CENTRE FOR GLOBAL INTELLIGENT CONTENT

DCU at FIRE 2013: Cross Language !ndian News Story Search

Piyush Arora, Jennifer Foster, Gareth J. F. Jones
CNGL Centre for Global Intelligent Content
School of Computing, Dublin City University, Ireland

WWW.CNGL.IE

Outline

CENTRE FOR GLOBAL INTELLIGENT CONTENT

- **Introduction**
- Our Approach
- Experimental Details
- Results
- Conclusions and Future Work

Introduction

CL!NSS FIRE'13 task: News story linking between English and Indian Languages documents.

Outline

CENTRE FOR GLOBAL INTELLIGENT CONTENT

- Introduction
- **Our Approach**
- Experimental Details
- Results
- Conclusion and Future Work

Our Approach

CENTRE FOR GLOBAL INTELLIGENT CONTENT

The approach used by us has 2 main steps:

- *Step-1*: Follow traditional cross-language information retrieval (CLIR) approach:
  - Index documents using Lucene search engine.
  - Translate input query from source to target language using machine translation (MT)
  - Rank documents for retrieval using Lucene search engine
- *Step-2*: Combine multiple runs using data fusion methods

Contd…

Novel features of our approach

- Query modifications using different features such as:
  - Summarize query documents to form focused queries prior to translation
  - Identify Named Entities (NEs) as candidates for transliteration
  - Combine MT translation with NEs transliterations to capture alternative translations
- Adding weighting to reflect publication date relationship between query and target documents

Outline

CENTRE FOR GLOBAL INTELLIGENT CONTENT

- Introduction
- Our Approach
- **Experimental Details**
- Results
- Conclusion and Future Work

Experimental Details

CENTRE FOR GLOBAL INTELLIGENT CONTENT

Pre-Processing and Indexing

- Index documents using Lucene.
- Used Lucene's inbuilt Hindi Analyzer
- Stopword list obtained by concatenating the following:
  1. FIRE Hindi stopword list
  2. Lucene internal stopword list
  3. Stopword list created by selecting all words with Document Frequency (DF) > 5000

Contd…

CENTRE FOR GLOBAL INTELLIGENT CONTENT

Cross Language Search

- Input queries translated separately using:
  - Bing
  - Google

System	NDCG@1	NDCG@5	NDCG@10	NDCG@20
Palkosvi	0.32	0.33	0.34	0.36
Bing	0.54	0.52	0.53	0.55
Google	**0.56**	**0.55**	**0.56**	**0.58**

Baseline Results

Main Features Used For Query Modifications

CENTRE FOR GLOBAL INTELLIGENT CONTENT

Summarizer: based on extraction of sentences weighted using various factors indicating importance to document

- Varying length of summary
  - Summary length half of query document
  - Summary length one third of query document
  - Summary of top 3 ranked sentences from query document.
- Use alternative translation services: Bing, Google

Summarizer Features

Main Features used for summarizer:

— skimming: position of a sentence in a paragraph.

— namedEntity: number of named entities in each sentence.

— TSISF: similar to TF-IDF function but works at sentence level.

— titleTerm: overlap between the sentences and the terms in the title of a document.

— clusterKeyword: relatedness between words in a sentence.

Contd…

CENTRE FOR GLOBAL INTELLIGENT CONTENT

Transliteration

English Word	Translated Word	Transliterated Word
Games	खेल	गेम्स
Commonwealth	राष्ट्रमंडल	कॉमनवेल्थ

Using Date

Adding a constant of 0.04 to the retrieved documents occurring in a window of 10 days of the query document.

Feature Selection

CENTRE FOR GLOBAL INTELLIGENT CONTENT

— *Using Google translation*

- Using 1/3 summary
- Using 3-sentence summary
- Using 3-sentence summary + all NE transliterated
- Using complete input query + all NE transliterated

— *Using Bing translation*

- Using 1/3 summary
- Using 3-sentence summary
- Using complete input query + all NE transliterated

System	NDCG@1	NDCG@5	NDCG@10	NDCG@20
1/3 summary	0.5408	0.5814	**0.5872**	0.5907
1/3 summary+ NE transliterated	0.5408	0.5757	0.5828	**0.5957**
3-sentence summary	**0.5918**	**0.5815**	0.5855	0.5897
Complete query +NE transliterated	0.5714	0.562	0.5743	0.591

Results Using Google Translation

System	NDCG@1	NDCG@5	NDCG@10	NDCG@20
3-sentence summary	**0.5612**	**0.556**	0.5623	**0.5734**
1/3 summary	0.551	0.555	**0.5639**	0.5721
Complete query + NE transliterated	0.5102	0.5315	0.5463	0.5574

Results Using Bing Translation

Data Fusion

CENTRE FOR GLOBAL INTELLIGENT CONTENT

Data Fusion: The scores of documents retrieved using different systems are normalized and then combined to give a final score.

Min-Max algorithm for normalization:

$$Normalized\ Score = \frac{unormalized\ score - minimum\ score}{maximum\ score - minimum\ score}$$

CombMNZ approach for data fusion:

$$Summation\ \ of\ individual\ retrieval\ results\ * \\ number\ of\ non\ zero\ retrievals$$

Top 3 feature/system combinations selected:

- ***Run-1:*** Using Google translation and 1/3 summary of input query.
- ***Run-2:*** Using Google translation and combining 1/3 summary with and without NE transliterated, 3-sentence summary and using whole query + incorporating date factor.
- ***Run-3:*** Combining all the features, i.e. including queries translated using both Google and Bing using complete query as well as 1/3 summary and 3-sentence summary with and without NE transliterated.

System	NDCG@1	NDCG@5	NDCG@10	NDCG@20
Run-1	0.5408	0.5814	0.5872	0.5907
Run-2	**0.6224**	**0.5835**	**0.5943**	**0.6022**
Run-3	**0.6224**	0.5733	0.5833	0.5956

Results on Training set

Outline

CENTRE FOR GLOBAL INTELLIGENT CONTENT

- Introduction
- Our Approach
- Experimental Details
- **Results**
- Conclusion and Future Work

Results on test set

Evaluation- Submitted runs blind – submission combinations selected using features that performed best on the training set

System	NDCG@1	NDCG@5	NDCG@10	NDCG@20
Run-1	0.74	0.66587	0.6759	0.6849
Run-2	0.74	0.6701	0.7047	0.7042
Run-3	**0.74**	**0.6809**	**0.7268**	**0.7249**

Results on Test set

Outline

CENTRE FOR GLOBAL INTELLIGENT CONTENT

- Introduction
- Our Approach
- Experimental Details
- Results
- **Conclusion and Future Work**

Conclusion & Future Work

CENTRE FOR GLOBAL INTELLIGENT CONTENT

Future Work:

- Handling abbreviations such as “MNK”, “YSR”, political party names, movie names, etc.
- Handling spelling variants.
- Normalizing text, handling language variations.
- Minimizing translation and transliteration error.
- Explore alternative scoring functions such as BM25.
- Weighting different features rather than linearly scoring them.

CENTRE FOR GLOBAL INTELLIGENT CONTENT

Thank You

Questions?

This research is supported by Science Foundation Ireland (SFI) as a part of the CNGL Centre for Global Intelligent Content at DCU (Grant NO: 12/CE/I2267)