Learning-Based Named Entity Recognition for Morphologically-Rich, Resource-Scarce Languages

Kazi Saidul Hasan **Md. Altaf ur Rahman** **Vincent Ng**

Human Language Technology Research Institute
University of Texas at Dallas

European Chapter of
the Association for Computational Linguistics, 2009

Goal

Named Entity Recognition (NER) for morphologically-rich, resource-scarce languages

- Use **Bengali** as our representative language
- Focus on identifying PERSON, ORGANIZATION, and LOCATION

Challenges

1. Scarcity of NE-labeled data
   - Unsupervised techniques for English NER (e.g., Collins and Singer (1999)) are unlikely to work well
     - Problem: Lack of capitalization in Bengali
2. Lack of publicly available gazetteers
3. Inaccurate POS tagger
   - Unsupervised POS induction techniques (e.g., distributional clustering) are unlikely to work well
     Problems: Distributional representation may not be reliable for Bengali because of
     - free word order
     - morphological inflections

Weaknesses of Existing Bengali NE Recognizers

- Use their own manually-constructed gazetteers
  - The results are not reproducible
- Use pseudo-affixes (created by extracting the first *n* and the last *n* characters of a word)
  - The process is ad-hoc and may not cover many useful affixes
- Typically adopt a pipelined NER architecture
  - Errors from the POS tagger are propagated to the NE recognizer

Our Approach

Still supervised, but

- we investigate two new features when used in a pipelined architecture for POS tagging and NER:
  1. affixes induced from an unannotated corpus
  2. semantic class information extracted from Wikipedia
- we propose a joint model for learning POS tagging and NER simultaneously

Outline

1. Two New Linguistic Features
   - Induced Affixes
   - Semantic Classes from Wikipedia
2. POS Tagging with New Features
   - Evaluation
3. Pipelined NER with New Features
   - Evaluation
4. Joint Model for POS Tagging and NER
   - Evaluation

Outline

1. Two New Linguistic Features
   - Induced Affixes
   - Semantic Classes from Wikipedia
2. POS Tagging with New Features
   - Evaluation
3. Pipelined NER with New Features
   - Evaluation
4. Joint Model for POS Tagging and NER
   - Evaluation

Affix Induction

Goal

For morphologically-rich languages, a lot of grammatical information is expressed via affixes. Unlike previous approaches, we learn affixes rather than using pseudo-affixes.

Approach: Keshava and Pitler's (2006) method

Assume that
(1) V is a vocabulary extracted from a large, unannotated corpus
(2) α and β are two character sequences
(3) $\alpha\beta$ is the concatenation of α and β

- if $\alpha\beta$, $\alpha \in V$, we extract β as a candidate suffix
- if $\alpha\beta$, $\beta \in V$, we extract α as a candidate prefix

Affix Induction (contd.)

Many of the induced affixes could be spurious.
Example: If both "can" and "candidate" are in V, then "didate" is extracted as an induced suffix

Solution

1. Assign a score to each induced affix
2. Select only those that score over a certain threshold

We use induced affix n-grams as features. These are like word n-grams except that these are made of affixes.

Outline

1. Two New Linguistic Features
   - Induced Affixes
   - Semantic Classes from Wikipedia
2. POS Tagging with New Features
   - Evaluation
3. Pipelined NER with New Features
   - Evaluation
4. Joint Model for POS Tagging and NER
   - Evaluation

Semantic Class Induction from Wikipedia

Goal

1. Generate a list of phrases and tokens that are potentially named entities from all the articles in the Bengali Wikipedia
2. Heuristically annotate each of them with one of four classes: PER (person), ORG (organization), LOC (location), or OTHERS

Generating an Annotated List of Phrases

Steps

1. Generate and annotate the title of each article
   (i) Use category information to annotate the title

How to read and write in Unicode Bangla | Must-read for contributors

পিথাগোরাস

উইকিপিডিয়া, মুক্ত বিশ্বকোষ থেকে

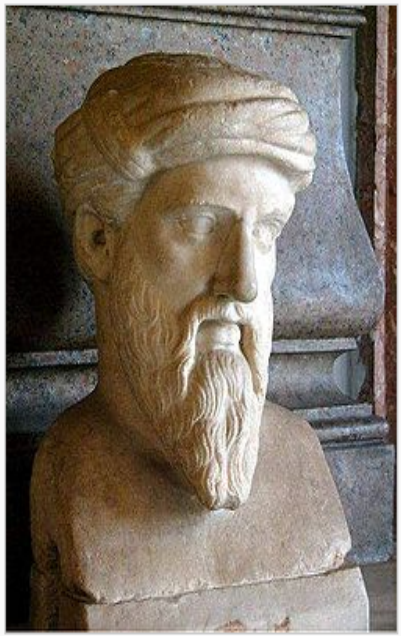
পিথাগোরাস বা **সামোসের পিথাগোরাস** (প্রাচীন গ্রিক ভাষায় Πυθαγόρας *পুথাগোরাস*) (খ্রিস্টপূর্ব ৫৮০ - খ্রিস্টপূর্ব ৫০০) একজন গ্রীক দার্শনিক ও গণিতবিদ।

পিথাগোরাস সবচেয়ে বেশি বিখ্যাত পিথাগোরাসের উপপাদ্যের জন্য। খ্রীস্টের জন্মের প্রায় ৬০০ বছর আগে পিথাগোরাস দর্শণ শাস্ত্রে প্রভুত অবদান রাখেন। তাকে বলা হয় - সংখ্যার জনক। পিথাগোরাস প্রথম নিজেকে দার্শনিক হিসাবে দাবী করেন। প্লেটো, এরিস্টটল ও কোপারনিকাসের অনেক অর্জন পিথাগোরাসের ধারণারই বিস্তৃতি।

তার কোনো লেখা পরে আর পাওয়া যায়নি ফলে তার অবদান সম্পর্কে নিশ্চিত হওয়া যায়নি।

৫২৯ খ্রিস্টপূর্বাব্দে তিনি ইতালীর ক্রোটোনা শহরে বসবাস শুরু করেন।

পিথাগোরাসকে বলা যায় সংখ্যা দার্শনিক। তিনি বিশ্বাস করতেন পৃথিবীর সকল কিছুর মূল হলো সংখ্যা। সংখ্যার সঙ্গে নৈতিকতা, রঙ, শুভ-অশুভকে তিনি মিরিয়ে ফেলেন। তার বিশ্বাস হলো মানুষের আত্মা অবিনশ্বর। ফলে হিন্দু, বৌদ্ধ ও অন্যান্য কয়েকটি ধর্মের মতো তিনিও জন্মান্তরবাদে বিশ্বাসী। পিথাগোরাস পৃথিবীর গোলকত্বে বিশ্বাস করতেন কারণ তার ধারণা ছিল বৃত্ত হলো পরম বক্ররেখা। সূর্য, চন্দ্র বা গ্রহদের নিজেদের একটা গতি আছে।



