CS697

UNSUPERVISED CROSS LINGUAL ALIGNMENT

Advisor: Prof. Satyadev Nandakumar

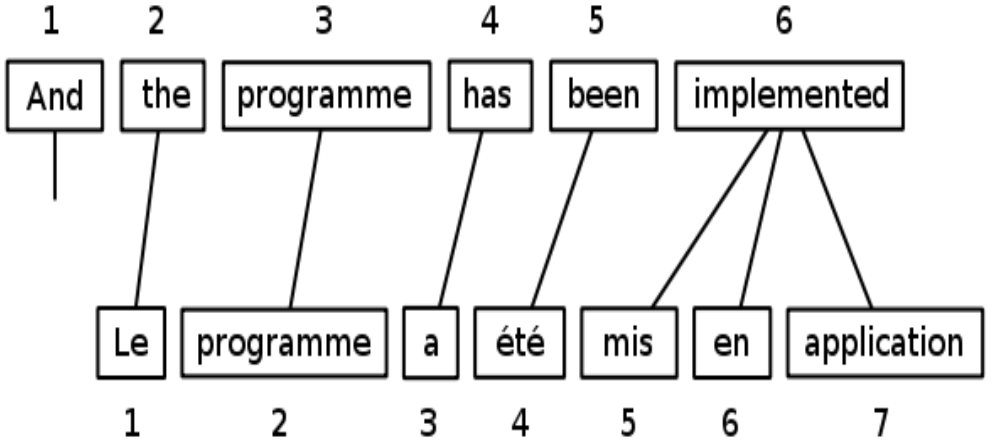
PRANJAL SINGH
10327511

CONTENT

- Motivation
- What we GAIN?
- Previous Work
- Dataset
- Introduction
- Formulation
- Problem
- Algorithm
- Scores
- Evaluation
- Base Results
- Reference

MOTIVATION

- Large volume of text in all languages
- Same information may not be present
- Alignment problem



Source: Wikipedia

WHAT WE GAIN???

- Can detect inconsistencies

- Same information can be broadcasted easily everywhere

PREVIOUS WORK

- **Cross Lingual Lexical Substitution:** *Apidianaki,* 2011
  - aims at providing for a target word in context, several alternative substitute words in another language
  - Unsupervised
  - identifies the senses of words by clustering their translations according to their semantic similarity
  - Evaluate on SemEval-2010 CLLS Dataset

PREVIOUS WORK

- **DbPedia:** *Auer,* 2008
  - Ongoing project
  - automatically extracts information from Wikipedia
  - normalizes the extracted information, links the information with other online data repositories
  - provides an interactive access.

