A Language-Independent Approach to Identify the Named Entities in under-resourced languages and Clustering Multilingual Documents

Kiran Kumar N
Santosh GSK
Vasudeva Varma

Motivation

Multilingual Document Clustering (MDC)

- Aggregates similar content across the languages.
- Helps in improving the usefulness of the content about a particular topic
  - -Cross Lingual Information Retrieval (CLIR)

during the period 200 BCE–20
dynasties that would trade exte
same time, Hinduism assertec
e fourth and fifth centuries CE,
Ganges Plain that would beco
ed on devotion rather than the

१७वीं शताब्दी के मध्यकाल में
प्राप्त किया। अंग्रेज दूसरे देशों
इंडिया कम्पनी के विरुद्ध असफ
बीसवी सदी के प्रारम्भ मे आधु
हुआ जिसने सामाजिक और रा
एक औपचारिक स्वरूप दिया।

భారత గణతంత్ర రాజ్యము న
భారత దేశ ప్రాముఖ్యత గత
నాలుగో స్థానంలో ఉంది. ప్ర
భారతదేశం, ప్రపంచం లోనే
ఆవిర్భవించినది. ..

Challenges

- Availability of bilingual dictionaries is limited.
- Coverage of Named Entities in any language dictionary is very less.
- Language independent tools don't exist.
  - Lemmatizers, POS taggers, etc. are language specific.

Traditional Document Clustering

- Documents represented as “bag of words" (BoW).
- Problem: All the terms present in a document are given equal importance.
- Ex: Documents sharing some collection of terms and representing the same topic may be falsely assigned to different clusters.

Contd…

- Reason: Lack of the identification of important terms, which represent the topic of that document.

- Solution: Giving high priority to the terms which helps in representing the topic of that document.

Multilingual Document Clustering (MDC)

- Our approach can be fundamentally broken down into two phases

1. Identification of the Named Entities

2. Clustering multilingual documents

Related Work

Identifying Named Entities (NEs) in multilingual documents

- Montalvo et al., 2006 used Freeling tool and common NE recognizer for English and Spanish to identify the NEs present in both the language.

- Romaric et al., 2004 performed linguistic analysis such as lemmatization, morphological analysis to recognize the NEs present in the data.

Contd…

- Negri and Magnini, 2004 have used the aligned English-Italian WordNet predicates present in Multi WordNet for Multilingual named entity recognition.

- In all the above systems discussed, the authors used language dependent resources or tools to extract the NEs present in the data.

- Hence, such systems face the problem of extendibility of their approaches.

Our Approach

1. Identification of the Named Entities

2. Clustering multilingual documents

Phase -1: NE Identification

- We propose a language-independent approach to identify the Named Entities present in under-resourced Indian languages (Hindi and Marathi)

- Named Entities present in English (a high resourced language) are utilized for this purpose.

Contd…

- All Named Entities present in English documents are identified using Stanford NER.

- In order to identify the NEs present in non-English documents, the NEs present in all English documents are utilized.

- All the non-English words after being translated into English are compared with the NEs in English documents and words which have an exact match are identified as the NEs of corresponding non-English documents.

Contd…

- NE separator function is used to represent each document in the dataset with two vectors namely a NE vector and a nonNE vector.
- The NE vector contains only NEs present in the document.
- Whereas, the nonNE vector contains the remaining words of that document.
- In both these vectors, the values are TFIDF scores.

Document Similarity

- Cosine similarity measure is used.
- Overall similarity between document d_i and d_j is calculated as:

$$Overall\_Sim(d_i, d_j) = \alpha * sim(d_i,d_j)^{NE} + \beta * sim(d_i,d_j)^{Category} \quad \text{-Eq. (1)}$$

Where $\alpha + \beta = 1$.

Contd…

$$sim(d_i,d_j) = \cos(v_i,v_j) = \frac{v_i.v_j}{|v_i|\ |v_j|} \quad \text{– Eq.(2)}$$

where v_i and v_j belong to NE, nonNE vectors of documents d_i and d_j respectively.

- Any two terms are compared using the Modified Levenshtein Edit Distance measure.

- Coefficients α and βare determined experimentally

Modified Levenshtein Edit Distance Measure (MLED)

- Replace the purpose of Lemmatizers.
- Helps in matching a word in its inflected form with its base form or other inflected forms.
- **Ex:** In English, the verb `walk' may appear in various inflected forms such as `walked', `walks', `walking'.

Contd…

- The rules are very intuitive and are based on three aspects:
    1. Minimum length of the two words
    2. Actual Levenshtein distance between the words
    3. Length of subset string match, starting from first letter.

