Project Presentation: CS498

AN EMPIRICAL STUDY OF CROSS-LINGUAL UNSUPERVISED ALIGNMENT BASED ON SYNTACTIC FEATURES

Advisor: Prof. Amitabha Mukerjee

Pranjal Singh
10511
spranjal@iitk.ac.in

MOTIVATION

- Ever increasing volume of text documents
- Quite important to identify similar contents in two languages
- Uses unsupervised learning which is less expensive as compared to fully supervised approach

OVERVIEW

- Bag of Words Model with terms as features

- Two Approaches:

1. **PLSA**

2. **ADIOS**

DATASET

- Manually built corpus on Coal Scam in Hindi and English
- English: ~52,000 tokens
- Hindi: ~36,000 tokens

SAMPLE TEXT(ENGLISH)

“The Supreme Court on Thursday expanded the scope of the coal scam investigations and issued notice to 7 states – Jharhand, Chhattisarh, Andhra Pradesh, Odisha, Madhya Pradesh and West Bengal.

While BJP also sought sacking of Law Minister Ashwani Kumar alleging his interference in preparation of CBI report to the Supreme Court on the coal scam, DMK sought resignation of P C Chacko as the Chairman of Joint”

SAMPLE TEXT(HINDI)

“कोयला घोटाले में शीर्ष उद्योगपति कुमार मंगलम बिड़ला के खिलाफ एफआइआर दर्ज होने के बाद मामले की आंच प्रधानमंत्री मनमोहन सिंह तक आने के बाद पूरी सरकार उनके बचाव में उतर आई है। वहीं, कोयला मंत्री श्रीप्रकाश जायसवाल ने कहा है कि पीएम को किसी से ईमानदारी का प्रमाणपत्र लेने की जरूरत नहीं है।

सीबीआइ ने कोयला ब्लॉक आवंटन के मामले में 14वीं एफआइआर में कुमार मंगलम बिड़ला के खिलाफ मामला दर्ज किया है। प्रधानमंत्री के समर्थन में पूरी तरह उतरते हुए शर्मा ने कहा कि अगर इसी तरह सरकार के हर फैसले पर सवाल उठते रहे तो मंत्रियों और नौकरशाहों के लिए कोई भी निर्णय करना मुश्किल हो जाएगा।सीबीआइ की एफआइआर के अनुसार”

METHODOLOGY

Feature Extraction(PLSA)

- Removed stop-words and symbols

- Represented document as term by document matrix (**Bag-of-Words Model**)

E.g. *document1* =$\{t_1, t_2, \ldots\ldots\ldots, t_n\}$

where t_i: term frequency in that document

- Stemming **NOT** used

SAMPLE

human	1	0	0	1	0	0	0	0	0
interface	1	0	1	0	0	0	0	0	0
computer	1	1	0	0	0	0	0	0	0
user	0	1	1	0	1	0	0	0	0
system	0	1	1	2	0	0	0	0	0
response	0	1	0	0	1	0	0	0	0
time	0	1	0	0	1	0	0	0	0
EPS	0	0	1	1	0	0	0	0	0
survey	0	1	0	0	0	0	0	0	1
trees	0	0	0	0	0	1	1	1	0
graph	0	0	0	0	0	0	1	1	1
minors	0	0	0	0	0	0	0	1	1

PROBLEMS

- Synonyms: separate words that have the same meaning.
  - E.g. ‘car’ & ‘automobile’
  - They tend to reduce recall
- Polysems: words with multiple meanings
  - E.g. ‘saturn’
  - They tend to reduce precision

→ The problem is more general: there is a disconnect between *topics* and *words*

PHILOSOPHY

‘… a more appropriate model should consider some *conceptual* dimensions instead of words.’

