# Whitepaper of NEWS 2010 Shared Task on Transliteration Mining

**A Kumaran**
Microsoft Research India
Bangalore, India

**Mitesh M. Khapra**
Indian Institute of Technology-Bombay
Mumbai, India

**Haizhou Li**
Institute for Infocomm
Research, Singapore

## Abstract

Transliteration is generally defined as phonetic translation of names across languages. Machine Transliteration is a critical technology in many domains, such as machine translation, cross-language information retrieval/extraction, etc. Recent research has shown that high quality machine transliteration systems may be developed in a language-neutral manner, using a reasonably sized good quality corpus (~15-25K parallel names) between a given pair of languages. In this shared task, we focus on acquisition of such good quality names corpora in many languages, thus complementing the machine transliteration shared task that is concurrently conducted in the same NEWS 2010 workshop. Specifically, this task focuses on mining the Wikipedia paired entities data (aka, inter-wiki-links) to produce high-quality transliteration data that may be used for transliteration tasks.

## 1 Task Description

The task is to develop a system for mining single word transliteration pairs from the standard Wikipedia paired topics (aka, Wikipedia Inter-Language Links, or WIL[1]) in one or more of the specified language pairs. The WIL's link articles on the same topic in multiple languages, and are traditionally used as a parallel language resource for many NLP applications, such as Machine Translation, Crosslingual Search, etc. Specific WIL's of interest for our task are those that contain proper names – either wholly or partly – which can yield rich transliteration data.

Each WIL consists of a topic in the source and the language pair, and the task is to identify parts of the topic (in the respective language titles) that are transliterations of each other. A seed data set (of about 1K transliteration pairs) would be provided for each language pair, and are the only resource to be used for developing a mining system. The participants are expected to produce a paired list of source-target single word named entities, for every WIL provided. At the evaluation time, a random subset of WIL's (about 1K WIL's) in each language pair that are hand labeled would be used to test the results produced by the participants.

Participants may use only the 1K seed data provided by the organizers to produce "standard" results; this restriction is imposed to provide a meaningful way of comparing the effective methods and approaches. However, "non-standard" runs would be permitted where participants may use more seed data or any language-specific resource available to them.

## 2 Important Dates

| SHARED TASK SCHEDULES | |
|---|---|
| Registration Opens | 1-Feb-2010 |
| Registration Closes | 13-Mar-2010 |
| Training Data Release | 26 -Feb-2010 |
| Test Data Release | 13-Mar-2010 |
| Results Submission Due | 20-Mar-2010 |
| Evaluation Results Announcement | 27-Mar-2010 |
| Short Papers Due | 5-Apr-2010 |
| Workshop Paper Submission Closes | 5-Apr-2010 |
| Workshop & Task Papers Acceptance | 6-May-2010 |
| CRC Due | 15-May-2010 |
| Workshop Date | 16-Jul-2010 |

## 3 Participation

1. Registration (1 Feb 2010)
   a. Prospective participants are to register to the NEWS-2010 Workshop homepage, for this specific task.
2. Training Data Release (26 Feb 2010)
   a. Registered participants are to obtain seed and Wikipedia data from the Shared Task organizers.

[1] Wikipedia's Interlanguage Links: *http://en.wikipedia.org/wiki/Help:Interlanguage_links*.

*Proceedings of the 2010 Named Entities Workshop, ACL 2010*, pages 29–38,
Uppsala, Sweden, 16 July 2010. ©2010 Association for Computational Linguistics

3. Evaluation Script (1 March 2010)
   a. A sample submission and an evaluation script will be released in due course.
   b. The participants must make sure that their output is produced in a way that the evaluation script may run and produce the expected output.
   c. The same script (with held out test data and the user outputs) would be used for final evaluation.

4. Testing data (13 March 2010)
   a. The test data would be a held out data of approximately 1K “gold-standard” mined data.
   b. The submissions (up to 10) would be tested against the test data, and the results published.

5. Results (27 March 2010)
   a. On the results announcement date, the evaluation results would be published on the Workshop website.
   b. Note that only the scores (in respective metrics) of the participating systems on each language pairs would be published, but no explicit ranking of the participating systems.
   c. Note that this is a shared evaluation task and not a competition; the results are meant to be used to evaluate systems on common data set with common metrics, and not to rank the participating systems. While the participants can cite the performance of their systems (scores on metrics) from the workshop report, they should not use any ranking information in their publications.
   d. Further, all participants should agree not to reveal identities of other participants in any of their publications unless you get permission from the other respective participants. If the participants want to remain anonymous in published results, they should inform the organizers at the time of registration. Note that the results of their systems would still be published, but with the participant identities masked. As a result, in this case, your organization name will still appear in the web site as one of participants, but it is not linked explicitly with your results.

