Mapping Hindi-English Text Re-use Document Pairs

Parth Gupta[1], Khushboo Singhal[2]

[1]NLE Lab, UPV, Spain

[2]IR-Lab, DA-IICT, India

December 4, 2011

Outline

What is Text Re-use?

Text Re-use tries to project the phenomena of Plagiarism where it refers to *"the unauthorized use or close imitation of the language and thoughts of another author and the representation of them as ones own original work, as by not crediting the author"*[1]

In Cross Language Text Re-use the source and suspicious documents are in different languages.

Why two different terms? "Plagiarism" and "Text Re-use"

[1] http://dictionary.reference.com/browse/plagiarism

Text Re-use Detection Systems

Stages of Text Re-use Detection Systems

Text Re-use Detection Systems

Stages of Text Re-use Detection Systems

- Pre-processing - Text Normalization, Language Identification, Translation etc

Text Re-use Detection Systems

Stages of Text Re-use Detection Systems

- Pre-processing - Text Normalization, Language Identification, Translation etc
- Candidate Retrieval - Select a small subset of source documents using CL/IR

Text Re-use Detection Systems

Stages of Text Re-use Detection Systems

- Pre-processing - Text Normalization, Language Identification, Translation etc
- Candidate Retrieval - Select a small subset of source documents using CL/IR
- Detailed Analysis - Compare each doc pairs in detail

Text Re-use Detection Systems

Stages of Text Re-use Detection Systems

- Pre-processing - Text Normalization, Language Identification, Translation etc
- Candidate Retrieval - Select a small subset of source documents using CL/IR
- Detailed Analysis - Compare each doc pairs in detail
- Post-Processing - if properly cited - remove etc

Text Re-use Detection Systems

Stages of Text Re-use Detection Systems

- Pre-processing - Text Normalization, Language Identification, Translation etc
- Candidate Retrieval - Select a small subset of source documents using CL/IR
- Detailed Analysis - Compare each doc pairs in detail
- Post-Processing - if properly cited - remove etc

CLiTR Task Definition

"Find the source document (English) for the given suspicious document (Hindi)

CLiTR Task Definition

"Find the source document (English) for the given suspicious document (Hindi) IF it contains text re-use."

CLiTR Task Definition

"Find the source document (English) for the given suspicious document (Hindi) IF it contains text re-use."

The task is at document level and not at fragment level!

CLiTR Task Definition

"Find the source document (English) for the given suspicious document (Hindi) IF it contains text re-use."

The task is at document level and not at fragment level!

Quite similar to the Candidate Retrieval Step: have to return the first document

Motivation

- Useful for

Motivation

- Useful for
  - Checking Text Authenticity

Motivation

- Useful for
  - Checking Text Authenticity
  - Finding near duplicate candidates

Motivation

- Useful for
  - Checking Text Authenticity
  - Finding near duplicate candidates
  - Finding Parallel data from Comparable data [in case of CL PD]

Motivation

- Useful for
  - Checking Text Authenticity
  - Finding near duplicate candidates
  - Finding Parallel data from Comparable data [in case of CL PD]

Motivation

Motivation

Why to study Indian Text Re-use?

Motivation

Why to study Indian Text Re-use?

- Diversity of Languages in India

Motivation

Why to study Indian Text Re-use?

- Diversity of Languages in India
- Prolific Web Data in Regional Languages: Wikipedia, blogs, personal pages, government pages etc.

Motivation

Why to study Indian Text Re-use?

- Diversity of Languages in India
- Prolific Web Data in Regional Languages: Wikipedia, blogs, personal pages, government pages etc.

Education in Regional Languages

Motivation

Why to study Indian Text Re-use?

- Diversity of Languages in India
- Prolific Web Data in Regional Languages: Wikipedia, blogs, personal pages, government pages etc.

Education in Regional Languages

- 27 June-2011 Outlook India Quoted:

Motivation

Why to study Indian Text Re-use?