(Gardenfors)

PLSA (PROBABILISTIC LATENT SEMANTIC ANALYSIS)

- To identify latent features
- Find out topics from a collection of text documents
- Creates clusters using probabilistic approach

PLSA

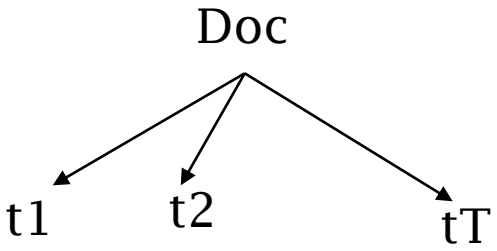
- Latent variables z are introduced and relate to documents d
- $|z| << |d|$, as the same z_i may be associated with more than one documents

- z performs as a bottleneck and results in dimensionality reduction

PLSA

$$P(doc) = P(term_1 \mid doc)P(term_2 \mid doc)\ldots P(term_L \mid doc)$$

$$= \prod_{l=1}^{L} P(term_l \mid doc) = \prod_{t=1}^{T} P(term_t \mid doc)^{X(term_t, doc)}$$

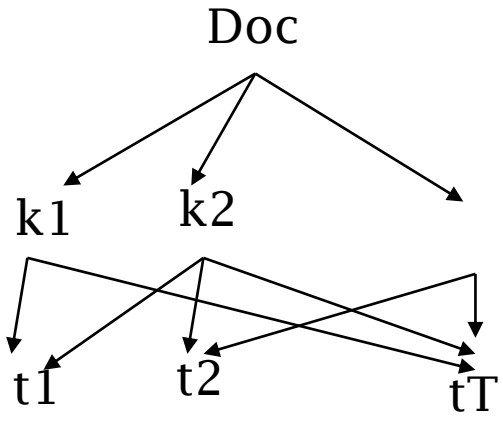


We know how to compute the parameter of this model, i.e., $P(term_t \mid doc)$

PLSA

Now let us have K topics as well:

$$P(term_t \mid doc) = \sum_{k=1}^{K} P(term_t \mid topic_k) P(topic_k \mid doc)$$



The same, written using shorthands:

$$P(t \mid doc) = \sum_{k=1}^{K} P(t \mid k) P(k \mid doc)$$

So by replacing this, for any doc in the collection,

$$P(doc) = \prod_{t=1}^{T} \{ \sum_{k=1}^{K} P(t \mid k) P(k \mid doc) \}^{X(t,doc)}$$

PLSA

- The parameters of this model are:
  P(t|k)
  P(k|doc)
- It is possible to derive the equations for computing these parameters by Maximum Likelihood.
- If we do so, what do we get?
  P(t|k) for all t and k, is a term by topic matrix
  (gives which terms make up a topic)
  P(k|doc) for all k and doc, is a topic by document matrix
  (gives which topics are in a document)

PLSA

- The log likelihood of this model is the log probability of the entire collection:

$$\sum_{d=1}^{N} \log P(d) = \sum_{d=1}^{N} \sum_{t=1}^{T} X(t,d) \log \sum_{k=1}^{K} P(t \mid k) P(k \mid d)$$

which is to be maximised w.r.t. parameters $P(t \mid k)$ and then also $P(k \mid d)$,

subject to the constraints that $\sum_{t=1}^{T} P(t \mid k) = 1$ and $\sum_{k=1}^{K} P(k \mid d) = 1.$

RESULTS(PLSA)

- [कोयला, आवंटन, कोल, जिंदल, ब्लॉक, कंपनियों, कंपनी, हिंडाल्को, ब्लाक]
- [coal, cbi, the, allocation, birla, fir, block, hindalco, alleged]