পিথাগোরাসের প্রতিকৃতি

পিথাগোরাসের উপপাদ্য

[সম্পাদনা]

সমকোণী ত্রিভুজের অতিভুজের ওপর অঙ্কিত বর্গক্ষেত্রের ক্ষেত্রফল অপর দুই বাহুর ওপর অঙ্কিত বর্গক্ষেত্রের ক্ষেত্রফলের সমষ্টির সমান।

এই নিবন্ধটি অসম্পূর্ণ। আপনি চাইলে এটিকে সমৃদ্ধ করতে পারেন।

প্র · আ · স

প্রাক-সক্রেটীয় দার্শনিকবৃন্দ

মিলেতুসীয় ধারা: থেলিস · আনাক্সিমান্দ্রোস · মিলেতুসের আনাক্সিমেনিস
পিথাগোরাসবাদী: **পিথাগোরাস** · ফিলোলাউস · আল্কমায়োন · আর্খুতাস · তিমায়োস
এফেসোসীয় ধারা: ইরাক্লেইতোস — **এলেয়ার দার্শনিক ধারা:** ক্সেনোফানিজ · পার্মেনিদেস · এলেয়ার জিনো · মেলিসোস
বহুত্ববাদী ধারা: আনাক্সাগোরাস · এম্পেদোক্লেস — **পরমাণুবাদী ধারা:** লেউকিপ্পোস · দেমোক্রিতোস
সোফিবাদ: প্রোতাগোরাস · গোর্গিয়াস · প্রোদিকোস · ইপ্পিয়াস
দিওগেনেস · ক্রেরেকুদেস

বিষয়শ্রেণীসমূহ: অসম্পূর্ণ । গ্রিক গণিতবিদ । গ্রিক দার্শনিক । খ্রিস্টপূর্ব ৫৮০-এ জন্ম । খ্রিস্টপূর্ব ৫০০-এ মৃত্যু

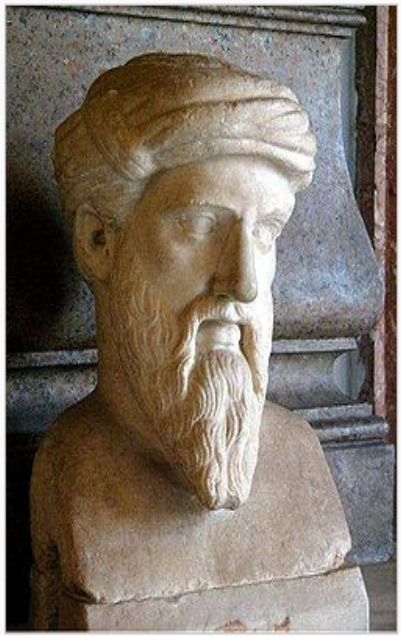
How to read and write in Unicode Bangla | Must-read for contributors

উইকিপিডিয়া
একটি মুক্ত বিশ্বকোষ

পিথাগোরাস

Pythagoras

উইকিপিডিয়া, মুক্ত বিশ্বকো



পিথাগোরাসের প্রতিকৃতি

পিথাগোরাস বা **সামোসের পিথাগোরাস** (প্রাচীন গ্রিক ভাষায় Πυθαγόρας *পুথাগোরাস*) (খ্রিস্টপূর্ব ৫৮০ - খ্রিস্টপূর্ব ৫০০) একজন গ্রীক দার্শনিক ও গণিতবিদ।

পিথাগোরাস সবচেয়ে বেশি বিখ্যাত পিথাগোরাসের উপপাদ্যের জন্য। খ্রীস্টের জন্মের প্রায় ৬০০ বছর আগে পিথাগোরাস দর্শণ শাস্ত্রে প্রভুত অবদান রাখেন। তাকে বলা হয় - সংখ্যার জনক। পিথাগোরাস প্রথম নিজেকে দার্শনিক হিসাবে দাবী করেন। প্লেটো, এরিস্টটল ও কোপারনিকাসের অনেক অর্জন পিথাগোরাসের ধারণারই বিস্তৃতি।

তার কোনো লেখা পরে আর পাওয়া যায়নি ফলে তার অবদান সম্পর্কে নিশ্চিত হওয়া যায়নি।

৫২৯ খ্রিস্টপূর্বাব্দে তিনি ইতালীর ক্রোটোনা শহরে বসবাস শুরু করেন।

পিথাগোরাসকে বলা যায় সংখ্যা দার্শনিক। তিনি বিশ্বাস করতেন পৃথিবীর সকল কিছুর মূল হলো সংখ্যা। সংখ্যার সঙ্গে নৈতিকতা, রঙ, শুভ-অশুভকে তিনি মিরিয়ে ফেলেন। তার বিশ্বাস হলো মানুষের আত্মা অবিনশ্বর। ফলে হিন্দু, বৌদ্ধ ও অন্যান্য কয়েকটি ধর্মের মতো তিনিও জন্মান্তরবাদে বিশ্বাসী। পিথাগোরাস পৃথিবীর গোলকত্বে বিশ্বাস করতেন কারণ তার ধারণা ছিল বৃত্ত হলো পরম বক্ররেখা। সূর্য, চন্দ্র বা গ্রহদের নিজেদের একটা গতি আছে।

পিথাগোরাসের উপপাদ্য

[সম্পাদনা]

সমকোণী ত্রিভুজের অতিভুজের ওপর অঙ্কিত বর্গক্ষেত্রের ক্ষেত্রফল অপর দুই বাহুর ওপর অঙ্কিত বর্গক্ষেত্রের ক্ষেত্রফলের সমষ্টির সমান।

এই নিবন্ধটি অসম্পূর্ণ। আপনি চাইলে এটিকে সমৃদ্ধ করতে পারেন।

প্র · আ · স

প্রাক-সক্রেটীয় দার্শনিকবৃন্দ

মিলেতুসীয় ধারা: থেলিস · আনাক্সিমান্দ্রোস · মিলেতুসের আনাক্সিমেনিস
পিথাগোরাসবাদী: **পিথাগোরাস** · ফিলোলাউস · আল্কমায়োন · আর্খুতাস · তিমায়োস
এফেসোসীয় ধারা: ইরাক্লেইতোস — **এলেয়ার দার্শনিক ধারা:** ক্সেনোফানিজ · পার্মেনিদেস · এলেয়ার জিনো · মেলিসোস
বহুত্ববাদী ধারা: আনাক্সাগোরাস · এম্পেদোক্লেস — **পরমাণুবাদী ধারা:** লেউকিপ্পোস · দেমোক্রিতোস
সোফিবাদ: প্রোতাগোরাস · গোর্গিয়াস · প্রোদিকোস · ইপ্পিয়াস
দিওগেনেস · ক্রেরেকুদেস