- Diversity of Languages in India
- Prolific Web Data in Regional Languages: Wikipedia, blogs, personal pages, government pages etc.

Education in Regional Languages

- 27 June-2011 Outlook India Quoted:
  - "Anna University, Chennai, offers Tamil medium professional courses such as BEs in mechanical and civil engineering"
  - "Gujarat Vidyapith, Ahmedabad, offers an MCA and an MSc (Microbiology) in Gujarati"

Motivation

Why to study Indian Text Re-use?

- Diversity of Languages in India
- Prolific Web Data in Regional Languages: Wikipedia, blogs, personal pages, government pages etc.

Education in Regional Languages

- 27 June-2011 Outlook India Quoted:
  - "Anna University, Chennai, offers Tamil medium professional courses such as BEs in mechanical and civil engineering"
  - "Gujarat Vidyapith, Ahmedabad, offers an MCA and an MSc (Microbiology) in Gujarati"
- 25 Aug-2011 MSN News:

Motivation

Why to study Indian Text Re-use?

- Diversity of Languages in India
- Prolific Web Data in Regional Languages: Wikipedia, blogs, personal pages, government pages etc.

Education in Regional Languages

- 27 June-2011 Outlook India Quoted:
  - "Anna University, Chennai, offers Tamil medium professional courses such as BEs in mechanical and civil engineering"
  - "Gujarat Vidyapith, Ahmedabad, offers an MCA and an MSc (Microbiology) in Gujarati"
- 25 Aug-2011 MSN News:
  - "AICTE exploring options of education in regional languages"

Approach

Phase-1

Transform Hindi Documents into English using available Resources like Bi-lingual Dictionary (D)[2], Wordnet (W)[3], Transliteration System (T)[4]

- D: Replace Hindi word by its English entry word
- W+D: Take all terms from wordnet and then look-up dictionary
- D+T: Look up in Dictionary, IF not then transliterate

[2] http://www.clt.iitb.ac.in/hdict/webinterface_user/index.php
[3] www.cfilt.iitb.ac.in/wordnet/webhwn/
[4] www.google.com/transliterate
[5] We user Terrier 3.5 implementation of BM25

Approach

Phase-1

Transform Hindi Documents into English using available Resources like Bi-lingual Dictionary (D)[2], Wordnet (W)[3], Transliteration System (T)[4]

- D: Replace Hindi word by its English entry word
- W+D: Take all terms from wordnet and then look-up dictionary
- D+T: Look up in Dictionary, IF not then transliterate

Phase-2

Calculate BM25 score between two documents[5]

[2] http://www.clt.iitb.ac.in/hdict/webinterface_user/index.php
[3] www.cfilt.iitb.ac.in/wordnet/webhwn/
[4] www.google.com/transliterate
[5] We user Terrier 3.5 implementation of BM25

Notions

- We consider this as a candidate retrieval task
- We look for high recall so that potential documents are identified for detailed analysis
- We do not want to translate the documents because

Notions

- We consider this as a candidate retrieval task
- We look for high recall so that potential documents are identified for detailed analysis
- We do not want to translate the documents because
  - It incurs computational overhead which is not feasible in real scenario where source pool is extremely large

Notions

- We consider this as a candidate retrieval task
- We look for high recall so that potential documents are identified for detailed analysis
- We do not want to translate the documents because
  - It incurs computational overhead which is not feasible in real scenario where source pool is extremely large e.g WWW

Notions

- We consider this as a candidate retrieval task
- We look for high recall so that potential documents are identified for detailed analysis
- We do not want to translate the documents because
  - It incurs computational overhead which is not feasible in real scenario where source pool is extremely large e.g WWW
  - It is constrained by the availability/quality of Translators.

Notions

- We consider this as a candidate retrieval task
- We look for high recall so that potential documents are identified for detailed analysis
- We do not want to translate the documents because
  - It incurs computational overhead which is not feasible in real scenario where source pool is extremely large e.g WWW
  - It is constrained by the availability/quality of Translators.
  - Even for detailed analysis task, it is upper bounded by quality of Machine Translation.