6. Short Papers on Task (5 April 2010)
   a. Each submitting site is required to submit a 4-page system paper (short paper) for its submissions, including their approach, data used and the results.
   b. All system short papers will be included in the proceedings. Selected short papers will be presented in the NEWS 2010 workshop. Acceptance of the system short-papers would be announced together with that of other papers.

## 4 Languages Involved

The task involves transliteration mining in the language pairs summarized in the following table.

| Source Language | Target Language | Track ID |
|---|---|---|
| English | Chinese | WM-EnCn |
| English | Hindi | WM-EnHi |
| English | Tamil | WM-EnTa |
| English | Russian | WM-EnRu |
| English | Arabic | WM-EnAr |

**Table 1:** Language Pairs in the shared task

## 5 Data Sets for the Task

The following datasets are used for each language pair, for this task.

| Training Data | Size | Remarks |
|---|---|---|
| Seed Data (Parallel) | ~1K | Paired names between source and target languages. |
| To-be-mined Wikipedia Inter-Wiki-Link Data (Noisy) | Variable | Paired named entities between source and target languages obtained directly from Wikipedia |
| Test Data | ~1K | This is a subset of Wikipedia Inter-Wiki-Link data, which will be hand labeled. |

**Table 2:** Datasets for the shared task

The first two sets would be provided by the organizers to the participants, and the third will be used for evaluation.

**To-Mine-Data WIL data**: All WIL’s from an appropriate download from Wikipedia would be provided. The WIL data might look like the samples shown in Tables 3 and 4, with the single-word transliterations highlighted. Note that there could be 0, 1 or more single-word transliterations from each WIL.

| # | English Wikipedia Title | Hindi Wikipedia Title |
|---|---|---|
| 1 | Indian National Congress | भारतीय राष्ट्रीय कांग्रेस |
| 2 | University of Oxford | ऑक्सफ़र्ड विश्वविद्यालय |
| 3 | Indian Institute of Science | भारतीय विज्ञान संस्थान |
| 4 | Jawaharlal Nehru University | जवाहरलाल नेहरू विश्वविद्यालय |

**Table 3:** Sample English-Hindi Wikipedia title pairs

| # | English Wikipedia Title | Russian Wikipedia Title |
|---|---|---|
| 1 | Mikhail Gorbachev | Горбачёв, Михаил Сергеевич |
| 2 | George Washington | Вашингтон, Джордж |
| 3 | Treaty of Versailles | Версальский договор |
| 4 | French Republic | Франция |

**Table 4:** Sample English-Russian Wikipedia title pairs

**Seed transliteration data:** In addition we provide approximately 1K parallel names in each language pair as seed data to develop any methodology to identify transliterations. For standard run results, only this seed data could be used, though for non-standard runs, more data or other linguistics resources may be used.

| English Names | Hindi Names |
|---|---|
| Village | विलॆज |
| Linden | लिन्डन |
| Market | मार्केट |
| Mysore | मैसूर |

**Table 5:** Sample English-Hindi seed data

| English Names | Russian Names |
|---|---|
| Gregory | Григорий |
| Hudson | Гудзон |
| Victor | Виктор |
| baranowski | барановский |

**Table 6:** Sample English-Russian seed data

**Test set:** We plan to randomly select ~1000 wikipedia links (from the large noisy Inter-wiki-links) as test-set, and manually extract the single word transliteration pairs associated with each of these WILs. Please note that a given WIL can provide 0, 1 or more single-word transliteration pairs. To keep the task simple, we consider as correct transliterations only those that are clear transliterations word-per-word (morphological variations one or both sides are not considered transliterations) These 1K test set will be a subset of Wikipedia data provided to the user. The gold dataset might look like the following (assuming the items 1, 2, 3 and 4 in Tables 3 and 4 were among the randomly selected WIL's from To-Mine-Data).

| WIL# | English Names | Hindi Names |
|---|---|---|
| 1 | Congress | कांग्रेस |
| 2 | Oxford | ऑक्सफ़र्ड |
| 3 | <Null> | <Null> |
| 4 | Jawaharlal | जवाहरलाल |
| 4 | Nehru | नेहरू |