বিষয়শ্রেণীসমূহ: অসম্পূর্ণ । গ্রিক গণিতবিদ । গ্রিক দার্শনিক । খ্রিস্টপূর্ব ৫৮০-এ জন্ম । খ্রিস্টপূর্ব ৫০০-এ মৃত্যু

How to read and write in Unicode Bangla | Must-read for contributors

উইকিপিডিয়া
একটি মুক্ত বিশ্বকোষ

পিথাগোরাস

Pythagoras

উইকিপিডিয়া, মুক্ত বিশ্বকো

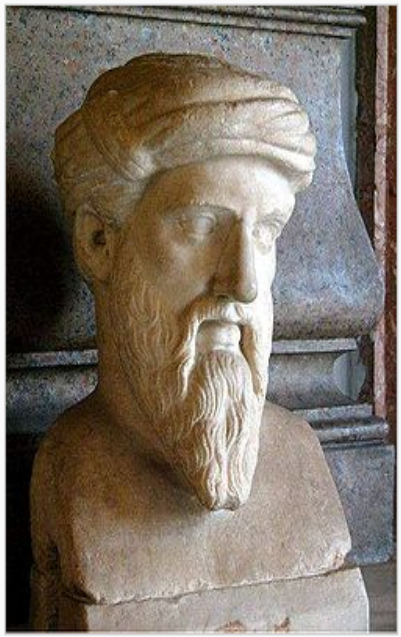
পিথাগোরাস বা **সামোসের পিথাগোরাস** (প্রাচীন গ্রিক ভাষায় Πυθαγόρας *পুথাগোরাস*) (খ্রিস্টপূর্ব ৫৮০ - খ্রিস্টপূর্ব ৫০০) একজন গ্রীক দার্শনিক ও গণিতবিদ।

পিথাগোরাস সবচেয়ে বেশি বিখ্যাত পিথাগোরাসের উপপাদ্যের জন্য। খ্রীস্টের জন্মের প্রায় ৬০০ বছর আগে পিথাগোরাস দর্শণ শাস্ত্রে প্রভূত অবদান রাখেন। তাকে বলা হয় - সংখ্যার জনক। পিথাগোরাস প্রথম নিজেকে দার্শনিক হিসাবে দাবী করেন। প্লেটো, এরিস্টটল ও কোপারনিকাসের অনেক অর্জন পিথাগোরাসের ধারণারই বিস্তৃতি।

তার কোনো লেখা পরে আর পাওয়া যায়নি ফলে তার অবদান সম্পর্কে নিশ্চিত হওয়া যায়নি।

৫২৯ খ্রিস্টপূর্বাব্দে তিনি ইতালীর ক্রোটোনা শহরে বসবাস শুরু করেন।

পিথাগোরাসকে বলা যায় সংখ্যা দার্শনিক। তিনি বিশ্বাস করতেন পৃথিবীর সকল কিছুর মূল হলো সংখ্যা। সংখ্যার সঙ্গে নৈতিকতা, রঙ, শুভ-অশুভকে তিনি মিরিয়ে ফেলেন। তার বিশ্বাস হলো মানুষের আত্মা অবিনশ্বর। ফলে হিন্দু, বৌদ্ধ ও অন্যান্য কয়েকটি ধর্মের মতো তিনিও জন্মান্তরবাদে বিশ্বাসী। পিথাগোরাস পৃথিবীর গোলকত্বে বিশ্বাস করতেন কারণ তার ধারণা ছিল বৃত্ত হলো পরম বক্ররেখা। সূর্য, চন্দ্র বা গ্রহদের নিজেদের একটা গতি আছে।



পিথাগোরাসের প্রতিকৃতি

পিথাগোরাসের উপপাদ্য [সম্পাদনা]

সমকোণী ত্রিভুজের অতিভুজের ওপর অঙ্কিত বর্গক্ষেত্রের ক্ষেত্রফল অপর দুই বাহুর ওপর অঙ্কিত বর্গক্ষেত্রের ক্ষেত্রফলের সমষ্টির সমান।

এই নিবন্ধটি অসম্পূর্ণ। আপনি চাইলে এটিকে সমৃদ্ধ করতে পারেন।

প্র · আ · স — প্রাক-সক্রেটীয় দার্শনিকবৃন্দ

নো · মেলিসোস
তোস

Categories: Greek mathematician & philosopher | Born: 580 BC | Death: 500 BC

বিষয়শ্রেণীসমূহ: অসম্পূর্ণ । গ্রিক গণিতবিদ । গ্রিক দার্শনিক । খ্রিস্টপূর্ব ৫৮০-এ জন্ম । খ্রিস্টপূর্ব ৫০০-এ মৃত্যু

How to read and write in Unicode Bangla | Must-read for contributors

উইকিপিডিয়া
একটি মুক্ত বিশ্বকোষ

পিথাগোরাস

Pythagoras

উইকিপিডিয়া, মুক্ত বিশ্বকো

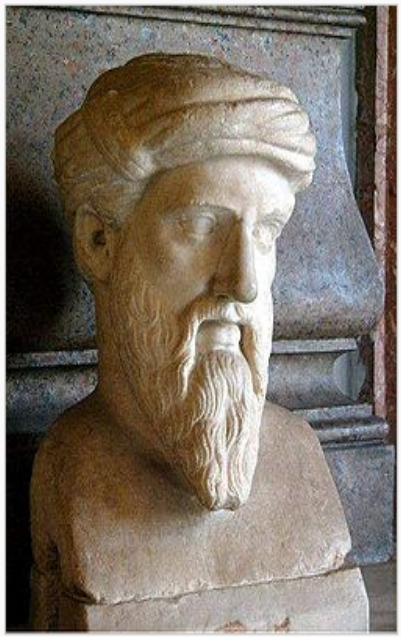
পিথাগোরাস বা **সামোসের পিথাগোরাস** (প্রাচীন গ্রিক ভাষায় Πυθαγόρας *পুথাগোরাস*) (খ্রিস্টপূর্ব ৫৮০ - খ্রিস্টপূর্ব ৫০০) একজন গ্রীক দার্শনিক ও গণিতবিদ।

পিথাগোরাস সবচেয়ে বেশি বিখ্যাত পিথাগোরাসের উপপাদ্যের জন্য। খ্রীস্টের জন্মের প্রায় ৬০০ বছর আগে পিথাগোরাস দর্শণ শাস্ত্রে প্রভূত অবদান রাখেন। তাকে বলা হয় - সংখ্যার জনক। পিথাগোরাস প্রথম নিজেকে দার্শনিক হিসাবে দাবী করেন। প্লেটো, এরিস্টটল ও কোপারনিকাসের অনেক অর্জন পিথাগোরাসের ধারণারই বিস্তৃতি।