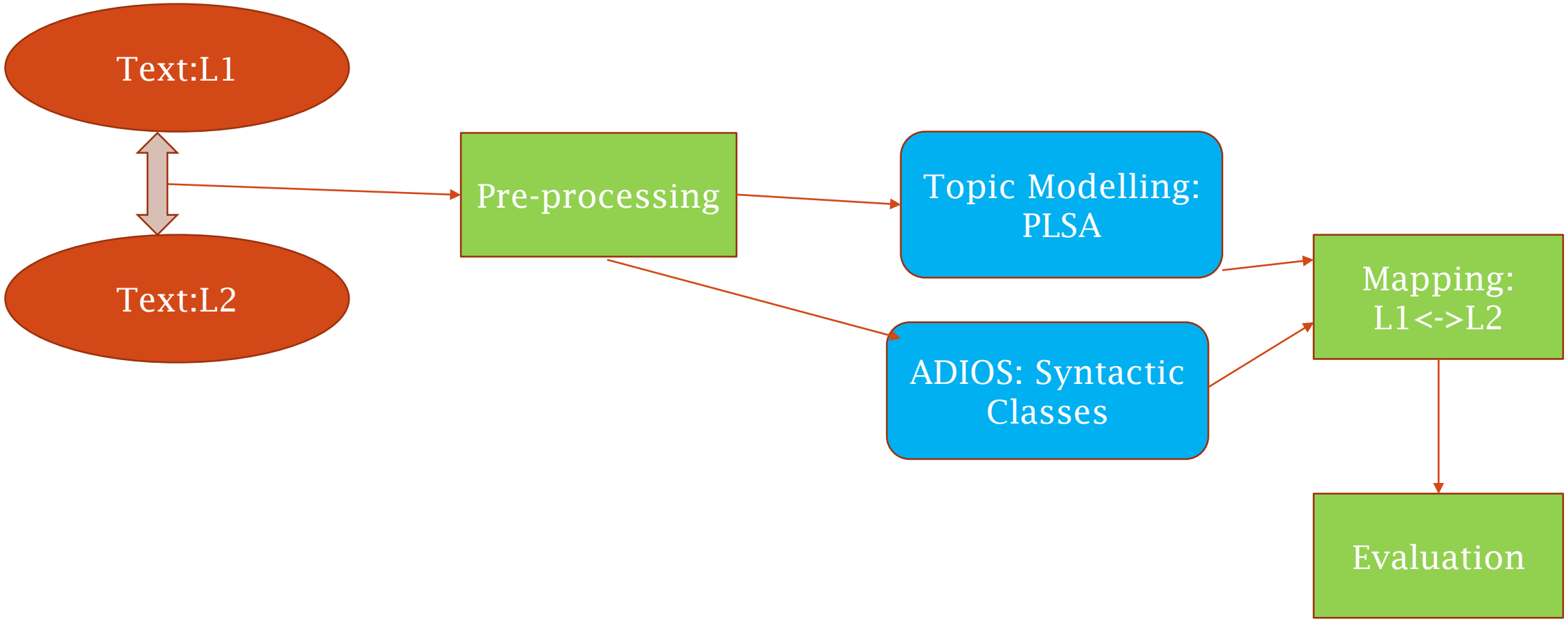
DATASET

- Manually built corpus on Coal Scam in Hindi and English
- Collected from various online Hindi and English newspaper and news channel websites such as *Dainik Jagran, Hindustan, Aaj Tak*, etc.

- English: ~52,000 tokens

- Hindi: ~36,000 tokens

INTRODUCTION



FEATURES(NLP)

- **Bag of Words Model**

Text (such as a sentence or a document) is represented as the bag (multiset) of its words, disregarding grammar and even word order but keeping multiplicity

L1: John likes to watch movies. Mary likes movies too.

L2: John also likes to watch football games.

```
{
   "John": 1,
   "likes": 2,
   "to": 3,
   "watch": 4,
   "movies": 5,
   "also": 6,
   "football": 7,
   "games": 8,
   "Mary": 9,
   "too": 10
}
```

Vector

L1: [1, 2, 1, 1, 2, 0, 0, 0, 1, 1]
L2: [1, 1, 1, 1, 0, 1, 1, 1, 0, 0]

FEATURES(NLP)

- **tf-idf Model (Term Frequency Inverse Document Frequency)**

tf(t,D): Raw frequency divided by the maximum raw frequency of any term in the document

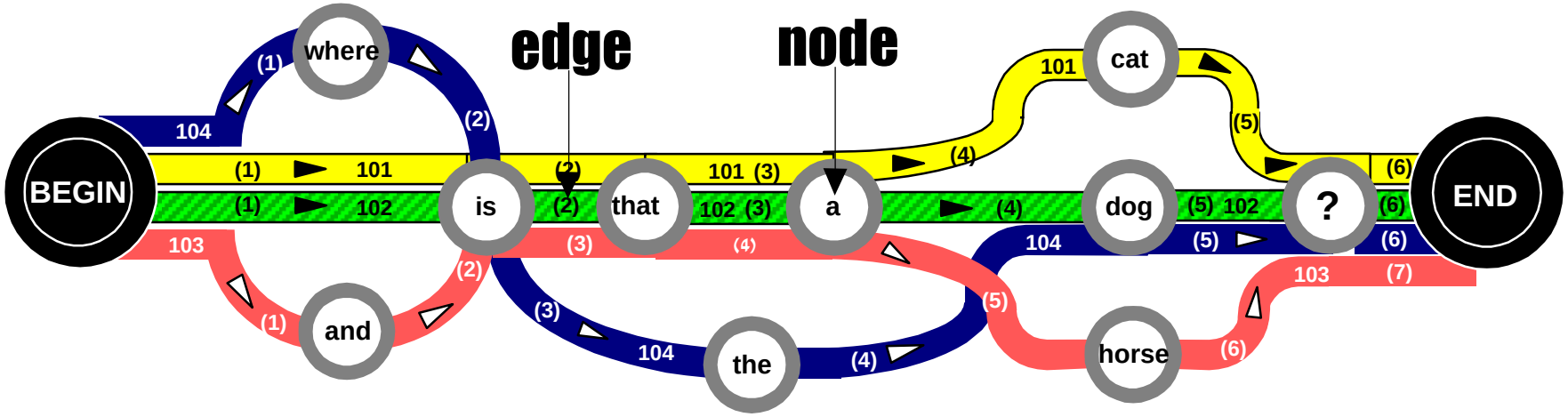
idf(t,D): It is obtained by dividing the total number of documents by the number of documents containing the term, and then taking the logarithm of that quotient