**Table 7:** Sample English-Hindi transliteration pairs mined from Wikipedia title pairs

| WIL# | English Names | Russian Names |
|---|---|---|
| 1 | Mikhail | Михаил |
| 1 | Gorbachev | Горбачёв |
| 2 | George | Джордж |
| 2 | Washington | Вашингтон |
| 3 | Versailles | Версальский |
| 4 | <Null> | <Null> |

**Table 8:** Sample English-Russian transliteration pairs mined from Wikipedia title pairs

**Evaluation:** The participants are expected to mine such single-word transliteration data for every specific WIL, though the evaluation would be done only against the randomly selected, hand-labeled test set. At evaluation time, the task organizers check every WIL in test set from among the user-provided results, to evaluate the quality of the submission on the 3 metrics described later.

Additional information on data use:

1. Seed data may have ownership and appropriate licenses may need to be procured for use.
2. To-be-mined Wikipedia data is extracted from Wikipedia (in Jan/Feb 2010), and distributed as-is. No assurances that they are correct, complete or consistent.

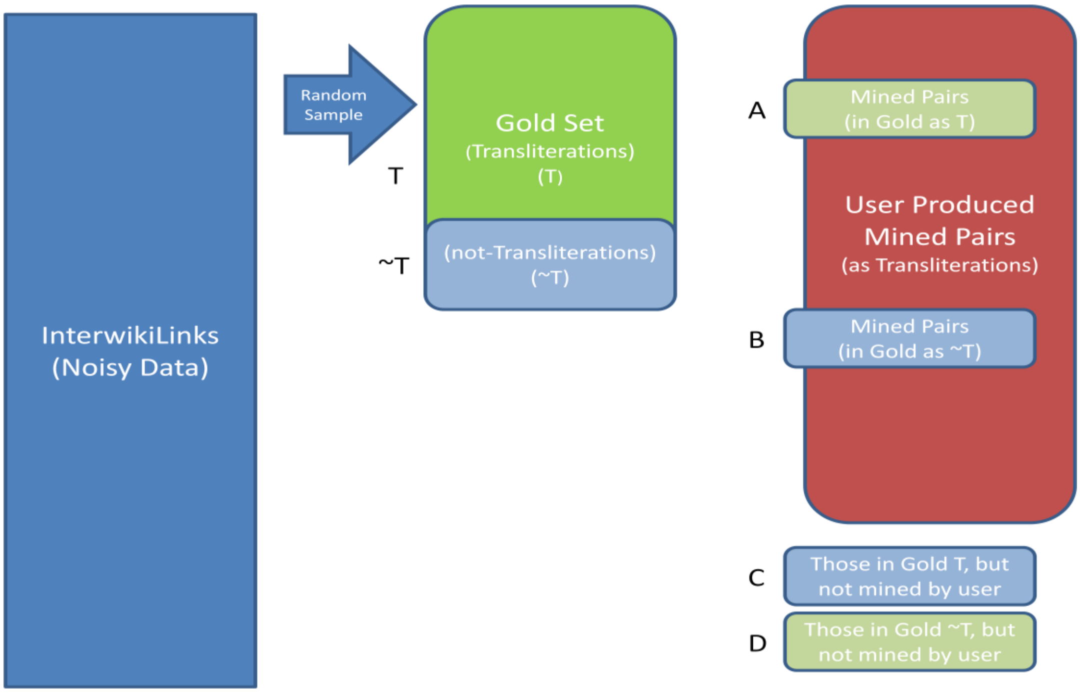
InterwikiLinks (Noisy Data)
Random Sample
T
~T
Gold Set (Transliterations) (T)
(not-Transliterations) (~T)
A
Mined Pairs (in Gold as T)
User Produced Mined Pairs (as Transliterations)
B
Mined Pairs (in Gold as ~T)
C
Those in Gold T, but not mined by user
D
Those in Gold ~T, but not mined by user

**Figure 1:** Overview of the mining task and evaluation

3. The hand-labeled test set is created by NEWS shared task organizers, and will be used for computing the metrics for a given submission.
4. We expect that the participants to use only the seed data (parallel names) provided by the Shared Task for a standard run to ensure a fair evaluation and a meaningful comparison between the effectiveness of approaches taken by various systems. At least one such run (using only the data provided by the shared task) is mandatory for all participants for a given task that they participate in.
5. If more data (either parallel names data or monolingual data), or any language-specific modules were used, then all such runs using extra data or resources must be marked as “Non-standard”. For such non-standard runs, it is required to disclose the size and characteristics of the data or the nature of languages resources used, in their paper.
6. A participant may submit a maximum of 10 runs for a given language pair (including one or more “standard” run). There could be more standard runs, without exceeding 10 (including the non-standard runs).