তার কোনো লেখা পরে আর পাওয়া যায়নি ফলে তার অবদান সম্পর্কে নিশ্চিত হওয়া যায়নি।

৫২৯ খ্রিস্টপূর্বাব্দে তিনি ইতালীর ক্রোটোনা শহরে বসবাস শুরু করেন।

পিথাগোরাসকে বলা যায় সংখ্যা দার্শনিক। তিনি বিশ্বাস করতেন পৃথিবীর সকল কিছুর মূল হলো সংখ্যা। সংখ্যার সঙ্গে নৈতিকতা, রঙ, শুভ-অশুভকে তিনি মিলিয়ে ফেলেন। তার বিশ্বাস হলো মানুষের আত্মা অবিনশ্বর। ফলে হিন্দু, বৌদ্ধ ও অন্যান্য কয়েকটি ধর্মের মতো তিনিও জন্মান্তরবাদে বিশ্বাসী। পিথাগোরাস পৃথিবীর গোলকত্বে বিশ্বাস করতেন কারণ তা নিজেদের একটা গতি আছে।



পিথাগোরাসের প্রতিকৃতি

PER	Born on, Death on
LOC	Cities of, Countries of

পিথাগোরাসের উপপাদ্য

সমকোণী ত্রিভুজের অতিভুজের ওপর অঙ্কিত বর্গক্ষেত্রের ক্ষেত্রফল অপর দুই বাহু

এই নিবন্ধটি অসম্পূর্ণ। আপনি চাইলে এটিকে সমৃদ্ধ করতে পারেন।

Categories: Greek mathematician & philosopher | Born: 580 BC | Death: 500 BC

ব · আ · স

নো · মেলিসোস
তোস

বিষয়শ্রেণীসমূহ: অসম্পূর্ণ । গ্রিক গণিতবিদ । গ্রিক দার্শনিক । খ্রিস্টপূর্ব ৫৮০-এ জন্ম । খ্রিস্টপূর্ব ৫০০-এ মৃত্যু

Generating an Annotated List of Phrases (contd.)

Steps

1. Generate and annotate the title of each article
   - (i) Use category information to annotate the title
   - (ii) Use a small set of keywords to annotate the title

Generating an Annotated List of Phrases (contd.)

NE Class	Keywords
PER	"born," "died," "one," "famous"
LOC	"city," "area," "population," "located," "part of"
ORG	"establish," "situate," "publish"

Table: Keywords for each named entity class

How to read and write in Unicode Bangla | Must-read for contributors

স্থানাংক: 22.5697, 88.3697

কলকাতা

উইকিপিডিয়া, মুক্ত বিশ্বকোষ থেকে

কলকাতা (ইংরেজি ভাষায়: Kolkata, পূর্বে Calcutta সহায়িকা•তথ্য) ভারতের পূর্বাঞ্চলীয় রাজ্য পশ্চিমবঙ্গের রাজধানী। হুগলী নদীর পূর্ব তীরে অবস্থিত এই শহরের পৌর অঞ্চলের জনসংখ্যা ৫০ লক্ষের কিছু বেশি। তবে কলকাতা ও তার পার্শ্ববর্তী জেলাগুলিতে বিস্তৃত কলকাতার মহানগরীয় অঞ্চলের জনসংখ্যা ১ কোটি ৪০ লক্ষের কাছাকাছি; এই জনসংখ্যার বিচারে কলকাতা ভারতের চতুর্থ বৃহত্তম শহর ও তৃতীয় বৃহত্তম মেট্রোপলিটান বা মহানগরীয় অঞ্চল।

১৯১১ সাল পর্যন্ত কলকাতা ব্রিটিশ ভারতের রাজধানী ছিল। একদা ভারতের আধুনিক শিক্ষা, বিজ্ঞান, শিল্প, সংস্কৃতি ও রাজনীতির কেন্দ্রভূমি কলকাতা মহানগরে ১৯৫৪ সালের পর থেকে তীব্র রাজনৈতিক সংঘাত ও তার সঙ্গে সঙ্গে অর্থনৈতিক মন্দা দেখা দেয়। তবে ২০০০ সালের পর থেকে এই শহর আবার আর্থিক ও বাণিজ্যিক সমৃদ্ধির পথে অগ্রসর হয় ও সাংস্কৃতিক হৃতগৌরব অনেকাংশে পুনরাধিকার করে। তবে ভারতের অন্যান্য মহানগরগুলির মতো নগরায়নজনিত দারিদ্র্য, পরিবেশ দূষণ ও যানজটের সমস্যা থেকে কলকাতাও একেবারে মুক্ত হতে পারেনি।

কলকাতা শহরের প্রসিদ্ধি তার বৈপ্লবিক আন্দোলনগুলির জন্য। ভারতের স্বাধীনতা আন্দোলন এবং পরবর্তীতে বামপন্থী বাণিজ্যিক ইউনিয়নগুলির আন্দোলন এ শহরের ইতিহাসের একটি বড় অংশ। এর সঙ্গে সঙ্গে আধুনিক ভারতে কলকাতা *সাংস্কৃতিক রাজধানী* ও *আনন্দ নগরী* (City of Joy *সিটি অব জয়*) নামে নন্দিত। রবীন্দ্রনাথ ঠাকুর, রোনাল্ড রস, সুভাষচন্দ্র বসু, মাদার তেরেসা, সত্যজিৎ রায়, সত্যেন্দ্রনাথ বসু, স্বামী বিবেকানন্দ রাজা রামমোহন রায় সি ভি রামন-সহ বহু বিশ্ববিশ্রুত ব্যক্তিত্বের বাসভূমি এই নগরী তার ঐতিহাসিক ঐতিহ্য ও সমৃদ্ধ সাংস্কৃতিক প্রেক্ষাপটের কারণে আজও বিশ্ববাসীর চোখে মর্যাদার আসনে অধিষ্ঠিত।

সূচিপত্র [আড়ালে রাখো]

১ নামকরণ
২ ইতিহাস
৩ ভূগোল
৩.১ শহরের গঠন
৪ অর্থনীতি
৫ রাজধানী ও নগর প্রশাসন
৬ যাতায়াত
৭ জনপরিসংখ্যান
৮ সংস্কৃতি
৯ শিক্ষা
১০ গণমাধ্যম

কলকাতা

পশ্চিমবঙ্গ • ভারত

ভিক্টোরিয়া মেমোরিয়াল

স্থানাঙ্ক: 22.5697, 88.3697

সময় অঞ্চল	IST (UTC+৫:৩০)
আয়তন	১,৫০০ বর্গ কিমি (৫৭৯ sq mi)
• উচ্চতা	• ৯ m (৩০ ft)
জেলা	কলকাতা জেলা
জনসংখ্যা	৪,৫৮০,৫৪৪ (২০০১)
• ঘনত্ব	•
• মেট্রো	৯,৯২৫ (এক্সপ্রেশন ত্রুটি: ≠-এর জন্য অপরিজ্ঞাত নেই)

