Identifying Similar and Co-referring Documents Across Languages

Pattabhi R K Rao T
AU-KBC Research Centre,
MIT Campus, Anna University,
Chennai-44, India.
pattabhi@au-kbc.org

Sobha L
AU-KBC Research Centre,
MIT Campus, Anna University,
Chennai-44, India.
sobha@au-kbc.org

Abstract

This paper presents a methodology for finding similarity and co-reference of documents across languages. The similarity between the documents is identified according to the content of the whole document and co-referencing of documents is found by taking the named entities present in the document. Here we use Vector Space Model (VSM) for identifying both similarity and co-reference. This can be applied in cross-lingual search engines where users get documents of very similar content from different language documents.

1 Introduction

In this age of information technology revolution, the growth of technology and easy accessibility has contributed to the explosion of text data on the web in different media forms such as online news magazines, portals, emails, blogs etc in different languages. This represents 80% of the unstructured text content available on the web. There is an urgent need to process such huge amount of text using Natural Language Processing (NLP) techniques. One of the significant challenges with the explosion of text data is to organize the documents into meaningful groups according to their content.

The work presented in this paper has two parts a) finding multilingual cross-document similarity and b) multilingual cross-document entity co-referencing. The present work analyzes the documents and identifies whether the documents are similar and co-referring. Two objects are said to be similar, when they have some common properties between them. For example, two geometrical figures are said to be similar if they have the same shape. Hence similarity is a measure of degree of resemblance between two objects.

Two documents are said to be similar if their contents are same. For example a document D1 describes about a bomb blast incident in a city and document D2 also describes about the same bomb blast incident, its cause and investigation details, then D1 and D2 are said to be similar. But if document D3 talks of terrorism in general and explains bomb blast as one of the actions in terrorism and not a particular incident which D1 describes, then documents D1 and D3 are dissimilar. The task of finding document similarity differs from the task of document clustering. Clustering is a task of categorization of documents based on domain/field. In the above example, documents D1, D2, D3 can be said to be in a cluster of crime domain. When documents are similar they share common noun phrases, verb phrases and named entities. While in document clustering, sharing of named entities and noun phrases is not essential but still there can be some noun phrases and named entities in common. Cross-document co-referencing of entities refers to the identification of same entities across the documents. When the named entities present in the documents which are similar and also co-referencing, then the documents are said to be co-referring documents.

The paper is further organized as follows. In section 2, the motivation behind this paper is explained and in 3 the methodology used is described. Results and discussions are dealt in section 4 and conclusion in section 5.

2 Motivation

Dekang Lin (1998) defines similarity from the information theoretic perspective and is applicable if the domain has probabilistic model. In the past decade there has been significant amount of work done on finding similarity of documents and organizing the documents according to their content. Similarity of documents are identified using different methods such as Self-Organizing Maps (SOMs) (Kohonen et al, 2000; Rauber, 1999), based on Ontologies and taxanomy (Gruber, 1993; Resnik, 1995), Vector Space Model (VSM) with similarity measures like Dice similarity, Jaccard's similarity, cosine similarity (Salton, 1989). Bagga (Bagga et al., 1998) have used VSM in their work for finding co-references across the documents for English documents. Chung and Allan (2004) have worked on cross-document co-referencing using large scale corpus, where they have said ambiguous names from the same domain (here for example, politics) are harder to disambiguate when compared to names from different domains. In their work Chung and Allan compare the effectiveness of different statistical methods in cross-document co-reference resolution task. Harabagiu and Maiorano (2000) have worked on multilingual co-reference resolution on English and Romanian language texts. In their system, "SWIZZLE" they use a data-driven methodology which uses aligned bilingual corpora, linguistic rules and heuristics of English and Romanian documents to find co-references. In the Indian context, obtaining aligned bilingual corpora is difficult. Document similarity between Indian languages and English is tough since the sentence structure differs and Indian languages are agglutinative in nature. In the recent years there has been some work done in the Indian languages, (Pattabhi et al, 2007) have used VSM for multilingual cross-document co-referencing, for English and Tamil, where no bilingual aligned corpora is used.

