ/NC করেছেন/VM
আহমদ সফা/PER ঢাকায়/LOC শেষ/JJ নিঃশ্বাস/NC ত্যাগ/NC করেছেন/VM
কবিগুরু/NP রবীন্দ্রনাথ ঠাকুর/PER রচিত/VM ভিখারিনী/BOOK গল্পটি/NC বাংলা/VM সাহিত্যের/NX প্রথম/JQ ছোটগল্প/JJ
জেফ্রি হান্টার/PER দ্য কেজ/MOV সিনেমায়/NC অভিনয়/NC করেছেন/VM
হুমায়ুন আহমেদ/PER দুই দুয়ারি/MOV মুভির/NC পরিচালক/NC
সপ্তম ফটো এজেন্সি/ORG প্রতিষ্ঠানটির/NC অবস্থান/NC নিউ ইয়র্ক সিটিতে/LOC
ক্রিস্টোফার মরিস/PER সপ্তম ফটো এজেন্সি/ORG প্রতিষ্ঠানের/NC প্রতিষ্ঠাতা/NC

Table 5: Sample texts with NER and POS tagging

Data preprocessing and Feature Construction (Cont.)

Left Window	Entity1	Middle Win-dow	Entity2	Right Window
[পল্লীকবি/NP]	PER	[ ]	LOC	[জেলায়/NC জন্মগ্রহণ/NC করেছেন/VM]
[ ]	PER	[প্রাতিষ্ঠানিক/JJ শিক্ষা/NC শুরু/NC হয়েছিলো/VM]	LOC	[ ]
[ ]	PER	[ ]	MOV	[চলচিত্রটি/NC পরিচালনা/NC করেছেন/VM]
[কবিগুরু/NP]	PER	[রচিত /□□]	BOOK	[গল্পটি/NC বাংলা/NP সাহিত্যের/NC প্রথম/JQ ছোটগল্প/JJ]
[ ]	PER	[রচিত/VM]	BOOK	[একটি/JQ গোয়েন্দা/NC উপন্যাস/NC]

Table 6: Lexical Features Extracted from Sample Sentences

DS based Training Corpus Labelling

পল্লীকবি/NP জসীম উদ্দীন/PER ফরিদপুর/LOC জেলায়/NC জন্মগ্রহণ/NC করেছেন/VM

Table 7: A sample formatted text

Entity 1	Entity 2	Relation Name
জসীম উদ্দীন	ফরিদপুর	জন্মস্থান
...	...	...
রবীন্দ্রনাথ ঠাকুর	কলকাতা	মৃত্যুস্থান

Table 8: A sample KB

Formatted Text	Relation Name
পল্লীকবি/NP জসীম উদ্দীন/PER ফরিদপুর/LOC জেলায়/NC জন্মগ্রহণ/NC করেছেন/VM	জন্মস্থান

Table 9: Labeled Corpus

Valid Pattern Extraction

After generating candidate patterns for each relation, calculated Conflict Score value for each pattern of each relation.

$$CSr_i, ptr_j = \frac{\#\,\text{conflict instances}\ (e1,\ e2)}{\#\,\text{of seed instances}\ (e1,\ e2)}$$

- Here, *e1* and *e2* represent entity pairs extracted from the corpus, and the **conflict instances** occur when *e1* matches with an entity in the corpus but does not match with the corresponding *e2* in the KB.
- The total number of **seed instances** refers to the number of entity pairs in the corpus that match the KB for the given relation and pattern.

In the training phase, we selected the top 80% of patterns with the lowest conflict scores as valid. Sentences containing these patterns were relabeled as 'NONE' to reduce the effect of noisy patterns on relation extraction accuracy.

Valid Pattern Extraction

Different K values were evaluated to identify the optimal K value for selecting top K% valid patterns based on conflict scores. Notably, when **K was set to 80**, the highest F1 score was achieved compared to other tested values.