Our Approach

1. Identification of the Named Entities

2. Clustering multilingual documents

Multilingual Document Clustering based on Named Entities (MDC_{NE})

- Steinbach et al. 2000 compared different clustering algorithms and concluded that Bisecting k-means performs better than the standard k-means and agglomerative hierarchical clustering.

- We used Bisecting k-means algorithm to form multilingual clusters where equation (1) is used in order to compare two documents efficiently.

- All Hindi, Marathi documents are mapped into English using Shabdanjali dictionary, Marathi-Hindi dictionary and Wiki dictionary.

Wiki dictionary

- Proper nouns play a pivotal role in measuring the similarity between two given documents.
- Dictionaries, in general, don't cover many proper nouns.
- We availed cross-lingual links of aligned Wikipedia titles and built a Wiki dictionary.
  - In order to handle proper nouns.

Our Approach: Advantages

- Our approach is scalable to other languages with relative ease.
  - Wikipedia acts as a conceptual interlingua with its cross lingual links
  - Avoided usage of language-specific tools by creating alternatives like MLED.
- It also addresses the future growth of multilingual information
  - Any first story or hot topic gets dynamically updated in Wikipedia.

Data sets

FIRE 2010 dataset available for the ad-hoc cross lingual document retrieval task.

- Consists of total 2182 documents
  - 650 English documents
  - 913 Hindi documents
  - 619 Marathi documents

Wikipedia Data

- Wikipedia releases periodic dumps of its data for different languages.
- We used the Sept,'10 release dump consisting of 2 million English articles, 55,537 Hindi articles and 52,300 Marathi articles

Experimental Evaluation

- We have randomly selected 90 documents from Hindi and Marathi dataset. T
- Three experts from the linguistics department are given 30 documents each to manually identify the NEs present in those documents.
- NE Identification system is then evaluated using precision and recall.
- $NE_{Precision} = NEs_{correctlyIdentified} / NEs_{totalNEsIdentified}$
- $NE_{recall} = NEs_{correctlyIdentified} / NEs_{totalNEsPresent}$

Experimental Evaluation

- We used F-Score, Purity and Normalized Mutual Information (NMI) to evaluate our clusters.

- Our Clutering basic involves training and testing phases:

Training Phase

- Training data constitutes around 60% (1320 documents) of the total documents (2182) in the dataset.

- Out of these1320 documents, 400 documents are in English, 550 are in Hindi and 370 in Marathi.

Contd…

- The α and β values are determined by conducting experiments on the training data using Eq. (1).
- Bisecting k-means algorithm is performed on the training data by varying the α value from 0.0 to 1.0 with 0.1 increment (β = 1- α).
- Finally, α and β are set to the value for which best cluster results are obtained.

Contd…

- In our experiments, it was found that setting $\alpha = 0.8$ and $\beta = 0.2$ has yielded good results

Testing Phase

- Test data constitutes around 40% (862documents) of the total documents in the dataset
  - 250 English documents
  - 363 Hindi documents
  - 249 Marathi documents

- Bisecting k-means algorithm is performed on the test data, after setting the α and β values obtained in training phase, using Eq. (1) in similarity calculation.

Results

- Evaluation results of NE Identification System using different dictionaries

NE Identification Measure	MA dictionaries	Wiki dictionaries
$NE_{Precision}$	70.23	78.34
NE_{Recall}	65.33	70.13

Here MA* = manually annotated dictionaries such as Shabdanjali dictionary and Marathi-Hindi dictionary

Contd…

- Evaluation of the Clustering schemes formed using different dictionaries

	MA dictionaries		Wiki dictionaries	
Evaluation Measure	**$MDC_{Keywords}$**	**MDC_{NE}**	**$MDC_{Keywords}$**	**MDC_{NE}**
F-Score	0.504	0.548	0.662	0.705
Purity	0.582	0.614	0.737	0.771
NMI	0.626	0.661	0.761	0.798

Conclusions

- From the results it can be concluded that Clustering based on Named Entities (MDC_{NE}) outperform the clustering based on all key words present in a document ($MDC_{Keywords}$).
- NEs alone are not sufficient for forming better clusters, NEs when combined along with the nonNEs have yielded better clustering results.
- Proposed approach is completely language independent
- Created alternatives like Wiki dictionary, MLED, etc. to ensure the accuracy.

Future Work

- We plan to extend the proposed approach which implements only static clustering to handle the dynamic clustering of multilingual documents.

- Also, we would like to consider comparable corpora of different languages to study the applicability of our approach.