One of the methods used in cross-lingual information retrieval (CLIR) is Latent Semantic Analysis (LSA) in conjunction with multilingual parallel aligned corpus. This approach works well for information retrieval task where it has to retrieve most similar document in one language to a query given in another language. One of the drawbacks of using LSA in multilingual space for the tasks of document clustering, document similarity is that it gives similar documents more based on the language than by topic of the documents in different languages (Chew et al, 2007). Another drawback of LSA is that the reduced dimension matrix is difficult to interpret semantically. The examples in Table 1, illustrate this.

	Before Reduction	After Reduction
1.	{(car),(truck),(flower)}	{(1.2810*car+0.5685*truck),(flower)}
2	{(car),(bottle),(flower)}	{(1.2810*car+0.5685*bottle),(flower)}

Table 1. LSA Example

In the first example the component _(1.2810*car+0.5685*truck)_ can be inferred as "Vehicle" but in cases such as in second example, the component _(1.2810*car+0.5685*bottle)_ does not have any interpretable meaning in natural language. In LSA the dimension reduction factor '_k_' has very important role to play and the value of 'k' can be found by doing several experiments. The process of doing dimension reduction in LSA is computationally expensive. When LSA is used, it reduces the dimensions statistically and when there is no parallel aligned corpus, this can not be interpreted semantically.

Hence, in the present work, we propose VSM which is computationally simple, along with cosine similarity measure to find document similarity as well as entity co-referencing. We have taken English and three Dravidian languages viz. Tamil, Telugu and Malayalam for analysis.

3 Methodology

In VSM, each document is represented by a vector which specifies how many times each term occurs in the document (the term frequencies). These counts are weighted to reflect the importance of each term and weighting is the inverse document frequency (idf). If a term t occurs in n documents in the collection then the "_idf_" is the inverse of log n. This vector of weighted counts is called a "bag of words" representation. Words such as "stop words" (or function words) are not included in the representation.

The documents are first pre-processed, to get syntactic and semantic information for each word in the documents. The preprocessing of documents involves sentence splitting, morph analysis, part-of-speech (POS) tagging, text chunking and named entity tagging. The documents in English are preprocessed using Brill's Tagger (Brill, 1994) for POS tagging and fn-TBL (Ngai and Florian, 2001) for text chunking. The documents in Indian languages are preprocessed, using a generic engine (Arulmozhi et al., 2006) for POS tagging, and text chunking based on TBL (Sobha and Vijay, 2006). For both English and Indian language documents the named entity tagging is done using Named Entity Recognizer (NER) which was developed based on conditional random field (CRF). The tagset used by the NER tagger is a hierarchical tagset, consists of mainly i) ENAMEX, ii) NUMEX and iii) TIMEX. Inside the ENAMEX there are mainly 11 subtype's viz. a) Person b) Organization c) Location d) Facilities e) Locomotives f) Artifacts g) Entertainment h) Cuisines i) Organisms j) Plants k) Disease. For the task of multilingual cross-document entities co-referencing, the documents are further processed for anaphora resolution where the corresponding antecedents for each anaphor are tagged in the document. For documents in English and Tamil, anaphora resolution is done using anaphora resolution system. For documents in Malayalam and Telugu anaphora resolution is done manually. After the preprocessing of documents, the language model is built by computing the term frequency – inverse document frequency (tf-idf) matrix. For the task of finding multilingual cross-document similarity, we have performed four different experiments. They are explained below:

E1: The terms are taken from documents after removing the stop words. These are raw terms where no preprocessing of documents is done; the terms are unique words in the document collection.

E2: The terms taken are the words inside the noun phrases, verb phrases and NER expressions after removing the stop words.

E3: The whole noun phrase/verb phrase/NER expression is taken to be a single term.

E4: The noun phrase/NER expression along with the POS tag information is taken as a single term.