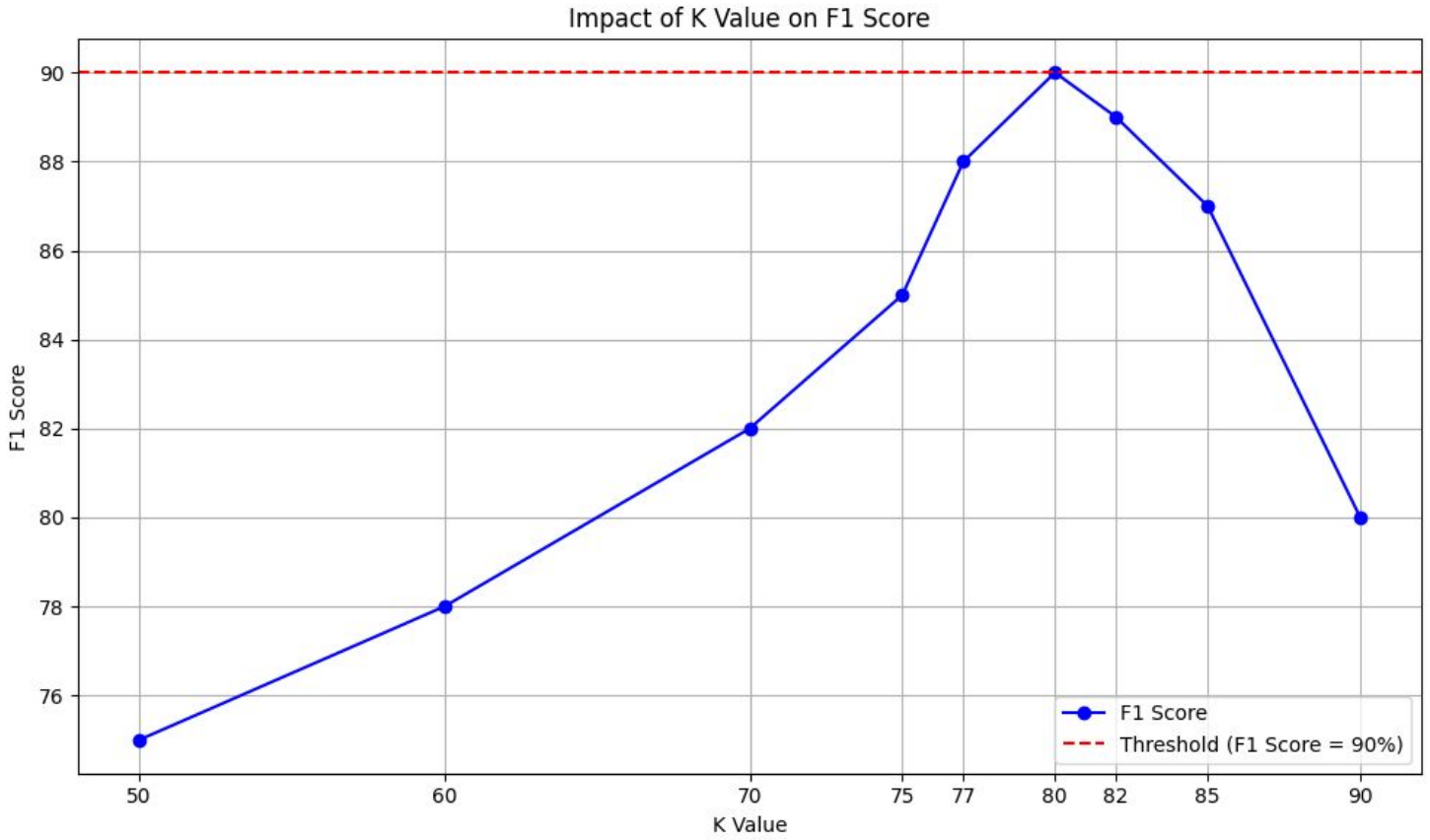


Fig 3: Impact of K Value on F1 Score for our Proposed Approach

Valid Pattern Extraction

Nonetheless, **A sentence with noisy pattern remains incorrectly labeled if both entities match with an entity pair in the KB,** a scenario the CS method cannot resolve. Our proposed method addresses this issue and enhances the efficiency of DS-based RE for Bangla text.

The probability score for each pattern of each relation,

$$PSr_i, ptr_j = \frac{\eta_{r_i, ptr_j}}{\eta_{r_i}}$$

where:

- η_{r_i, ptr_j} represents the total number of seed instances for pattern ptr_j of relation r_i.
- η_{r_i} denotes the total number of seed instances for relation r_i.

Valid Pattern Extraction

We set a threshold $\phi_{r_p} = 0.09$ for each pattern based on probability scores. Patterns with $PS(r_i, ptr_j) > \phi_{r_p}$ were considered valid, filtering out previously misclassified patterns.

Relation Name	Pattern	CS	PS
জন্মস্থান (Place of Birth)	প্রাতিষ্ঠানিক শিক্ষা শুরু হয়েছিলো	0.282	0.076
চলচিত্র পরিচালক (Movie Director)	চলচিত্রে অভিনয় করেছিলেন	0.562	0.029
লেখক (Writer)	বইটি উৎসর্গ করেছেন	0.342	0.022
প্রতিষ্ঠাতা (Company Founder)	কোম্পানির সভাপতির দায়িত্ব পালন করেছেন	0.442	0.134

Table 10: CS and PS Values for Patterns of each relation

Valid Pattern Extraction

Relation Name	Pattern	CS	PS
জন্মস্থান (Place of Birth)	প্রাতিষ্ঠানিক শিক্ষা শুরু হয়েছিলো	0.282	0.076
চলচিত্র পরিচালক (Movie Director)	চলচিত্রে অভিনয় করেছিলেন	0.562	0.029
লেখক (Writer)	বইটি উৎসর্গ করেছেন	0.342	0.022
প্রতিষ্ঠাতা (Company Founder)	কোম্পানির সভাপতির দায়িত্ব পালন করেছেন	0.442	0.134

CS Filtering: Patterns were initially deemed valid based on conflict scores.

Post-PS Filtering:

- **Green-marked Patterns**: Correctly identified as invalid.
- **Red-marked Patterns**: Incorrectly marked as invalid.

Outcome: Majority of incorrect patterns were filtered out, leading to the successful retrieval of valid patterns.