উইকিপিডিয়া
একটি মুক্ত বিশ্বকোষ

পরিভ্রমন
- প্রধান পাতা
- বাংলা উইকিপিডিয়া সম্প্রদায়
- সমসাময়িক ঘটনা
- সাম্প্রতিক পরিবর্তনসমূহ
- অজানা যেকোনো পৃষ্ঠা
- সাহায্য
- দান করুন

অনুসন্ধান

চলো | অনুসন্ধান

হাতিয়ার
- সংযোগকারী পৃষ্ঠাসমূহ
- সম্পর্কিত পরিবর্তন
- বিশেষ পৃষ্ঠাসমূহ
- ছাপার যোগ্য সংস্করণ
- স্থায়ী সংযোগ
- এই নিবন্ধটি উদ্ধৃত করুন

অন্যান্য ভাষাসমূহ
- Aragonés
- العربية
- অসমীয়া
- Azərbaycan
- Беларуская
- Български
- ব্রহ্মণ
- Brezhoneg

Kolkata

কলকাতা

উইকিপিডিয়া, মুক্ত বিশ্বকোষ থেকে

স্থানাংক: 22.5697, 88.3697

কলকাতা (ইংরেজি ভাষায়: Kolkata, পূর্বে Calcutta সহায়িকা•তথ্য) ভারতের পূর্বাঞ্চলীয় রাজ্য পশ্চিমবঙ্গের রাজধানী। হুগলী নদীর পূর্ব তীরে অবস্থিত এই শহরের পৌর অঞ্চলের জনসংখ্যা ৫০ লক্ষের কিছু বেশি। তবে কলকাতা ও তার পার্শ্ববর্তী জেলাগুলিতে বিস্তৃত কলকাতার মহানগরীয় অঞ্চলের জনসংখ্যা ১ কোটি ৪০ লক্ষের কাছাকাছি; এই জনসংখ্যার বিচারে কলকাতা ভারতের চতুর্থ বৃহত্তম শহর ও তৃতীয় বৃহত্তম মেট্রোপলিটান বা মহানগরীয় অঞ্চল।

১৯১১ সাল পর্যন্ত কলকাতা ব্রিটিশ ভারতের রাজধানী ছিল। একদা ভারতের আধুনিক শিক্ষা, বিজ্ঞান, শিল্প, সংস্কৃতি ও রাজনীতির কেন্দ্রভূমি কলকাতা মহানগরে ১৯৫৪ সালের পর থেকে তীব্র রাজনৈতিক সংঘাত ও তার সঙ্গে সঙ্গে অর্থনৈতিক মন্দা দেখা দেয়। তবে ২০০০ সালের পর থেকে এই শহর আবার আর্থিক ও বাণিজ্যিক সমৃদ্ধির পথে অগ্রসর হয় ও সাংস্কৃতিক হৃতগৌরব অনেকাংশে পুনরাধিকার করে। তবে ভারতের অন্যান্য মহানগরগুলির মতো নগরায়নজনিত দারিদ্র্য, পরিবেশ দূষণ ও যানজটের সমস্যা থেকে কলকাতাও একেবারে মুক্ত হতে পারেনি।

কলকাতা শহরের প্রসিদ্ধি তার বৈপ্লবিক আন্দোলনগুলির জন্য। ভারতের স্বাধীনতা আন্দোলন এবং পরবর্তীতে বামপন্থী বাণিজ্যিক ইউনিয়নগুলির আন্দোলন এ শহরের ইতিহাসের একটি বড় অংশ। এর সঙ্গে সঙ্গে আধুনিক ভারতে কলকাতা *সাংস্কৃতিক রাজধানী* ও *আনন্দ নগরী* (City of Joy *সিটি অব জয়*) নামে নন্দিত। রবীন্দ্রনাথ ঠাকুর, রোনাল্ড রস, সুভাষচন্দ্র বসু, মাদার তেরেসা, সত্যজিৎ রায়, সত্যেন্দ্রনাথ বসু, স্বামী বিবেকানন্দ রাজা রামমোহন রায় সি ভি রামন-সহ বহু বিশ্ববিশ্রুত ব্যক্তিত্বের বাসভূমি এই নগরী তার ঐতিহাসিক ঐতিহ্য ও সমৃদ্ধ সাংস্কৃতিক প্রেক্ষাপটের কারণে আজও বিশ্ববাসীর চোখে মর্যাদার আসনে অধিষ্ঠিত।

সূচিপত্র [আড়ালে রাখো]

১ নামকরণ
২ ইতিহাস
৩ ভূগোল
৩.১ শহরের গঠন
৪ অর্থনীতি
৫ রাজধানী ও নগর প্রশাসন
৬ যাতায়াত
৭ জনপরিসংখ্যান
৮ সংস্কৃতি
৯ শিক্ষা
১০ গণমাধ্যম

কলকাতা

পশ্চিমবঙ্গ • ভারত

ভিক্টোরিয়া মেমোরিয়াল

স্থানাঙ্ক: 22.5697, 88.3697

সময় অঞ্চল	IST (UTC+৫:৩০)
আয়তন	১,৫০০ বর্গ কিমি (৫৭৯ sq mi)
• উচ্চতা	• ৯ m (৩০ ft)
জেলা	কলকাতা জেলা
জনসংখ্যা	৪,৫৮০,৫৪৪ (২০০১)
• ঘনত্ব	•
মেট্রো	৯,৯১৫

এক্সপ্রেশন ত্রুটি: *-এর জন্য অপারেন্ড নেই

How to read and write in Unicode Bangla | Must-read for contributors

উইকিপিডিয়া
একটি মুক্ত বিশ্বকোষ

কলকাতা

Kolkata

উইকিপিডিয়া, মুক্ত বিশ্বকোষ থেকে

স্থানাংক: 22.5697, 88.3697

কলকাতা (ইংরেজি ভাষায়: Kolkata, পূর্বে Calcutta সহায়িকা•তথ্য) ভারতের পূর্বাঞ্চলীয় রাজ্য পশ্চিমবঙ্গের রাজধানী। হুগলী নদীর পূর্ব তীরে অবস্থিত এই শহরের পৌর অঞ্চলের জনসংখ্যা ৫০ লক্ষের কিছু বেশি। তবে কলকাতা ও তার পার্শ্ববর্তী [illegible] কাতার ম[illegible] কোটি ৪০ লক্ষের কাছাকাছি; এই জনসংখ্যার [illegible] তর চতুর্থ [illegible] মেট্রোপলিটান বা মহানগরীয় অঞ্চল।