The first experiment is the standard VSM implementation. The rest three experiments differ in the way the terms are taken for building the VSM. For building the VSM model which is common for all language document texts, it is essential that there should be translation/transliteration tool. First the terms are collected from individual language documents and a unique list is formed. After that, using the translation/transliteration tool the equivalent terms in language L2 for language L1 are found. The translation is done using a bilingual dictionary for the terms present in the dictionary. For most of the NERs only transliteration is possible since those are not present in the dictionary. The transliteration tool is developed based on the phoneme match it is a rule based one. All the Indian language documents are represented in roman notation (wx-notation) for the purpose of processing.

After obtaining equivalent terms in all languages, the VSM model is built. Let S1 and S2 be the term vectors representing the documents D1 and D2, then their similarity is given by equation (1) as shown below.

$$\mathrm{Sim}(S1,S2) = \sum_{tj} (W_{1j} \text{ x } W_{2j}) \qquad \text{-- (1)}$$

Where,

tj is a term present in both vectors S1and S2.
W_{1j} is the weight of term tj in S1 and
W_{2j} is the weight of term tj in S2.

The weight of term tj in the vector S1 is calculated by the formula given by equation (2), below.

$$Wij=(tf*\log(N/df))/[\mathrm{sqrt}(S_{i1}^2+S_{i2}^2+\ldots\ldots+S_{in}^2)] \text{ --(2)}$$

Where,

tf = term frequency of term t_j
N=total number of documents in the collection
df = number of documents in the collection that the term t_j occurs in.
sqrt represents square root

The denominator $[\mathrm{sqrt}(S_{i1}^2+S_{i2}^2+\ldots\ldots+S_{in}^2)]$ is the cosine normalization factor. This cosine normalization factor is the Euclidean length of the vector $S_{i,}$ where 'i' is the document number in the collection and S_{in}^2 is the square of the product of (tf*log(N/df)) for term t_n in the vector S_i.

For the task of multilingual cross-document entity co-referencing, the words with-in the anaphor tagged sentences are considered as terms for building the language model.

4 Results and Discussion

The corpus used for experiments is collected from online news magazines and online news portals. The sources in English include "The Hindu", "Times of India", "Yahoo News", "New York Times", "Bangkok Post", "CNN", "WISC", "The

Independent”. The sources for Tamil include “Dinamani”, “Dinathanthi”, “Dinamalar”, “Dinakaran”, and “Yahoo Tamil”. The work was primarily done using English and Tamil. Later on this was extended for Malayalam and Telugu. The data sources for Malayalam are “Malayala Manorama”, “Mathrubhumi”, “Deshabhimani”, “Deepika” and sources for Telugu include “Eenadu”, “Yahoo Telugu” and “Andhraprabha”. First we discuss about English and Tamil and Later Telugu and Malayalam.

The domains of the news taken include sports, business, politics, tourism etc. The news articles were collected using a crawler, and hence we find in the collection, a few identical news articles because they appear in different sections of the news magazine like in Front page section, in state section and national section.

The dataset totally consists of 1054 English news articles, 390 Tamil news articles. Here we discuss results in two parts; in the first part results pertaining to document similarity are explained. In second part we discuss results on multilingual cross-document entity co-referencing.

4.1 Document Similarity

The data collection was done in four instances, spread in a period of two months. At the first instance two days news was crawled from different news sources in English as well as Tamil. In the first set 1004 English documents and 297 Tamil documents were collected.

In this set when manually observed (human judgment) it was found that there are 90 similar documents forming 31 groups, rest of the documents were not similar. This is taken as gold standard for the evaluation of the system output.

As explained in the previous section, on this set the four experiments were performed. In the first experiment (E1), no preprocessing of the documents was done except that the stop words were removed and the language model was built. In this it was observed that the number of similar documents is 175 forming 25 groups. Here it was observed that along with actual similar documents, system also gives other not similar documents (according to gold standard) as similar ones. This is due to the fact there is no linguistic information given to the system, hence having words alone does not tell the context, or in which sense it is used. And apart from that named entities when split don’t give exact meaning, for example in name of hotels “Leela Palace” and “Mysore Palace”, if split into words yields three words, “Leela”, “Mysore”, and “Palace”. In a particular document, an event at hotel Leela Palace is described and the hotel is referred as Leela Palace or by Palace alone. Another document describes about Dussera festival at Mysore Palace. Now here the system identifies both these documents to be similar even though both discuss about different events. The precision of the system was observed to be 51.4%, where as the recall is 100% since all the documents which were similar in the gold standard is identified. Here while calculating the precision; we are considering the number of documents that are given by the system as similar to the number of documents similar according to the gold standard.