Relation Extraction

- After Relabeling the sentences with noisy patterns to NONE, we trained our model using **ensemble method**. Combined model outputs via majority voting or averaging to improve accuracy and robustness.

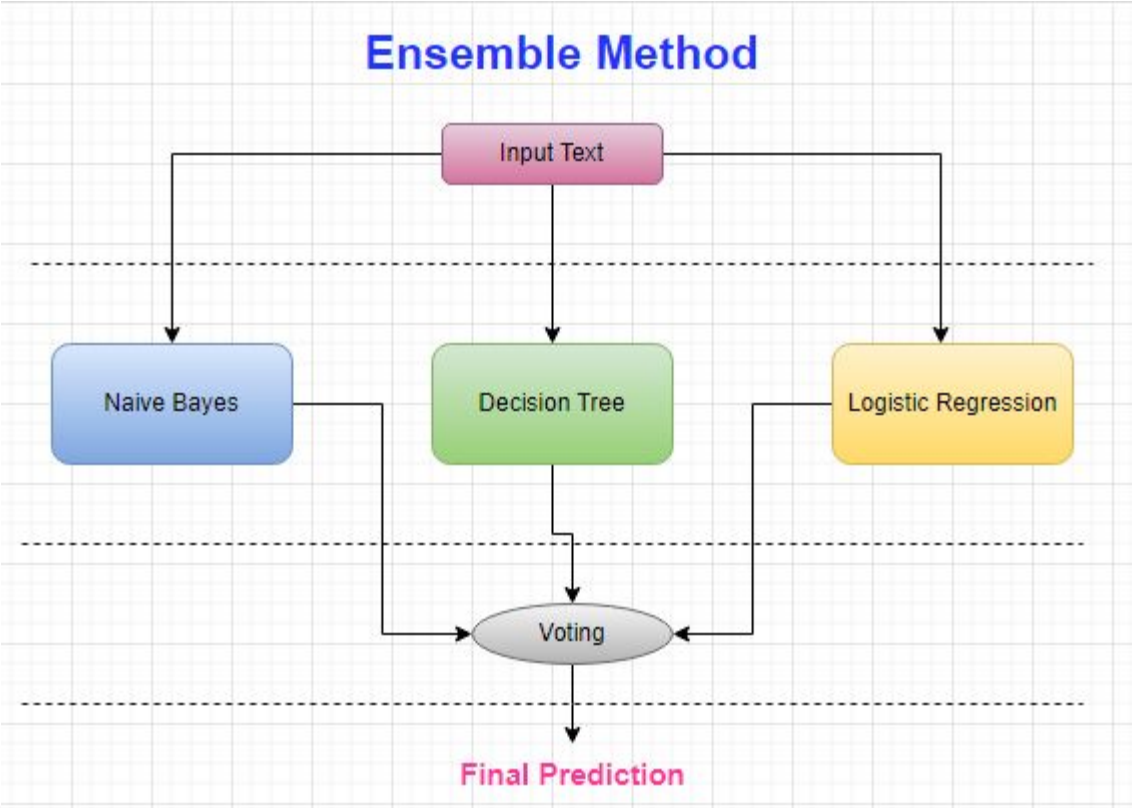


Fig 2: Ensemble Method

Experimental Result

Method	Logistic Regression	SVM	Decision Tree	Ensemble
Baseline	0.840769	0.842463	0.83275	0.850972
Baseline + CS	0.892496	0.894496	0.883203	0.902425
Baseline + PS	0.894761	0.899521	0.887914	0.901880
Baseline + CS + PS	0.899416	0.902916	0.895634	**0.912463**

Table 11: Comparison of F1 Scores for Different Classifiers and Methods

Method	**Accuracy**	**Precision**	**Recall**	**F1 Score**
Baseline	0.858369	0.863187	0.858369	0.850972
Baseline + CS	0.902361	0.904230	0.902361	0.902425
Baseline + PS	0.902652	0.912145	0.906652	0.901880
Baseline + CS + PS	0.903090	0.910362	0.904390	**0.912463**

Table 12: Performance Comparison of Different Methods

Experimental Result (Cont.)

Relation	Precision	Recall	F1-Score	Support
NONE	0.74	0.70	0.72	125
Movie Actor	0.84	0.94	0.89	125
Movie Director	0.88	0.82	0.85	125
Place of Birth	0.91	0.96	0.93	125
Company Founder	1.00	0.93	0.96	57
Company Location	1.00	1.00	1.00	125
Place of Death	1.00	0.92	0.96	125
Writer	1.00	1.00	1.00	125
accuracy			0.91	932
macro avg	0.92	0.91	0.92	932
weighted avg	0.91	0.91	0.91	932

Table 13: Relation Extraction Report for Proposed Methodology (Base+CS+PS)

Experimental Result (Cont.)