## 6 Paper Format

All paper submissions to NEWS 2010 should follow the ACL 2010 paper submission policy (http://acl2010.org/papers.html), including paper format, blind review policy and title and author format convention. Shared task system short papers are also in two-column format without exceeding four (4) pages plus any extra page for references. However, there is no need for double-blind requirements, as the users may refer to their runs and metrics in the published results.

## 7 Evaluation Metrics

We plan to measure the quality of the mining task using the following measures:

1. $\text{Precision}_{\text{CorrectTransliterations}}$ ($P_{\text{Trans}}$)
2. $\text{Recall}_{\text{CorrectTransliteration}}$ ($R_{\text{Trans}}$)
3. $\text{F-Score}_{\text{CorrectTransliteration}}$ ($F_{\text{Trans}}$).

Please refer to the following figures for the explanations:

A = True Positives (TP) = Pairs that were identified as "Correct Transliterations" by the participant and were indeed "Correct Transliterations" as per the gold standard
B = False Positives (FP) = Pairs that were identified as "Correct Transliterations" by the participant but they were "Incorrect Transliterations" as per the gold standard.
C = False Negatives (FN) = Pairs that were identified as "Incorrect Transliterations" by the participant but were actually "Correct Transliterations" as per the gold standard.

D = True Negatives (TN) = Pairs that were identified as "Incorrect Transliterations" by the participant and were indeed "Incorrect Transliterations" as per the gold standard.

1. **Recall$_{\textbf{CorrectTransliteration}}$ (R$_{\textbf{Trans}}$)**
The recall is going to be computed using the sample as follows:

$$R_{Trans} = \frac{TP}{TP + FN} = \frac{A}{A + C} = \frac{A}{T}$$

2. **Precision$_{\textbf{CorrectTransliteration}}$ (P$_{\textbf{Trans}}$)**
The precision is going to be computed using the sample as follows:

$$P_{Trans} = \frac{TP}{TP + FP} = \frac{A}{A + B}$$

3. **F-Score (F)**

$$F = \frac{2 * P_{Trans} * R_{Trans}}{P_{Trans} + R_{Trans}}$$

## 8 Contact Us

If you have any questions about this share task and the database, please contact one of the organizers below:

**Dr. A. Kumaran**
Microsoft Research India
Bangalore 560080 INDIA
a.kumaran@microsoft.com

**Mitesh Khapra**
Indian Institute of Technology-Bombay
Mumbai, INDIA
MKhapra@cse.iitb.ac.in.

**Dr Haizhou Li**
Institute for Infocomm Research
Singapore, SINGAPORE 138632
hli@i2r.a-star.edu.sg.

# Appendix A: Seed Parallel Names Data

- File Naming Conventions:
    - NEWS09_Seed_XXYY_1K.xml,
        - XX: Source Language
        - YY: Target Language
        - 1K: number of parallel names

- File Formats:
    - All data would be made available in XML formats (Appendix A).

- Data Encoding Formats:
    - The data would be in Unicode, in UTF-8 encoding. The results are expected to be submitted in UTF-8 format only, and in the XML format specified.

**File: NEWS2009_Seed_EnHi_1000.xml**

```
<?xml version="1.0" encoding="UTF-8"?>
       <SeedCorpus
           CorpusID = "NEWS2009-Seed-EnHi-1K"
           SourceLang = "English"
           TargetLang = "Hindi"
           CorpusType = "Seed"
           CorpusSize = "1000"
           CorpusFormat = "UTF8">
              <Name ID=”1”>
                     <SourceName>eeeeee1</SourceName>
                     <TargetName ID="1">hhhhhh1_1</TargetName>
                     <TargetName ID="2">hhhhhh1_2</TargetName>
                     ...
                     <TargetName ID="n">hhhhhh1_n</TargetName>
              </Name>
              <Name ID=”2”>
                     <SourceName>eeeeee2</SourceName>
                     <TargetName ID="1">hhhhhh2_1</TargetName>
                     <TargetName ID="2">hhhhhh2_2</TargetName>
                     ...
                     <TargetName ID="m">hhhhhh2_m</TargetName>
              </Name>
              ...
              <!-- rest of the names to follow -->
              ...
       </SeedCorpus>
```