Hence to overcome the above discussed problem, we did the second experiment (E2) where only words which occur inside the noun phrases, verb phrases and named entities are considered as terms for building the language model. Here it is observed that the number of similar documents is 140 forming 30 groups. This gives a precision of 64.2% and 100% recall. Even though we find a significant increase in the precision but still there are large number of false positives given by the system. A document consists of noun phrases and verb phrases, when the individual tokens inside these phrases are taken; it is equivalent to taking almost the whole document. This reduces the noise. The problem of “Leela Palace” and “Mysore Palace” as explained in the previous paragraph still persists here.

In the third experiment (E3) the whole noun phrase, verb phrase and named entity is considered as a single term for building the language model. Here the phrases are not split into individual tokens; the whole phrase is a single term for language model. This significantly reduces the number of false positives given by the system. The system identifies 106 documents as similar documents forming 30 groups. Now the precision of the system is 84.9%. In this experiment, the problem of “Leela Palace” and “Mysore Palace” is solved. Though this problem was solved the precision of the system is low, hence we performed the fourth (E4) experiment.

In the fourth experiment (E4), the part-of-speech (POS) information is given along with the phrase for building the language model. It is observed that the precision of the system increases. The number of similar documents identified is 100 forming 31 groups. This gives a precision of 90% and a recall of 100%.

Another important factor which plays a crucial role in implementation of language model or VSM is the threshold point. What is the threshold point that is to be taken? For obtaining an answer for this question, few experiments were performed by setting the threshold at various points in the range 0.75 to 0.95. When the threshold was set at 0.75 the number of similar documents identified by the system was larger, not true positives but instead false positives. Hence the recall was high and precision was low at 50%. When the threshold was moved up and set at 0.81, the number of similar documents identified was more accurate and the number of false positives got reduced. The precision was found to be 66%. When the threshold was moved up still further and set at 0.90, it was found that the system identified similar documents which were matching with the human judgment. The precision of the system was found to be 90%. The threshold was moved up further to 0.95, thinking that the precision would further improve, but this resulted in documents which were actually similar to be filtered out by the system. Hence the threshold chosen was 0.9, since the results obtained at this threshold point had matched the human judgment. For the experiments E1, E2, E3 and E4 explained above, the threshold is fixed at 0.9.

A new set of data consisting of 25 documents from 5 days news articles is collected. This is completely taken from single domain, terrorism. These news articles describe specifically the Hyderabad bomb blast, which occurred on August 25th 2007. All these 25 documents were only English documents from various news magazines. This data set was collected specifically to observe the performance of the system, when the documents belonging to single domain are given. In the new data set, from terrorism domain, human judgment for document similarity was found to have 13 similar documents forming 3 groups. While using this data set the noun phrases, verb phrases and named entities along with POS information were taken as terms to build the language model and the threshold was set at 0.9, it was observed that the system finds 14 documents to be similar forming 3 groups. Here, out of 14 similar documents, only 12 documents match with the human judgment and one document which ought to be identified was not identified by the system. The document which was not identified described about the current event, that is, bomb blast on 25th August in the first paragraph and then the rest of the document described about the similar events that occurred in the past. Hence the similarity score obtained for this document with respect to other documents in the group was 0.84 which is lower than the threshold fixed. Hence the recall of the system is 92.3% and the precision of the system is 85.7%.

Another data set consisting of 114 documents was taken from tourism domain. The documents were both in Tamil and English, 79 documents in Tamil and 35 documents in English. This data set describes various pilgrim places and temples in Southern India. The human annotators have found 21 similar documents which form a group of three. These similar documents describe about Lord Siva's and Lord Murugan's temples. The system obtained 25 documents as similar and grouped into three groups. Out of 25 documents obtained as similar, four were dissimilar. These dissimilar documents described non-Siva temples in the same place. In these dissimilar documents the names of offerings, festivals performed were referred by the same names as in the rest of the documents of the group, hence these documents obtained similarity score of 0.96 with respect to other documents in the group. Here we get a precision of 84% and a recall of 100%.