$$\textbf{tf-idf= tf(t,D) * idf(t,D)}$$

ADIOS (AUTOMATIC DISTILLATION OF STRUCTURE)

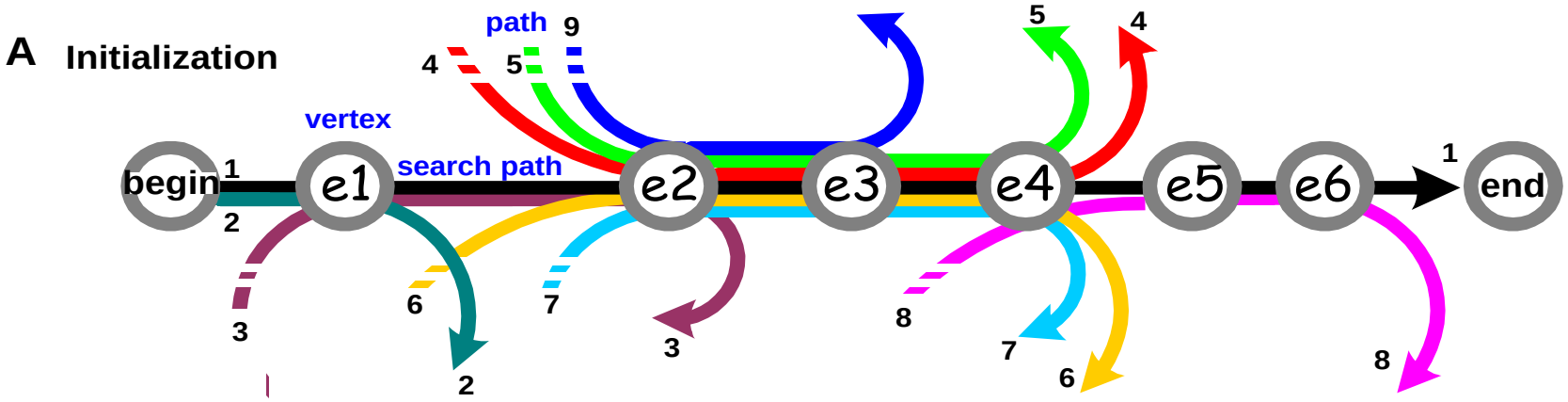
- ADIOS capable of learning complex syntax, generating grammatical novel sentences
- Proving useful in other fields that call for structure discovery from raw data, such as bioinformatics
- Composed of three main elements
  - A representational data structure
  - A segmentation criterion (MEX)
  - A generalization ability