Notions

- We consider this as a candidate retrieval task
- We look for high recall so that potential documents are identified for detailed analysis
- We do not want to translate the documents because
  - It incurs computational overhead which is not feasible in real scenario where source pool is extremely large e.g WWW
  - It is constrained by the availability/quality of Translators.
  - Even for detailed analysis task, it is upper bounded by quality of Machine Translation.
- We want to investigate the performance of available resources for candidate retrieval

Results

Table: Results

(a) Results on training data

Method	Precision	Recall	F-Measure
D	0.4545	0.6923	0.5488
W+D	0.1717	0.2615	0.2073
D+T	0.5051	**0.7692**	0.6097

[6] $\theta = 15.0$

Results

Table: Results

(a) Results on training data

Method	Precision	Recall	F-Measure
D	0.4545	0.6923	0.5488
W+D	0.1717	0.2615	0.2073
D+T	0.5051	**0.7692**	0.6097
D+T+θ[6]	0.6069	0.6769	**0.6400**

[6] $\theta = 15.0$

Results

Table: Results

(a) Results on training data

Method	Precision	Recall	F-Measure
D	0.4545	0.6923	0.5488
W+D	0.1717	0.2615	0.2073
D+T	0.5051	**0.7692**	0.6097
D+T+θ[6]	0.6069	0.6769	**0.6400**

(b) Results on test data

Method	Precision	Recall	F-Measure
D (Run-1)	0.342	0.580	0.430
W+D	NA	NA	NA
D+T (Run-2)	0.474	**0.804**	**0.596**

[6] $\theta = 15.0$

Results

Table: Results

(a) Results on training data

Method	Precision	Recall	F-Measure
D	0.4545	0.6923	0.5488
W+D	0.1717	0.2615	0.2073
D+T	0.5051	**0.7692**	0.6097
D+T+θ[6]	0.6069	0.6769	**0.6400**

(b) Results on test data

Method	Precision	Recall	F-Measure
D (Run-1)	0.342	0.580	0.430
W+D	NA	NA	NA
D+T (Run-2)	0.474	**0.804**	**0.596**
D+T+θ (Run-3)	0.439	0.607	0.509

[6] $\theta = 15.0$

Results

Table: Results

(a) Results on training data

Method	Precision	Recall	F-Measure
D	0.4545	0.6923	0.5488
W+D	0.1717	0.2615	0.2073
D+T	0.5051	**0.7692**	0.6097
D+T+θ[6]	0.6069	0.6769	**0.6400**

(b) Results on test data

Method	Precision	Recall	F-Measure
D (Run-1)	0.342	0.580	0.430
W+D	NA	NA	NA
D+T (Run-2)	0.474	**0.804**	**0.596**
D+T+θ (Run-3)	0.439	0.607	0.509

[6] $\theta = 15.0$

What for precision?

We introduce a similarity threshold empirically

(c) Effect of threshold θ on performance evaluation

θ Value	Precision	Recall	F-Measure
0.0	0.5051	0.7692	0.6097
9.0	0.5376	0.7692	0.6329
10.0	0.5371	0.7231	0.6164
15.0	**0.6069**	**0.6769**	**0.6400**
20.0	0.6635	0.5461	0.5991

Table: Results

Analysis

- Bi-lingual dictionary
  - itself was able to fetch recall around 0.58
  - if morphological analyzer is induced to see the root word in the dictionary, performance can even increase e.g. plurals, morphological variations etc
- Wordnet
  - Taking all the wordnet entries introduce a heavy topic drift
  - Only selected sense should be taken depending on the domain
- Transliteration
  - Certainly helps in improving recall

Conclusions

- Available resources are efficient for locating text re-use at document level
- Machine Translation is not necessary

Work Ahead

- Develop detailed analysis phase avoiding Machine Translation
- Work on the precision i.e. what not to call text re-use

Thank You! ☺