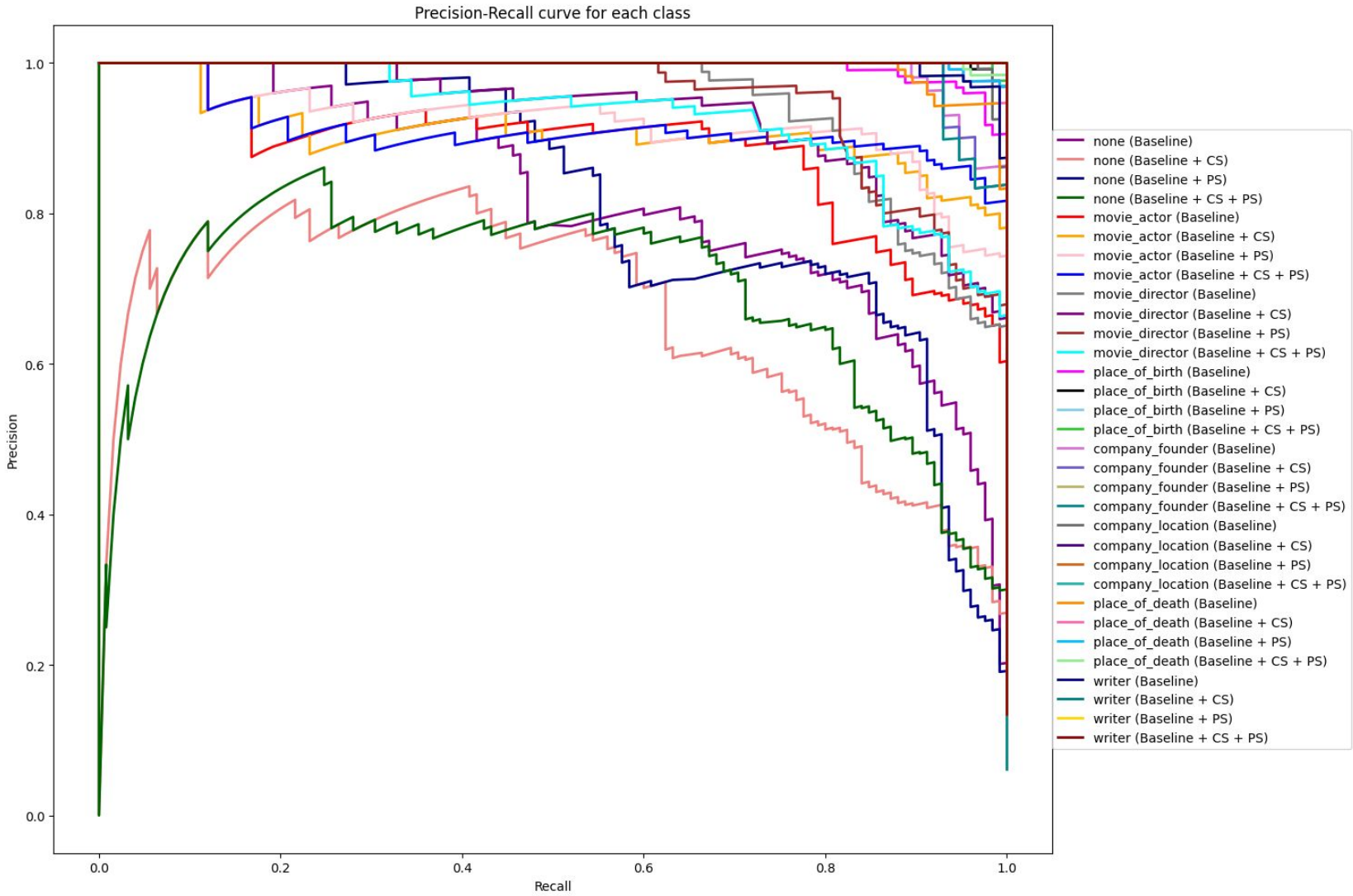


Fig 4: Precision-Recall curve for all seven relations for each method

Experimental Result (Cont.)