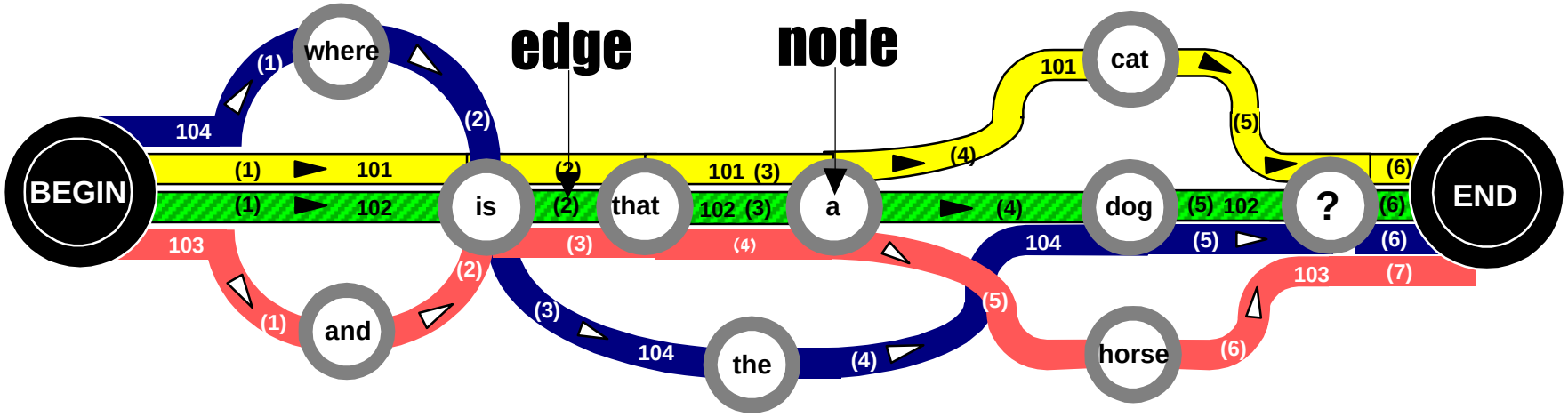
- [कोयला, घोटाले, सीबीआई, दर्ज, सरकार, सीबीआइ, ब्लॉक, मामले, पूर्व]
- [minister, coal, prime, bjp, the, scam, government, party, issue]

- [जांच, कोर्ट, कोयला, सरकार, रिपोर्ट, सीबीआई, सुप्रीम, प्रधानमंत्री, मंत्री, घोटाले]
- [cbi, coal, the, court, ministry, report, probe, files, agency, government]

ADIOS (AUTOMATIC DISTILLATION OF STRUCTURE)

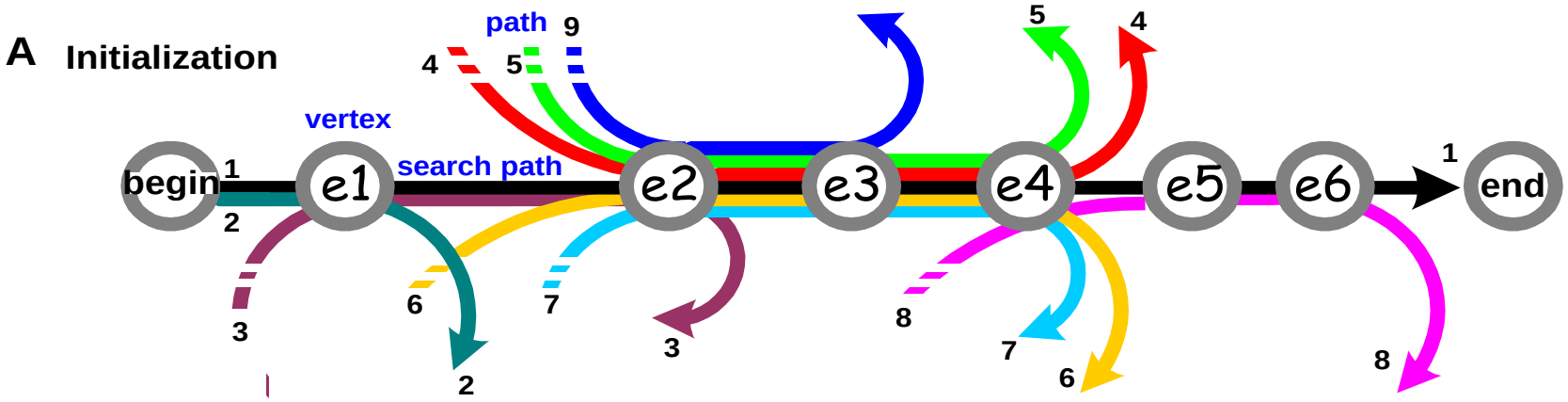
- ADIOS capable of learning complex syntax, generating grammatical novel sentences
- Proving useful in other fields that call for structure discovery from raw data, such as bioinformatics
- Composed of three main elements
  - A representational data structure
  - A segmentation criterion (MEX)
  - A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