located

population

area

১৯১১ [illegible] তা ব্রিটিশ ভারতের রাজধানী ছিল। একদা ভারতের আধুনিক শিক্ষা, বিজ্ঞান, শিল্প, সংস্কৃতি ও রা[illegible] কলকাতা মহানগরে ১৯৫৪ সালের পর থেকে তীব্র রাজনৈতিক সংঘাত ও তার সঙ্গে সঙ্গে অর্থনৈ[illegible] দয়। তবে ২০০০ সালের পর থেকে এই শহর আবার আর্থিক ও বাণিজ্যিক সমৃদ্ধির পথে অগ্রসর হয় ও সাংস্কৃতিক হৃতগৌরব অনেকাংশে পুনরাধিকার করে। তবে ভারতের অন্যান্য মহানগরগুলির মতো নগরায়নজনিত দারিদ্র্য, পরিবেশ দূষণ ও যানজটের সমস্যা থেকে কলকাতাও একেবারে মুক্ত হতে পারেনি।

কলকাতা শহরের প্রসিদ্ধি তার বৈপ্লবিক আন্দোলনগুলির জন্য। ভারতের স্বাধীনতা আন্দোলন এবং পরবর্তীতে বামপন্থী [illegible] আন্দোলন এ শহরের ইতিহাসের একটি বড় অংশ। এর সঙ্গে সঙ্গে আধুনিক ভারতে [illegible] *ধানী* ও *আনন্দ নগরী* (City of Jo[illegible] নন্দিত। রবীন্দ্রনাথ ঠাকুর, [illegible] বসু, মাদার তেরেসা, সত্যজিৎ রায়, [illegible] বিবেকানন্দ রাজা রামমোহন রায় সি ভি রামন-সহ বহু বিশ্ববিশ্রুত ব্যক্তিত্বের বাসভূমি [illegible] ক ঐতিহ্য ও সমৃদ্ধ সাংস্কৃতিক প্রেক্ষাপটের কারণে আজও বিশ্ববাসীর চোখে মর্যাদার আসনে অধিষ্ঠিত।

famous

part of

সূচিপত্র [আড়ালে রাখো]

১ নামকরণ
২ ইতিহাস
৩ ভূগোল
৩.১ শহরের গঠন
৪ অর্থনীতি
৫ রাজধানী ও নগর প্রশাসন
৬ যাতায়াত
৭ জনপরিসংখ্যান
৮ সংস্কৃতি
৯ শিক্ষা
১০ গণমাধ্যম

কলকাতা

পশ্চিমবঙ্গ • ভারত

ভিক্টোরিয়া মেমোরিয়াল

স্থানাঙ্ক: 22.5697, 88.3697

সময় অঞ্চল	IST (UTC+৫:৩০)
আয়তন	১,৫০০ বর্গ কিমি (৫৭৯ sq mi)
• উচ্চতা	• ৯ m (৩০ ft)
জেলা	কলকাতা জেলা
জনসংখ্যা	৪,৫৮০,৫৪৪ (২০০১)
• ঘনত্ব	•
মেট্রো	৯,৯১[illegible]

এক্সপ্রেশন ত্রুটি: *-এর জন্য অপরিচিত (নেই)

How to read and write in Unicode Bangla | Must-read for contributors

উইকিপিডিয়া
একটি মুক্ত বিশ্বকোষ

Kolkata

কলকাতা

উইকিপিডিয়া, মুক্ত বিশ্বকোষ থেকে

স্থানাংক: 22.5697, 88.3697

কলকাতা (ইংরেজি ভাষায়: Kolkata, পূর্বে Calcutta সহায়িকা•তথ্য) ভারতের পূর্বাঞ্চলীয় রাজ্য পশ্চিমবঙ্গের রাজধানী। হুগলী নদীর পূর্ব তীরে অবস্থিত এই শহরের পৌর অঞ্চলের জনসংখ্যা ৫০ লক্ষের কিছু [illegible] কলকাতা ও তার পার্শ্ববর্তী [illegible] কাতার ম[illegible] কোটি কাছাকাছি; এই জনসংখ্যার [illegible] তর চতুর্থ [illegible] মেট্রোপ[illegible] মহানগরীয় অঞ্চল।

১৯১১ [illegible] তা ব্রিটিশ ভারতের রাজধানী ছিল। একদা ভারতের আধুনিক শিক্ষা, বিজ্ঞান, শিল্প, সংস্কৃতি ও রা[illegible] কলকাতা মহানগরে ১৯৫৪ সালের পর থেকে তীব্র রাজনৈতিক সংঘাত ও তার সঙ্গে সঙ্গে অর্থনৈ[illegible] দয়। তবে ২০০০ সালের পর থেকে এই শহর আবার আর্থিক ও বাণিজ্যিক সমৃদ্ধির পথে অগ্রসর হয় ও সাংস্কৃতিক হৃতগৌরব অনেকাংশে পুনরাধিকার করে। তবে ভারতের অন্যান্য মহানগরগুলির মতো নগরায়নজনিত দারিদ্র্য, পরিবেশ দূষণ ও যানজটের সমস্যা থেকে কলকাতাও একেবারে মুক্ত হতে পারেনি।

কলকাতা শহরের প্রসিদ্ধি তার বৈপ্লবিক আন্দোলনগুলির জন্য। ভারতের স্বাধীনতা আন্দোলন এবং [illegible] বামপন্থী [illegible] আন্দোলন এ শহরের ইতিহাসের একটি বড় অংশ। এর সঙ্গে সঙ্গে [illegible] ভারতে [illegible] *ধানী* ও *আনন্দ নগরী* (City of Jo[illegible] [illegible] ঠাকুর, [illegible] বসু, মাদার তেরেসা, সত্যজিৎ রায়, [illegible] বিবেকানন্দ রাজা রামমোহন রায় সি ভি রামন-সহ বহু বিশ্ববিশ্রুত ব্যক্তিত্বের বাসভূমি [illegible] ক ঐতিহ্য ও সমৃদ্ধ সাংস্কৃতিক প্রেক্ষাপটের কারণে আজও বিশ্ববাসীর চোখে মর্যাদার আসনে অধিষ্ঠিত।

located | population | area | famous | part of

weight = 3

weight = 1

সূচিপত্র [আড়ালে রাখো]

১ নামকরণ
২ ইতিহাস
৩ ভূগোল
৩.১ শহরের গঠন
৪ অর্থনীতি
৫ রাজধানী ও নগর প্রশাসন
৬ যাতায়াত
৭ জনপরিসংখ্যান
৮ সংস্কৃতি
৯ শিক্ষা
১০ গণমাধ্যম