A new data set consisting of 46 documents was taken from various news magazines. This set consists of 24 English documents, 11 Tamil documents, 7 Malayalam documents and 4 Telugu documents. This data set describes the earthquake in Indonesia on 12th September 2007 and tsunami warning in other countries. The news articles were collected on two days 13th and 14th September 2007.

The documents collected were in different font encoding schemes. Hence before doing natural language processing such as morph-analysis, POS tagging etc, the documents were converted to a common roman notation (wx-notation) using the font converter for each encoding scheme.

Here we have used multilingual dictionaries of place; person names etc for translation. The language model is built by taking noun phrases and verb phrases along with POS information were as terms. In this set human annotators have found 45 documents to be similar and have grouped them into one group. The document which was identified as dissimilar describes about a Tamil film shooting at Indonesia being done during the quake time. The system had identified all the 46 documents including the film shooting document in the collection to be similar and put into one group. The "*film shooting*" document consisted of two paragraphs about the quake incident, other two paragraphs consisted of statement by the film producer stating that the whole crew is safe and the shooting is temporarily suspended for next few days. Since this document also contained the content describing the earthquake found in other documents of the group, the system identified this *"film shooting"* document to be similar. Here one interesting point which was found was that all the documents gave a very high similarity score greater than 0.95. Hence the precision of the system is 97.8% and recall 100%.

The summary of all these experiments with different dataset is shown in the table 2 below.

SNo	**Dataset**	**Precision %**	**Recall %**
1	English 1004 and Tamil 297 documents	90.0	100.0
2	English 25 – terrorism domain documents	85.7	92.3
3	35 English Docs and Tamil 79 docs - Tourism domain	84.0	100.0
4	46 Docs on Earth Quake incident – 24 English, 11 Tamil, 7 Malayalam, 4 Telugu	97.8	100.0
	Average	89.3 %	98.07%

Table 2. Summary of Results for Document similarity for four different data sets

4.2 Document Co-referencing

The documents that were identified as similar ones are taken for entity co-referencing. In this work the identification of co-referencing documents is done for English and Tamil. In this section first we discuss the co-referencing task for English documents in terrorism domain, then for documents in English and Tamil in Tourism domain. In the end of this section we discuss about documents in English and Tamil, which are not domain specific.

In the first experiment, the document collection in terrorism domain is taken for co-referencing task. This data set of 25 documents in terrorism domain consists of 60 unique person names. In this work we consider only person names for entity co-referencing. In this data set, 14 documents are identified as similar ones by the system. These 14 documents consist of 26 unique person names. .

The language model is built using only named entity terms and the noun, verb phrases occurring in the same sentence where the named entity occurs. POS information is also provided with the terms. Here we find that out of 26 entities, the system co-references correctly for 24 entities, even though the last names are same. The results obtained for these named entities is shown in the below table Table 3.

Entity Name	**No. of links containing the entity**	**Correct Responses obtained**	**Total Responses obtained**	**Precision %**	**Recall %**
Y S Rajasekhar Reddy	7	7	7	100	100
Indrasena Reddy	1	1	1	100	100
K Jana Reddy	1	1	1	100	100
Shivaraj Patil	2	2	2	100	100
Manmohan Singh	4	4	4	100	100
Abdul Shahel Mohammad	1	1	2	50	100
Mohammad Abdullah	1	1	2	50	100
Mohammad Amjad	1	1	1	100	100
Mohammad Yunus	1	1	1	100	100
Ibrahim	1	1	1	100	100
Dawood Ibrahim	1	1	1	100	100
Madhukar Gupta	3	3	3	100	100
N Chandrababu Naidu	2	2	2	100	100
Tasnim Aslam	2	2	2	100	100
Mahender Agrawal	1	1	1	100	100
Somnath Chatterjee	2	2	2	100	100
Pervez Musharaff	2	2	2	100	100
Sonia Gandhi	2	2	2	100	100
Taslima	1	1	1	100	100