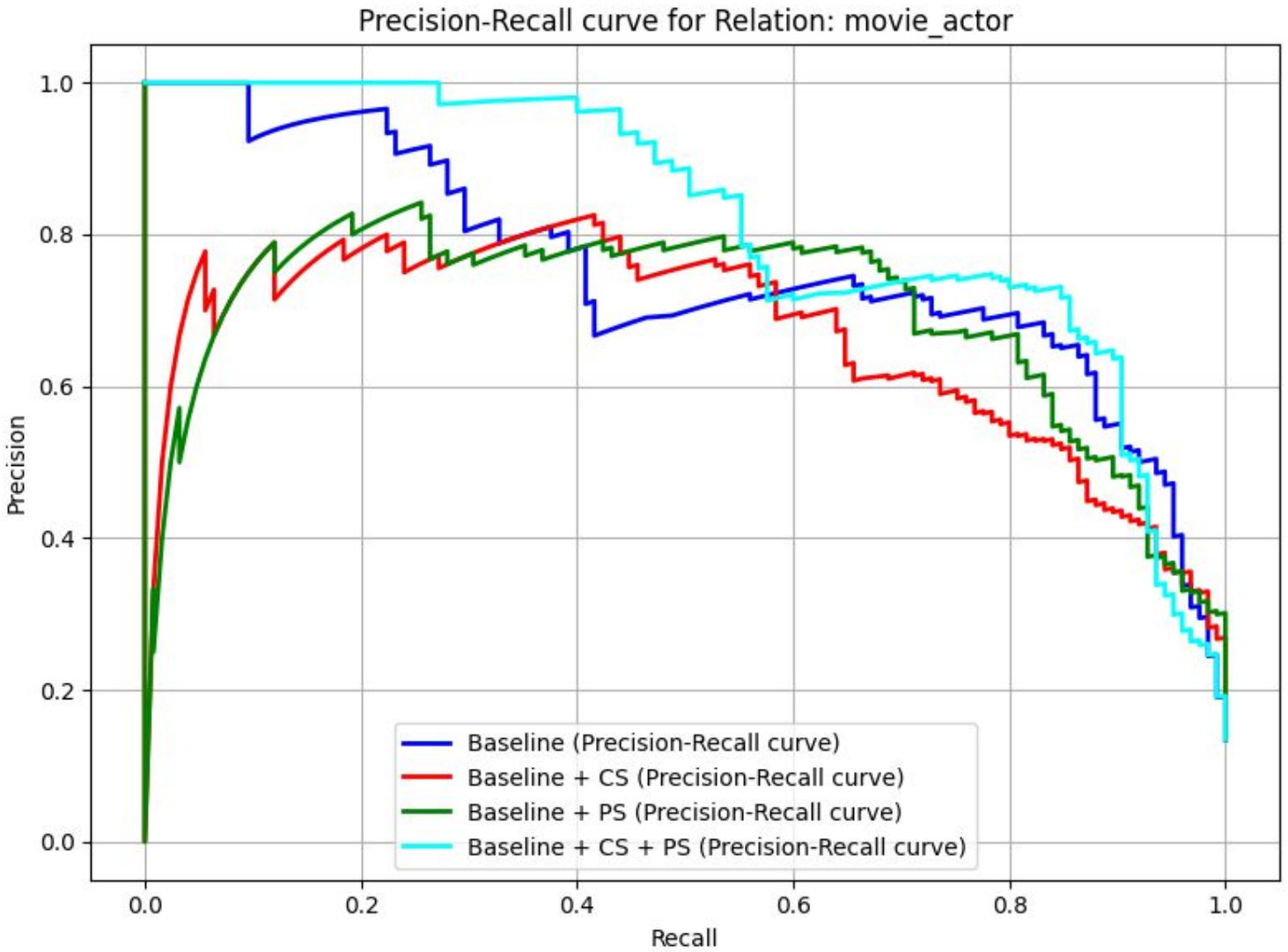


Fig 5: Precision-Recall curve for **Movie Actor** relations for each method

Experimental Result (Cont.)

Method of Retrieving Relevant Patterns	Precision	Recall	F1 Score
Baseline	0.876658	0.871245	0.861100
Baseline + CS	0.905728	0.905579	0.904445
Baseline + PS	0.910692	0.907725	0.903962
Baseline + CS + PS	0.913658	0.913090	0.912426

Table 14: Precision, Recall, and F1 Score for Relation Movie Actor

The individual P-R curves provide detailed insights into the strengths of our proposed method.

Experimental Result (Cont.)