কলকাতা

পশ্চিমবঙ্গ • ভারত

ভিক্টোরিয়া মেমোরিয়াল

স্থানাঙ্ক: 22.5697, 88.3697

সময় অঞ্চল	IST (UTC+৫:৩০)
আয়তন	১,৫০০ বর্গ কিমি (৫৭৯ sq mi)
• উচ্চতা	• ৯ m (৩০ ft)
জেলা	কলকাতা জেলা
জনসংখ্যা	৪,৫৮০,৫৪৪ (২০০১)
• ঘনত্ব	•
মেট্রো	৯,৯১[illegible]

এক্সপ্রেশন ত্রুটি: *-এর জন্য অপারেন্ড নেই

How to read and write in Unicode Bangla | Must-read for contributors

উইকিপিডিয়া
একটি মুক্ত বিশ্বকোষ

পরিভ্রমণ
- প্রধান পাতা
- বাংলা উইকিপিডিয়া সম্প্রদায়
- সমসাময়িক ঘটনা
- সাম্প্রতিক পরিবর্তনসমূহ
- অজানা যেকোনো পৃষ্ঠা
- সাহায্য
- দান করুন

অনুসন্ধান

চলো | অনুসন্ধান

হাতিয়ার
- সংযোগকারী পৃষ্ঠাসমূহ
- সম্পর্কিত পরিবর্তন
- বিশেষ পৃষ্ঠাসমূহ
- ছাপার যোগ্য সংস্করণ
- স্থায়ী সংযোগ
- এই নিবন্ধটি উদ্ধৃত করুন

অন্যান্য ভাষাসমূহ
- Aragonés
- العربية
- অসমীয়া
- Azərbaycan
- Беларуская
- Български
- ব্রহ্মপুত্র
- Brezhoneg

কলকাতা

Kolkata

উইকিপিডিয়া, মুক্ত বিশ্বকোষ থেকে

স্থানাংক: 22.5697, 88.3697

কলকাতা (ইংরেজি ভাষায়: Kolkata, পূর্বে Calcutta সহায়িকা•তথ্য) ভারতের পূর্বাঞ্চলীয় রাজ্য পশ্চিমবঙ্গের রাজধানী। হুগলী নদীর পূর্ব তীরে অবস্থিত এই শহরের পৌর অঞ্চলের জনসংখ্যা ৫০ লক্ষের কিছু [illegible] কলকাতা ও তার পার্শ্ববর্তী [illegible] কাতার ম[illegible] কোটি কাছাকাছি; এই জনসংখ্যার [illegible] তর চতুর্থ [illegible] মেট্রোপ[illegible] মহানগরীয় অঞ্চল।

located | population | weight = 3

১৯১১ [illegible] তা ব্রিটিশ ভারতের রাজধানী ছিল। একদা ভারতের আধুনিক শিক্ষা, বিজ্ঞান, শিল্প, সংস্কৃতি ও রা[illegible] কলকাতা মহানগরে ১৯৫৪ সালের পর থেকে তীব্র রাজনৈতিক সংঘাত ও তার সঙ্গে সঙ্গে অর্থনৈ[illegible] দয়। তবে ২০০০ সালের পর থেকে এই শহর আবার আর্থিক ও বাণিজ্যিক সমৃদ্ধির পথে অগ্রসর হয় ও সাংস্কৃতিক হৃতগৌরব অনেকাংশে পুনরাধিকার করে। তবে ভারতের অন্যান্য মহানগরগুলির মতো নগরায়নজনিত দারিদ্র্য, পরিবেশ দূষণ ও যানজটের সমস্যা থেকে কলকাতাও একেবারে মুক্ত হতে পারেনি।

area

কলকাতা শহরের প্রসিদ্ধি তার বৈপ্লবিক আন্দোলনগুলির জন্য। ভারতের স্বাধীনতা আন্দোলন এবং [illegible] বামপন্থী [illegible] আন্দোলন এ শহরের ইতিহাসের একটি বড় অংশ। এর সঙ্গে সঙ্গে [illegible] ভারতে [illegible] *ধানী* ও *আনন্দ নগরী* (City of Jo[illegible] [illegible] ঠাকুর, [illegible] বসু, মাদার তেরেসা, সত্যজিৎ রায়, [illegible] বিবেকানন্দ রাজা রামমোহন রায় সি ভি রামন-সহ বহু বিশ্ববিশ্রুত ব্যক্তিত্বের বাসভূমি [illegible] ক ঐতিহ্য ও সমৃদ্ধ সাংস্কৃতিক প্রেক্ষাপটের কারণে আজও বিশ্ববাসীর চোখে মর্যাদার আসনে আধাষ্ঠিত।

famous | part of | weight = 1

weighted sum for PERSON, ws_p
weighted sum for ORGANIZATION, ws_o
weighted sum for LOCATION, ws_l

সূচিপত্র [আড়ালে রাখো]

১ নামকরণ
২ ইতিহাস
৩ ভূগোল
৩.১ শহরের গঠন
৪ অর্থনীতি
৫ রাজধানী ও নগর প্রশ[illegible]
৬ যাতায়াত
৭ জনপরিসংখ্যান
৮ সংস্কৃতি
৯ শিক্ষা
১০ গণমাধ্যম

কলকাতা
পশ্চিমবঙ্গ • ভারত

ভিক্টোরিয়া মেমোরিয়াল

স্থানাঙ্ক: 22.5697, 88.3697

IST (UTC+৫:৩০)

১,৫০০ বর্গ কিমি (৫৭৯ sq mi)
• ৯ m (৩০ ft)

কলকাতা জেলা

৪,৫৮০,৫৪৪ (২০০১)

• ঘনত্ব
মোট ৯,৯২৩ (এক্সপ্রেশন ত্রুটি: ≤-এর জন্য অপরিবর্তিত নেই)

Generating an Annotated List of Phrases (contd.)

Steps

1. Generate and annotate the title of each article
   - (i) Use category information to annotate the title
   - (ii) Use a small set of keywords to annotate the title
2. Getting more location names
3. Generating and annotating the tokens in the titles

Generating an Annotated List of Phrases (contd.)