# Appendix B: Wikipedia InterwikiLinks Data

- File Naming Conventions:
    - NEWS09_Wiki_XXYY_nnnn.xml,
        - XX: Source Language
        - YY: Target Language
        - nnnn: size of paired entities culled from Wikipedia (“25K”, “10000”, etc.)
- File Formats:
    - All data would be made available in XML formats (Appendix A).
- Data Encoding Formats:
    - The data would be in Unicode, in UTF-8 encoding. The results are expected to be submitted in UTF-8 format only, and in the XML format specified.

**File: NEWS2009_Wiki_EnHi_10K.xml**

```
<?xml version="1.0" encoding="UTF-8"?>
       <WikipediaCorpus
           CorpusID = "NEWS2009-Wiki-EnHi-10K"
           SourceLang = "English"
           TargetLang = "Hindi"
           CorpusType = "Wiki"
           CorpusSize = "10000"
           CorpusFormat = "UTF8">
              <Title ID=”1”>
                     <SourceEntity>e1 e2 … en</SourceEntity>
                     <TargetEntity>h1 h2 … hm</TargetEntity>
              </Title>
              <Title ID=”2”>
                     <SourceEntity>e1 e2 … ei</SourceEntity>
                     <TargetEntity>h1 h2 … hj</TargetEntity>
              </Title>
              ...
              <!-- rest of the titles to follow -->
              ...
       </ WikipediaCorpus>
```

## Appendix C: Results Submission - Format

- File Naming Conventions:
    - NEWS09_Result_XXYY_gggg_nn_description.xml
        - XX: Source
        - YY: Target
        - gggg: Group ID
        - nn: run ID.
        - description: Description of the run
- File Formats:
    - All results would be submitted in XML formats (Appendix B).
- Data Encoding Formats:
    - The data would be in Unicode, in UTF-8 encoding. The results are expected to be submitted in UTF-8 format only.

**Example: NEWS2009_EnHi_TUniv_01_HMMBased.xml**

```
<?xml version="1.0" encoding="UTF-8"?>
       <WikipediaMiningTaskResults
           SourceLang = "English"
           TargetLang = "Hindi"
           GroupID = "Trans University"
           RunID = "1"
           RunType = "Standard"
           Comments = "SVD Run with params: alpha=xxx beta=yyy">
              <Title ID="1">
                     <MinedPair ID="1">
                            <SourceName>e1</SourceName>
                            <TargetName>h1</TargetName>
                     </MinedPair>
                     <MinedPair ID="2">
                            <SourceName>e2</SourceName>
                            <TargetName>h2</TargetName>
                     </MinedPair>
              <!—followed by other pairs mined from this title-->
              </Title>
              <Title ID="2">
                     <MinedPair ID="1">
                            <SourceName>e1</SourceName>
                            <TargetName>h1</TargetName>
                     </MinedPair>
```

```
        <MinedPair ID="2">
            <SourceName>e2</SourceName>
            <TargetName>h2</TargetName>
        </MinedPair>
    <!—followed by other pairs mined from this title-->
    </Title>
    ...
    <!-- All titles in the culled data to follow -->
    ...
</WikipediaMiningTaskResults>
```

## Appendix D: Sample Eng-Hindi Interwikilink Data

```
<?xml version="1.0" encoding="UTF-8"?>
<WikipediaCorpus CorpusID = "NEWS2009-Wiki-EnHi-Sample"
    SourceLang = "English"
    TargetLang = "Hindi"
    CorpusType = "Wiki" CorpusSize = "3"
    CorpusFormat = "UTF8">
        <Title ID="1">
        <SourceEntity>Indian National Congress</SourceEntity>
        <TargetEntity>भारतीय राष्ट्रीय कांग्रेस</TargetEntity>
    </Title>
    <!-- {Congress, कांग्रेस} should be identified by the paricipants-->
    <Title ID="2">
        <SourceEntity>University of Oxford</SourceEntity>
        <TargetEntity>ऑक्सफ़र्ड विश्वविद्यालय</TargetEntity>
    </Title>
    <!-- {Oxford, ऑक्सफ़र्ड} should be identified by the paricipants-->
    <Title ID="3">
        <SourceEntity>Jawaharlal Nehru University</SourceEntity>
        <TargetEntity>जवाहरलाल नेहरू विश्वविद्यालय</TargetEntity>
    </Title>
    <!-- {Jawaharlal, जवाहरलाल} and {Nehru, नेहरू} should be
                    identified by the paricipants-->
    <Title ID="4">
        <SourceEntity>Indian Institute Of Science</SourceEntity>
        <TargetEntity>भारतीय विज्ञान संस्थान</TargetEntity>
    </Title>
    <!--There are no transliteration pairs here -->
</WikipediaCorpus>
```