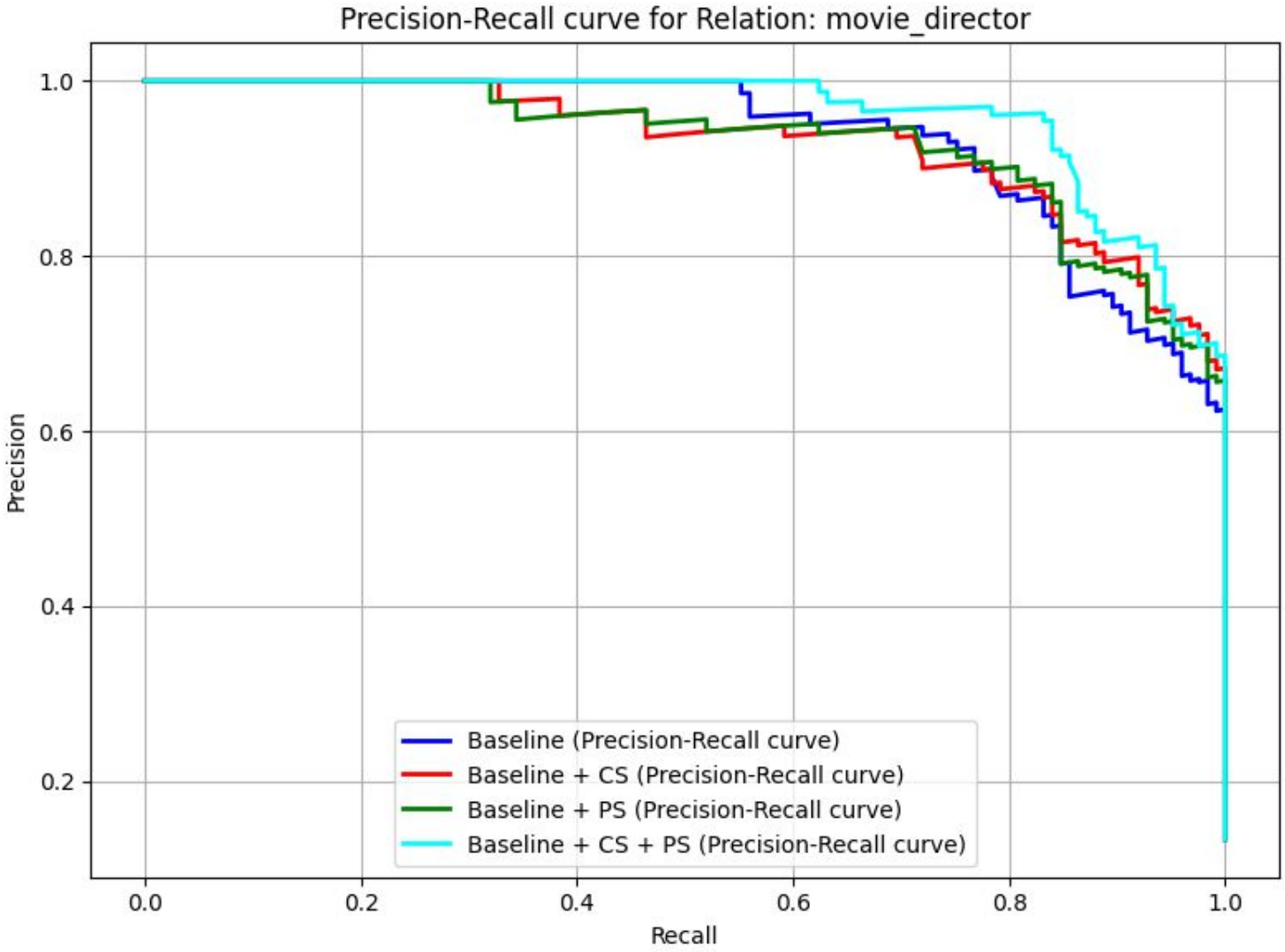


Fig 6: Precision-Recall curve for **Movie Director** relations for each method

Experimental Result (Cont.)

Method of Retrieving Relevant Patterns	Precision	Recall	F1 Score
Baseline	0.870300	0.865880	0.855359
Baseline + CS	0.900572	0.899142	0.898607
Baseline + PS	0.909125	0.907725	0.907410
Baseline + CS + PS	0.919216	0.915236	0.911896

Table 15: Precision, Recall, and F1 Score for Relation Movie Director

Experimental Result (Cont.)