Nasrin					
Bandaru Dattatreya	1	1	1	100	100
L K Advani	2	2	2	100	100
Average				95.2	100

Table 3. Results for entity co-referencing for English documents in terrorism domain

The system identifies the entity names ending with “Reddy” correctly. These names in the documents occur along with definite descriptions which helps the system in disambiguating these names. For example “*Y S Rajasekhar Reddy*” in most cases is referred to as “*Dr. Reddy*” along with the definite description “*chief minister*”. Similarly the other name “*K Jana Reddy*” occurs with the definite description “*Home minister*”. Since here we are taking full noun phrases as terms for building language model, this helps obtaining good results. For entities such as “Abdul Shahel Mohammad” and “Mohammad Abdullah”, it is observed that the both names are referred in the documents as “Mohammad” and surrounding phrases do not have any distinguishing phrases such as definite descriptions, which differentiate these names. Both these entities have been involved in masterminding of the Hyderabad bomb blast. Hence the system couldn’t disambiguate between these two named entities and identifies both to be same, hence it fails here.

In the second experiment, the data set in Tourism domain consisting of 79 Tamil Documents and 35 English documents is taken for the task of co-referencing. In this data set 25 documents were identified as similar. Now these similar documents of 25 are considered for entity co-referencing task. There are 35 unique names of Gods. Here in this domain, one of the interesting points is that, there are different names to refer to a single God. For example Lord Murugan, is also referred by other names such as “Subramanyan”, “Saravana”, “Karttikeyan”, “Arumukan” etc. Simialrly for Lord Siva is referred by “Parangirinathar”, “Dharbaraneswara” etc. It is observed that in certain documents the alias names are not mentioned along with common names. In these instances even human annotators found it tough for co-referencing, hence the system could not identify the co-references. This problem of alias names can be solved by having a thesaurus and using it for disambiguation. The results obtained for these named entities are shown in the table 4, below.

Entity Name	**No. of links containing the entity**	**Correct Responses obtained**	**Total Responses obtained**	**Precision %**	**Recall %**
Murugan	7	7	8	87.5	100
Shiva	10	9	9	100	90
Parvathi	10	9	11	81.8	90
Nala	5	5	5	100	100
Damayanthi	2	2	2	100	100
Narada	3	3	3	100	100
Saneeswarar	6	6	7	85.7	100
Deivayani	4	4	4	100	100
Vishnu	2	2	2	100	100
Vinayaka	3	3	3	100	100
Indra	2	2	2	100	100
Thirunavukkarasar	1	1	1	100	100
Mayan	2	2	2	100	100
Average				96.5	98.4

Table 4. Results for entity co-referencing for English and Tamil Documents in Tourism domain

The co-referencing system could disambiguate a document which was identified as similar by the system and dissimilar by the human annotator.

Another experiment is performed where both English and Tamil Documents are taken for entity co-referencing. In this experiment we have taken the data set in which there are 1004 English documents and 297 Tamil documents. The documents are not domain specific. Here 100 documents are identified as similar ones, which contains of 64 English and 36 Tamil documents. Now we consider these 100 similar documents for entity co-referencing. In the 100 similar documents, there are 520 unique named entities. The table (Table 5) below shows results of few interesting named entities in this set of 100 similar documents.

Entity Name	**No. of links containing the entity**	**Correct Responses obtained**	**Total Responses obtained**	**Precision %**	**Recall %**
Karunanidhi	7	7	7	100	100
Manmohan Singh	15	14	16	87.5	93.3
Sonia Gandhi	54	54	58	93.1	100
Shivaraj Patil	8	8	10	80	100
Prathibha Patil	24	24	26	92.3	100
Lalu Prasad	5	5	5	100	100

Atal Bihari Vajpayee	4	4	4	100	100
Abdul Kalam	22	22	22	100	100
Sania Mirza	10	10	10	100	100
Advani	8	8	8	100	100
Average				95.3	99.3