Annotating the tokens in the titles

- Assign each token the same NE label as that of its title
- Ambiguous case:

Anna	*University*
ORG	ORG

Anna	*Frank*
PER	PER

- Solution: Label "Anna" with its most frequent NE class

Applying the Annotated List to a Text

Example

Smith College is in Massachusetts

Applying the Annotated List to a Text

Example

Smith College is in Massachusetts

Applying the Annotated List to a Text

Example

Smith	*College*	*is*	*in*	*Massachusetts*
ORG	ORG			

Applying the Annotated List to a Text

Example

Smith	*College*	*is*	*in*	*Massachusetts*
ORG	ORG	OTHERS	OTHERS	LOC

Outline

1. Two New Linguistic Features
   - Induced Affixes
   - Semantic Classes from Wikipedia
2. POS Tagging with New Features
   - Evaluation
3. Pipelined NER with New Features
   - Evaluation
4. Joint Model for POS Tagging and NER
   - Evaluation

POS Tagging

Goal

Investigate whether the two new features can improve a baseline supervised POS tagger

Experimental setup

- Corpus: 78K word tokens
- Tagset: IIIT Hyderabad's POS tagset with 26 tags
- Learning algorithm: CRF
- 5-fold cross-validation (CV)

Features

Baseline

- word n-grams, pseudo-affixes

Results

Experiment	Overall	Seen	Unseen
Baseline	89.8	92.9	72.0

Table: 5-fold cross-validation accuracies

Features

Baseline+Induced Affixes

- word n-grams, pseudo-affixes
- induced affix n-grams

Results

Experiment	Overall	Seen	Unseen
Baseline	89.8	92.9	72.0
Baseline+Induced Affixes	90.5	93.3	74.6

Table: 5-fold cross-validation accuracies

Features

Baseline+Induced Affixes+Wiki

- word n-grams, pseudo-affixes
- induced affix n-grams
- induced Wiki n-grams

Results

Experiment	Overall	Seen	Unseen
Baseline	89.8	92.9	72.0
Baseline+Induced Affixes	90.5	93.3	74.6
Baseline+Induced Affixes+Wiki	90.8	93.5	75.5

Table: 5-fold cross-validation accuracies

Outline

1. Two New Linguistic Features
   - Induced Affixes
   - Semantic Classes from Wikipedia
2. POS Tagging with New Features
   - Evaluation
3. Pipelined NER with New Features
   - Evaluation
4. Joint Model for POS Tagging and NER
   - Evaluation

NER

Goal

Investigate whether the two features can improve a pipelined NE recognizer that uses induced POS tags as features

Experimental setup

- Corpus: 78K word tokens
- Learning algorithm: CRF
- IOB convention
- 5-fold cross-validation (CV)

Features

Baseline

- word n-grams, pseudo-affixes, **POS n-grams**

Results

Experiment	R	P	F
Baseline	60.9	74.4	67.0
Person	66.1	74.0	69.9
Organization	29.8	44.9	35.8
Location	52.6	80.4	63.6

Table: 5-fold cross-validation results

Features

Baseline+Induced Affixes

- word n-grams, pseudo-affixes, **POS n-grams**
- induced affix n-grams

Results

Experiment	R	P	F
Baseline	60.9	74.4	67.0
Person	66.1	74.0	69.9
Organization	29.8	44.9	35.8
Location	52.6	80.4	63.6
Baseline+Induced Affixes	60.4	73.3	66.2
Person	65.7	72.6	69.0
Organization	31.7	46.4	37.7
Location	51.4	80.0	62.6

Table: 5-fold cross-validation results

Features

Baseline+Induced Affixes+Wiki

- word n-grams, pseudo-affixes, **POS n-grams**
- induced affix n-grams
- induced Wiki n-grams

Results

Experiment	R	P	F
Baseline	60.9	74.4	67.0
Person	66.1	74.0	69.9
Organization	29.8	44.9	35.8
Location	52.6	80.4	63.6
Baseline+Induced Affixes	60.4	73.3	66.2
Person	65.7	72.6	69.0
Organization	31.7	46.4	37.7
Location	51.4	80.0	62.6
Baseline+Induced Affixes+Wiki	63.2	75.1	68.7
Person	66.4	75.1	70.5
Organization	30.7	43.8	36.1
Location	60.0	79.6	68.5

Table: 5-fold cross-validation results

Results

Experiment	R	P	F
Baseline	60.9	74.4	67.0
Person	66.1	74.0	69.9
Organization	**29.8**	44.9	35.8
Location	52.6	80.4	63.6
Baseline+Induced Affixes	60.4	73.3	66.2
Person	65.7	72.6	69.0
Organization	**31.7**	46.4	37.7
Location	51.4	80.0	62.6
Baseline+Induced Affixes+Wiki	63.2	75.1	68.7
Person	66.4	75.1	70.5
Organization	**30.7**	43.8	36.1
Location	60.0	79.6	68.5

Table: 5-fold cross-validation results

Results

Experiment	R	P	F
Baseline	60.9	74.4	67.0
Person	66.1	74.0	69.9
Organization	29.8	44.9	35.8
Location	52.6	80.4	63.6
Baseline+Induced Affixes	60.4	73.3	66.2
Person	**65.7**	72.6	69.0
Organization	31.7	46.4	37.7
Location	51.4	80.0	62.6
Baseline+Induced Affixes+Wiki	63.2	75.1	68.7
Person	**66.4**	75.1	70.5
Organization	30.7	43.8	36.1
Location	60.0	79.6	68.5

Table: 5-fold cross-validation results

Outline

1. Two New Linguistic Features
   - Induced Affixes
   - Semantic Classes from Wikipedia
2. POS Tagging with New Features
   - Evaluation
3. Pipelined NER with New Features
   - Evaluation
4. Joint Model for POS Tagging and NER
   - Evaluation

Joint Model

Goal

Jointly predict a word's POS and NE tag to help avoid error propagation

Model assumption

A Bengali word is part of an NE if and only if it is a proper noun. For our evaluation corpus, this assumption is correct 97.3% of the time.

Joint Model (contd.)

Model

- Jointly predicts a word's POS and NE tag
- Trained using CRF
  - Features: POS and NE taggers' features minus NE tagger's POS-related features
  - Joint tag: If a word is not a proper noun, its class is simply its POS tag. Otherwise, its class is its NE tag

Example:

Dhaka	*is*	*the*	*capital*	*of*	*Bangladesh*
B-LOC	VBZ	DT	NN	IN	B-LOC

NER Results for Joint Modeling

Experiment	**R**	**P**	**F**
Baseline	54.7	81.7	65.5
Baseline+Induced Affixes	56.7	88.9	69.3
Baseline+Induced Affixes+Wiki	61.7	86.3	**71.9**

Table: 5-fold cross-validation joint modeling results for NER

Comparison between Joint and Pipelined Model

Experiment	R	P	F
Baseline	54.7	81.7	65.5
Baseline+Induced Affixes	56.7	88.9	**69.3**
Baseline+Induced Affixes+Wiki	61.7	86.3	**71.9**

Table: 5-fold cross-validation joint modeling results for NER

Experiment	R	P	F
Baseline	60.9	74.4	67.0
Baseline+Induced Affixes	60.4	73.3	**66.2**
Baseline+Induced Affixes+Wiki	63.2	75.1	**68.7**

Table: 5-fold cross-validation results for NER

Summary

- Investigate the pipelined NER architecture and proposed two new features:
  - induced affixes
  - semantic class information from Wikipedia
- Proposed a joint model for learning POS tagging and NER simultaneously