ADIOS: THE MODEL



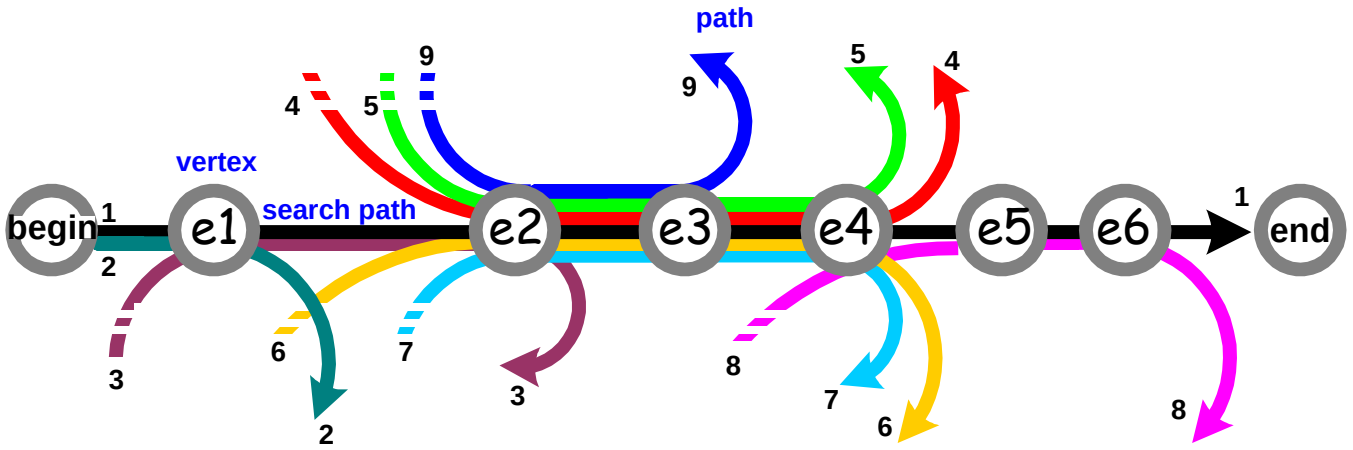
Is that a dog?

ADIOS

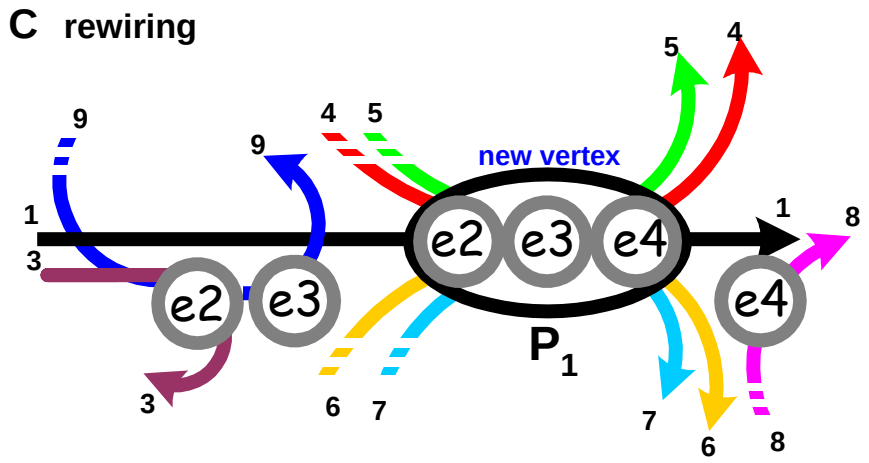


- Identifying patterns becomes easier on a graph
  - Sub-paths are automatically aligned