Table 5. Results for entity co-referencing for English and Tamil Documents not of any specific domain

5 Conclusion

The VSM method is a well known statistical method, but here it has been applied for multilingual cross-document similarity, which is a first of its kind. Here we have tried different experiments and found that using phrases with its POS information as terms for building language model is giving good performance. In this we have got an average precision of 89.3 and recall of 98.07% for document similarity. Here we have also worked on multilingual cross-document entity co-referencing and obtained an average precision of 95.6 % and recall of 99.2 %. The documents taken for multilingual cross-document co-referencing are similar documents identified by the similarity system. Considering similar documents, helps indirectly in getting contextual information for co-referencing entities, because obtaining similar documents removes documents which are not in the same context. Hence this helps in getting good precision. Here we have worked on four languages viz. English, Tamil, Malayalam and Telugu. This can be applied for other languages too. Multilingual document similarity and co-referencing, helps in retrieving similar documents across languages.

References

Arulmozhi Palanisamy and Sobha Lalitha Devi. 2006. *HMM based POS Tagger for a Relatively Free Word Order Language*, Journal of Research on Computing Science, Mexico. 18:37-48.

Bagga, Amit and Breck Baldwin. 1998. *Entity-Based Cross-Document Coreferencing Using the Vector Space Model*, Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and the 17th International Conference on Computational Linguistics (COLING-ACL'98):79-85.

Brill, Eric. 1994. *Some Advances in transformation Based Part of Speech Tagging,* Proceedings of the Twelfth International Conference on Artificial Intelligence (AAAI-94), Seattle, WA

Peter A. Chew, Brett W. Bader, Tamara G. Kolda, Ahmed Abdelali. 2007. *Cross-Language Information Retrieval Using PARAFAC2,* In the Proceedings Thirteenth International Conference on Knowledge Discovery and Data Mining (KDD' 07), San Jose, California.:143-152.

Chung Heong Gooi and James Allan. 2004. *Cross-Document Coreference on a Large Scale Corpus,* Proceedings of HLT-NAACL: 9-16.

Dekang Lin. 1998. *An Information-Theoretic Definition of Similarity,* Proceedings of International Conference on Machine Learning, Madison, Wisconsin, July.

T. R. Gruber. 1993. *A translation approach to portable ontologies,* Knowledge Acquisition, 5(2):199–220.

Harabagiu M Sanda and Steven J Maiorano. 2000. *Multilingual Coreference Resolution,* Proceedings of 6th Applied Natural Language Processing Conference: 142–149.

Kohonen, Teuvo Kaski, Samuel Lagus, Krista Salojarvi, Jarkko Honkela, Jukka Paatero,Vesa Saarela, Anti. 2000. *Self organisation of a massive document collection*, IEEE Transactions on Neural Networks, 11(3): 574-585.

G. Ngai and R. Florian. 2001. *Transformation-Based Learning in the Fast Lane*, Proceedings of the NAACL'2001, Pittsburgh, PA: 40-47

R K Rao Pattabhi, L Sobha, and Amit Bagga. 2007. *Multilingual cross-document co-referencing,* Proceedings of 6th Discourse Anaphora and Anaphor Resolution Colloquium (DAARC), March 29-30, 2007, Portugal:115-119

Rauber, Andreas Merkl, Dieter. 1999. *The SOMLib digital library system,* In the Proceedings of the 3rd European Conference on Research and Advanced Technology for Digital Libraries (ECDL'99), Paris, France. Berlin: 323-341.

P. Resnik. 1995. *Using information content to evaluate semantic similarity in taxonomy,* Proceedings of IJCAI: 448–453.

Salton, Gerald. 1989. *Automatic Text Processing: The Transformation, Analysis and Retrieval of Information by Computer,* Reading, MA: Addison Wesley

Sobha L, and Vijay Sundar Ram. 2006. *Noun Phrase Chunker for Tamil,* Proceedings of the First National Symposium on Modeling and Shallow Parsing of Indian Languages (MSPIL), IIT Mumbai, India: 194-198.