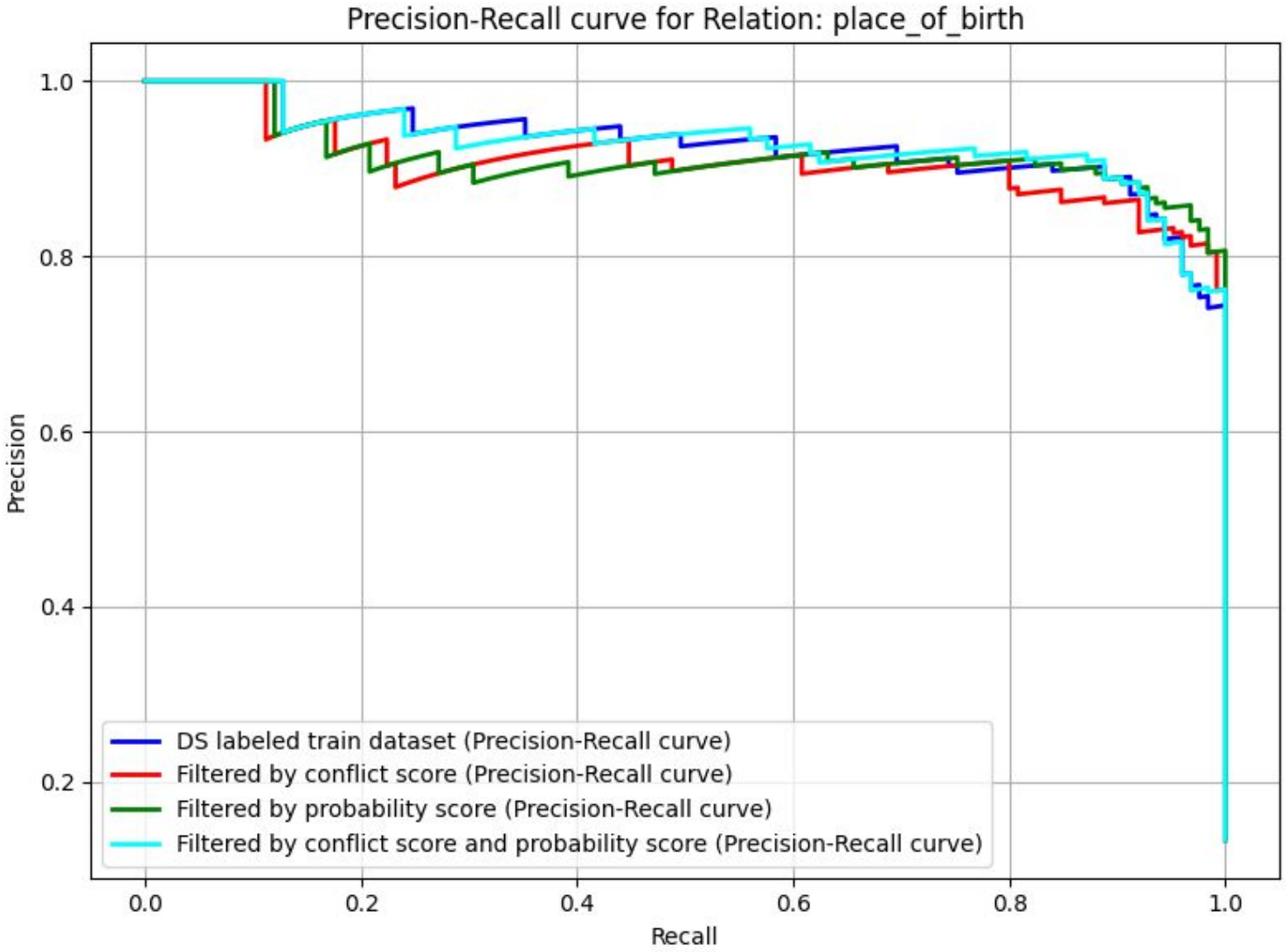


Fig 7: Precision-Recall curve for **Place of Birth** relations for each method

Experimental Result (Cont.)

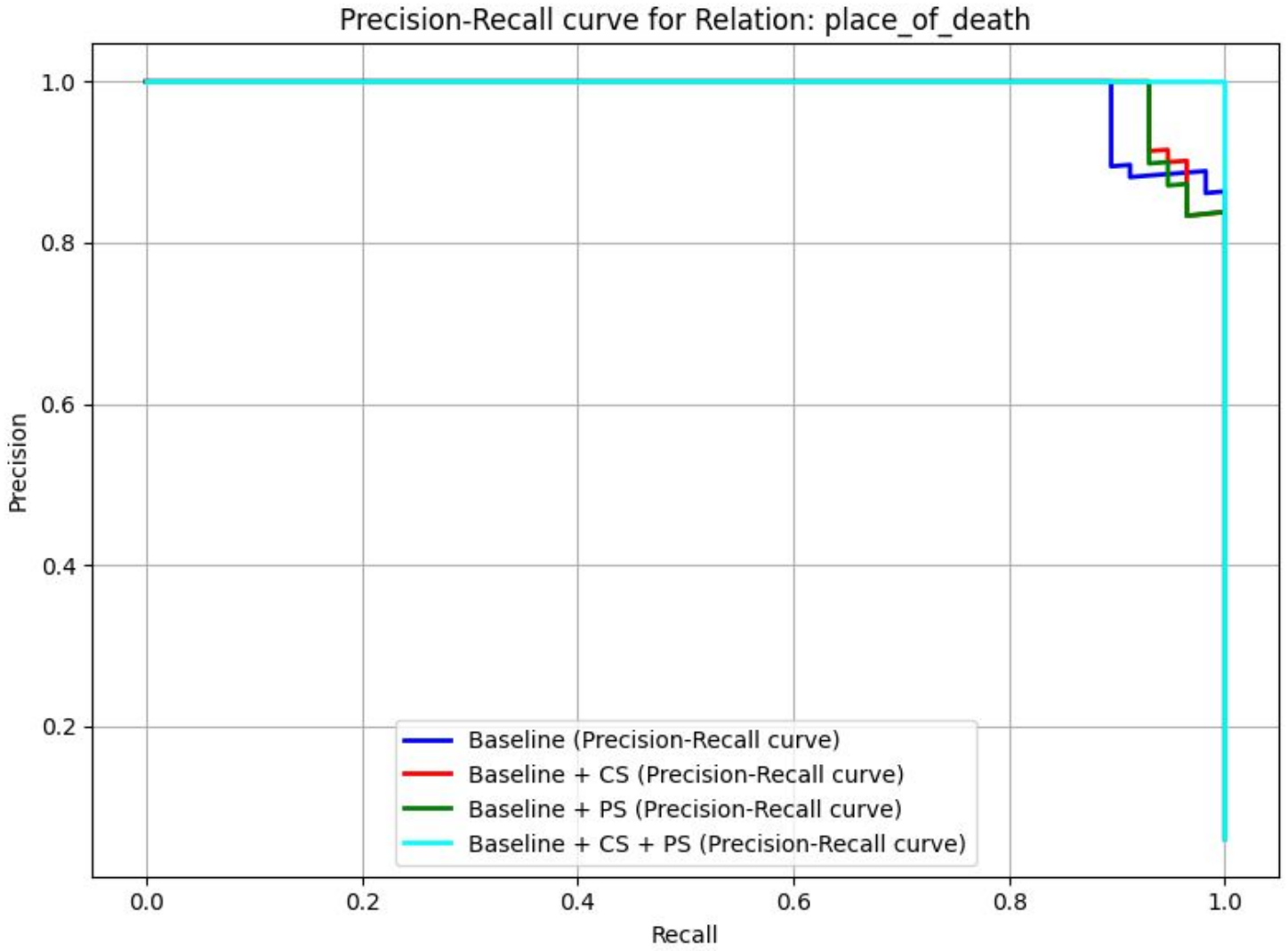


Fig 8: Precision-Recall curve for **Place of Death** relations for each method

Experimental Result (Cont.)