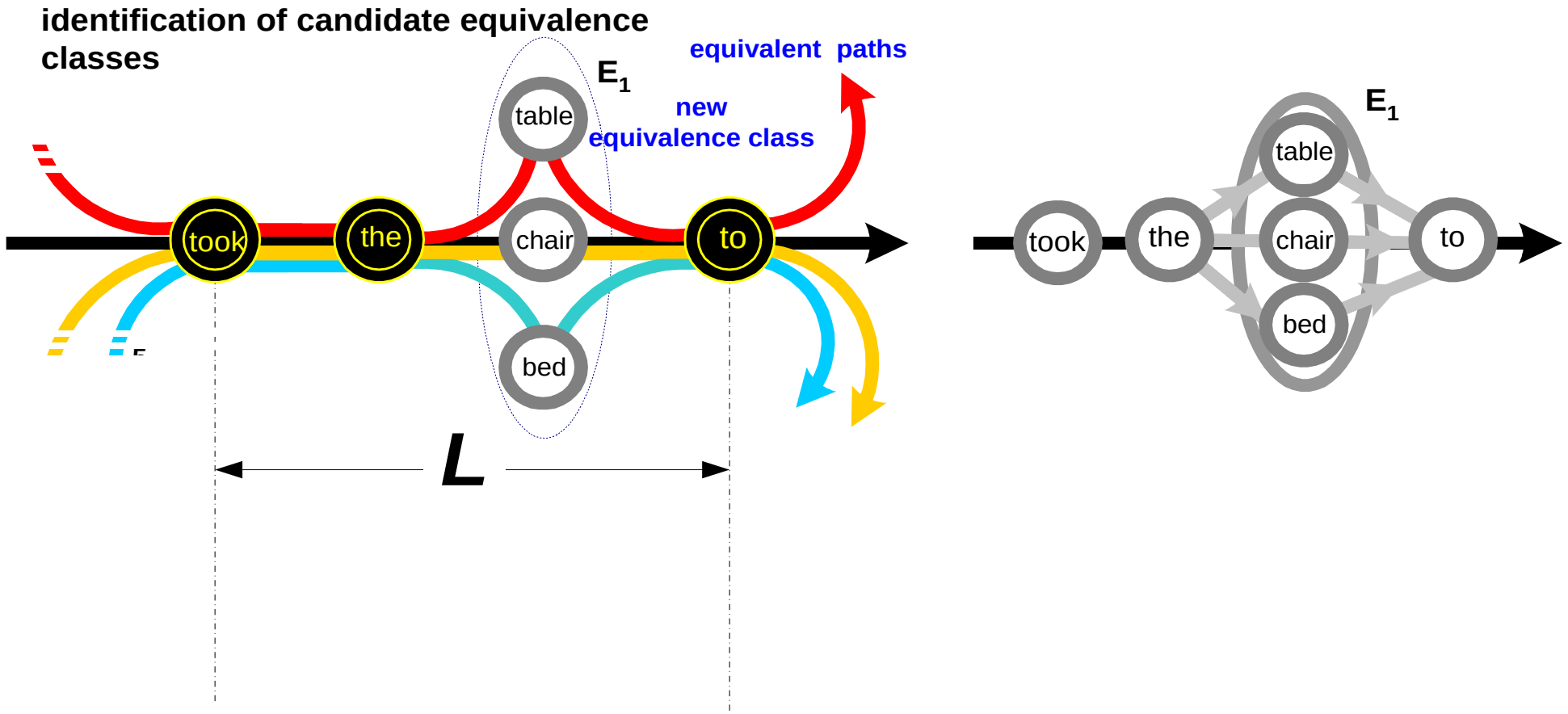
ADIOS



Once a pattern is identified as significant, the sub-paths it subsumes are merged into a new vertex and the graph is rewired accordingly. Repeating this process, leads to the formation of complex, hierarchically structured patterns.



ADIOS


identification of candidate equivalence classes
equivalent paths
E1
new equivalence class
table
took
the
chair
to
bed
L
E1
table
took
the
chair
to
bed

EXAMPLE

$$\left[\left[\left[\textit{the}\left[\begin{array}{c}\textit{ball}\\ \textit{door}\\ \textit{box}\\ \textit{square}\end{array}\right]\right]\left[\textit{circle}\right]\right] \rightarrow \left[\begin{array}{c}\textit{move}\\ \textit{came}\\ \textit{got}\end{array}\right] \rightarrow \textit{into}\right.$$

Ref: Nayak and Mukerjee, COLING 2012

FORMULATION

- Have 2 corpus focussed on a single topic Z
- Used PLSA and ADIOS to derive corresponding clusters in both the languages
- Can serve as a ID to the corresponding lines in the two languages or as an identifier to complex expressions