## Appendix E: Eng-Hindi Gold Mined Data (wrt the above WIL Data)

```
<?xml version="1.0" encoding="UTF-8"?>
<WikipediaMiningTaskResults
    SourceLang = "English"
    TargetLang = "Hindi"
    GroupID = "Gold-Standard"
    RunID = ""
    RunType = ""
    Comments = "">
    <Title ID="1">
        <MinedPair ID="1">
            <SourceName>Congress</SourceName>
            <TargetName> कांग्रेस</TargetName>
        </MinedPair>
    </Title>
    <Title ID="2">
        <MinedPair ID="1">
```

```
                        <SourceName>Oxford</SourceName>
                        <TargetName> ऑक्सफ़र्ड</TargetName>
                </MinedPair>
        </Title>
        <Title ID="3">
                <MinedPair ID="1">
                        <SourceName>Jawaharlal</SourceName>
                        <TargetName> जवाहरलाल</TargetName>
                </MinedPair>
                <MinedPair ID="2">
                        <SourceName>Nehru</SourceName>
                        <TargetName> नेहरू</TargetName>
                </MinedPair>
        </Title>
        <Title ID="4">
        </Title>
</WikipediaMiningTaskResults>
```

## Appendix F: English-Hindi Sample Submission and Evaluation

```
<?xml version="1.0" encoding="UTF-8"?>
<WikipediaMiningTaskResults
        SourceLang = "English"
        TargetLang = "Hindi"
        GroupID = "Gold-Standard"
        RunID = ""
        RunType = ""
        <Title ID="1">
                <MinedPair ID="1">
                        <SourceName>Congress</SourceName>
                        <TargetName> कांग्रेस</TargetName>
                </MinedPair>
The participant mined all correct transliteration pairs
        </Title>
        <Title ID="2">
                <MinedPair ID="1">
                        <SourceName>Oxford</SourceName>
                        <TargetName> ऑक्सफ़र्ड</TargetName>
                </MinedPair>
                <MinedPair ID="1">
                        <SourceName>University</SourceName>
                        <TargetName>विश्वविद्यालय</TargetName>
                </MinedPair>
The participant mined an incorrect transliteration pair {University,विश्वविद्यालय}
        </Title>
        <Title ID="3">
                <MinedPair ID="1">
                        <SourceName>Jawaharlal</SourceName>
                        <TargetName> जवाहरलाल</TargetName>
                </MinedPair>
The participant missed the correct transliteration pair {Nehru, नेहरू}
        </Title>
        <Title ID="4">
                <MinedPair ID="1">
                        <SourceName>Indian</SourceName>
                        <TargetName>भारतीय</TargetName>
                </MinedPair>
The participant mined an incorrect transliteration pair {Indian, भारतीय}
        </Title>
</WikipediaMiningTaskResults>
```

## Sample Evaluation

**T =** |**{**(Congress, कांग्रेस), (Oxford, ऑक्सफ़र्ड), (Jawaharlal, जवाहरलाल),(Nehru, नेहरू)**}** | **= 4**
**A = TP =** | **{**(Congress, कांग्रेस), (Oxford, ऑक्सफ़र्ड), (Jawaharlal, जवाहरलाल)}| = **3**
**B = FP =** |**{**(Indian**,** भारतीय), (University**,** विश्वविद्यालय) **}**| = **2**
**C = FN =** |**{**(Nehru, नेहरू)}| = **1**

$$R_{Trans} = \frac{TP}{TP + FN} = \frac{A}{A + C} = \frac{A}{T} = \frac{3}{4} = 0.75$$

$$P_{Trans} = \frac{TP}{TP+FP} = \frac{A}{A+B} = \frac{3}{5} = 0.60$$

$$F = \frac{2 * P_{Trans} * R_{Trans}}{P_{Trans} + R_{Trans}} = \frac{2 * 0.6 * 0.75}{0.6 + 0.75} = 0.67$$