Input Text: সত্যজিৎ রায় ১৯৯১ সালে কলকাতায় জন্মগ্রহণ করেছিলেন।

```
Relation extracted: জন্মস্থান (সত্যজিৎ রায়,কলকাতা)
Confidence score: 92%
```

Input Text: বাক্সরহস্য সত্যজিৎ রায়ের লিখা একটি গোয়েন্দা উপন্যাস।

```
Relation extracted: লেখক (সত্যজিৎ রায়,বাক্সরহস্য)
Confidence score: 89%
```

Conclusion

- This thesis successfully tackled the challenge of extracting valid patterns for **distant supervision-based relation extraction** in low-resource languages like Bangla.
- A novel approach was introduced to **mitigate the issues of noisy data and scarce linguistic resources.**
- Through **rigorous experimentation**, the method achieved a strong **F1 score of 91%**, demonstrating its effectiveness in extracting meaningful patterns from noisy datasets and advancing Bangla NLP.

Conclusion (Cont.)

Limitations:

Our location mnemonics are not sufficiently comprehensive. For example, in Location mnemonics, there is a record –
বাংলাদেশ > নরসিংদী > রায়পুরা.

However, in our corpus, the text "শামসুর রহমান পাড়াতলী গ্রামে জন্মগ্রহণ করেছেন" is present.

Here, পাড়াতলী (Paratolee) is a village in রায়পুরা (Raypura) which is the Upazila of নরসিংদী (Narshindi) District in বাংলাদেশ. Due to the incompleteness of our location mnemonics, the relation extracted for this sentence is labeled as 'None' instead of Place of Birth (জন্মস্থান). Expanding the mnemonics can address this issue.

Conclusion (Cont.)

Limitations:

Moreover, This model encounters challenges in extracting relations from certain texts. For instance, in the sentence "বিশ্বকবি তার গীতাঞ্জলি কাব্যগ্রন্থের জন্য নোবেল পুরস্কার পেয়েছেন", the model identifies the entities as (বিশ্বকবি, গীতাঞ্জলি). These entity pairs are not present in our knowledge base. Instead, the knowledge base contains an instance like (রবীন্দ্রনাথ ঠাকুর, গীতাঞ্জলি). This limitation can be addressed by using mnemonics for person entities, similar to the location mnemonics we developed.

Conclusion (Cont.)

Future Work:

Future work will focus on two key areas to address the identified limitations.

- **Expand dataset collection** to cover a wider range of domains.
- **Develop mnemonics for person entities** to enhance identification and extraction.

Conclusion (Cont.)

Person Name	Mnemonics
রবীন্দ্রনাথ ঠাকুর	বিশ্বকবি
জসীমউদ্দীন	পল্লীকবি
শেখ মুজিবর রহমান	বঙ্গবন্ধু
কাজী নজরুল ইসলাম	বিদ্রোহী কবি
লালন ফকির	বাউল সম্রাট
উইলিয়াম শেক্সপিয়ার	বার্ড অফ অ্যাভন
আইজাক নিউটন	ফাদার অফ গ্রাভিটি
উইনস্টন চার্চিল	ব্রিটিশ বুলডগ

Table 16: Sample Mnemonics for Person entity

Paper Submission

List of papers derived from this thesis work:

[1] *Retrieving Top K% Relevant Patterns for Distant Supervision-Based Relation Extraction for Bangla Sentences* [Under Review] International Conference on Signal Processing, Information, Communication and Systems 2024 (**SPICSCON 2024**).

[2] *Bangla-REX: A Distinct Dataset for Bangla Relation Extraction* [Under Review] **Data in Brief**, **2024** (**Journal Rank: Scopus Q2; ESCI Web of Science; Impact Factor: 1.2**

References

[1] Castelli, Vittorio and Imed Zitouni. “Relation Extraction.” *NLP of Semitic Languages* (2014).

[2] Ali, Manzoor, Mohammad Saleem and Axel-Cyrille Ngonga Ngomo. “Unsupervised Relation Extraction Using Sentence Encoding.” *Extended Semantic Web Conference* (2021).

[3] Mintz, Mike, Steven Bills, Rion Snow, and Dan Jurafsky. "Distant supervision for relation extraction without labeled data." In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pp. 1003-1011. 2009.

[4] I. Augenstein, D. Maynard, and F. Ciravegna, “Relation extraction from the web using distant supervision,” in International Conference Knowledge Engineering and Knowledge Management, 2014. [Online]. Available: https: //api.semanticscholar.org/CorpusID:711016

[5] T. Mahfuz, T. F. Suha, and M. M. Anwar, “Reducing wrong labels using conflict score in distant supervision for relation extraction in bangla language,” in 2020 IEEE Asia-Pacific Conference on Computer Science and Data Engi neering (CSDE). IEEE, 2020, pp. 1–6.

Thank You