FORMULATION

1. Direct Mapping using bilingual dictionaries for same meaning tokens

2. Predicted Mapping using Clusters derived above

3. Semantic Mapping such as synonymy using Word-Net resource and POS tags in both the languages.

FORMULATION

- Problem reduces to a matrix solving problem
- Rows contain sentences in one language and column contains those in the other language
- Row is *n*-dimensional vector; *n* is the number of tokens in that sentence

PROBLEM

- Reorder rows and columns such that scores for each of the cell of the matrix maximizes
- A general representation for word-to-word alignment *L* for a given cross-lingual text with *N* words(a_1, a_2,…, a_N) from one language and *M* words (b_1, b_2,…..,b_M) from other language can be written as:

$$L = L_1, L_2, \ldots\ldots, L_p$$

$$L_p = [a_{x1}\ ,\ b_{x2}\ ]$$

x_1 *is in* $\{1,\ldots\ldots,N\}$

x_2 *is in* $\{1,\ldots\ldots,M\}$

ALGORITHM

1. Items which are separated by a white space or a punctuation will be called a word. They are the ones which need to be aligned.

2. Any extra tokens which have no corresponding tokens in the other language will be substituted by COPY token which means, no subtraction in score for that.

ALGORITHM

3. Sometimes two lines may mean the same thing or a part of the line may mean the same, so the two sub-parts will be said as aligned.

4. Alignment scores will also be given when two numbers are aligned in both languages, say for example, when dates are aligned

SCORES

- Direct mapping is certainly the best alignment and should be given the maximum score

- Predicted mapping take into account both semantics and syntax, its score will be less than direct mapping score

SCORES

- Semantic mapping score will also be less than direct mapping and will depend on the amount of semantic similarity present

- For semantic mapping, words can be extracted from English and Hindi WordNet

EVALUATION

- Precision
- Recall
- F-Score

- **Precision(also called positive predictive value) is the fraction of retrieved instances that are relevant.**
- **Recall(also known as sensitivity) is the fraction of relevant instances that are retrieved.**
- **F-Score is the *harmonic mean* of precision and recall.**

EVALUATION

- We have alignment *A* derived by our algorithm and *G* is the Gold Standard Alignment.

$$P_X = \frac{|A_X \cap G_X|}{A_X}$$

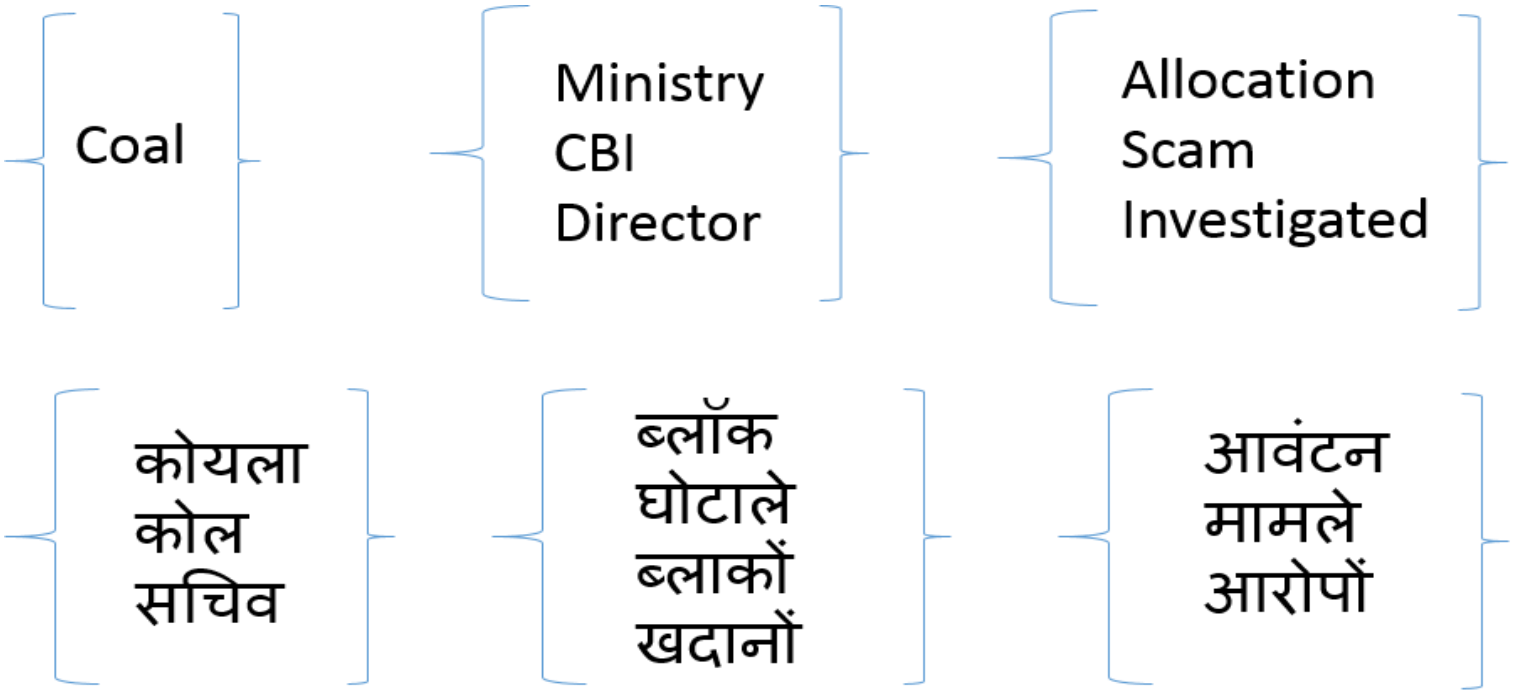
$$R_X = \frac{|A_X \cap G_X|}{G_X}$$

$$F_X = \frac{2P_X R_X}{P_X + R_X}$$

where X =C,P, C-Certain, P-Probable

BASE-RESULTS

- ADIOS



BASE-RESULTS

- **PLSA**

[कोयला, आवंटन, कोल, जिंदल, ब्लॉक, कंपनियों, कंपनी, हिंडाल्को, ब्लाक]

[coal, cbi, the, allocation, birla, fir, block, hindalco, alleged]

[कोयला, घोटाले, सीबीआई, दर्ज, सरकार, सीबीआइ, ब्लॉक, मामले, पूर्व]

[minister, coal, prime, bjp, the, scam, government, party, issue]

[जांच, कोर्ट, कोयला, सरकार, रिपोर्ट, सीबीआई, सुप्रीम, प्रधानमंत्री, मंत्री, घोटाले]

[cbi, coal, the, court, ministry, report, probe, files, agency, government]

FUTURE WORK

- Large Corpus for better vocabulary
- Include unsupervised results into an existing supervised system
- Phonemes for better alignment of Nouns
- N-gram clustering
- Improve predicted mapping score

REFERENCES

- Marianna Apidianaki. Unsupervised cross-lingual lexical substitution. In *Proceedings of the First Workshop on Unsupervised Learning in NLP*, EMNLP '11, pages 13{23, troudsburg, PA, USA, 2011. Association for Computational Linguistics.
- Thomas Hofmann. Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning*, 42(1):177{196, 2001.
- Heng Ji. Cross-lingual predicate cluster acquisition to improve bilingual event extraction by inductive learning. In *Proceedings of the Workshop on Unsupervised and Minimally Supervised Learning of Lexical Semantics*, UMSLLS '09, pages 27{35, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics.
- Zach Solan, David Horn, Eytan Ruppin, and Shimon Edelman. Unsupervised learning of natural languages. *Proceedings of the National Academy of Sciences of the United States of America*, 102(33):11629{11634, 2005.
- Jorg Tiedemann. Word to word alignment strategies. In *Proceedings of the 20th International Conference on Computational Linguistics*, COLING '04, Stroudsburg, PA, USA, 2004. Association for Computational Linguistics.
- Paola Virga and Sanjeev Khudanpur. Transliteration of proper names in cross-lingual information retrieval. In *Proceedings of the ACL 2003 Workshop on Multilingual and Mixed-language Named Entity Recognition - Volume 15*, MultiNER '03, pages 57{64, Stroudsburg, PA, USA, 2003. Association for Computational Linguistics.

THANK YOU!!