ADIOS: THE MODEL



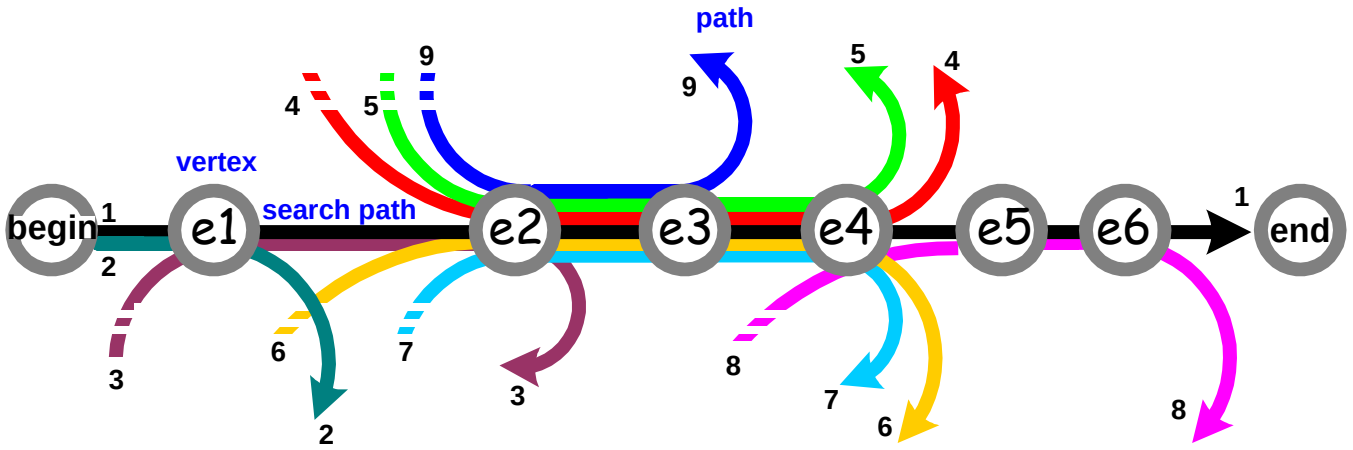
Is that a dog?

ADIOS

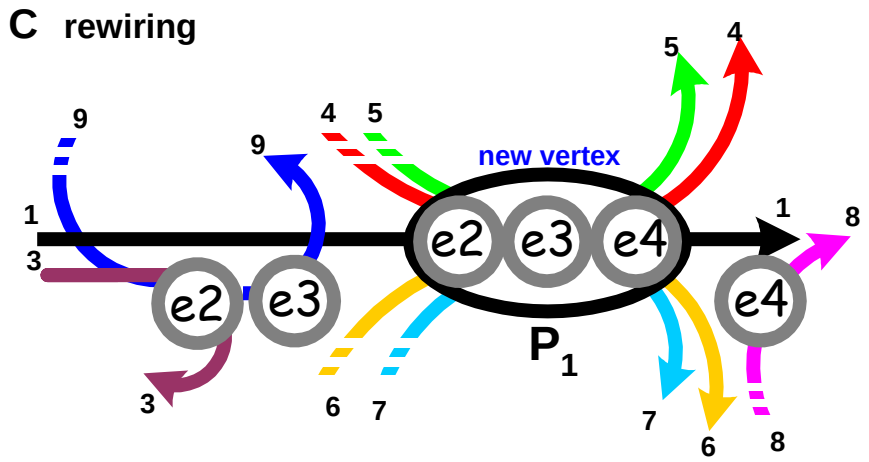


- Identifying patterns becomes easier on a graph
  - Sub-paths are automatically aligned

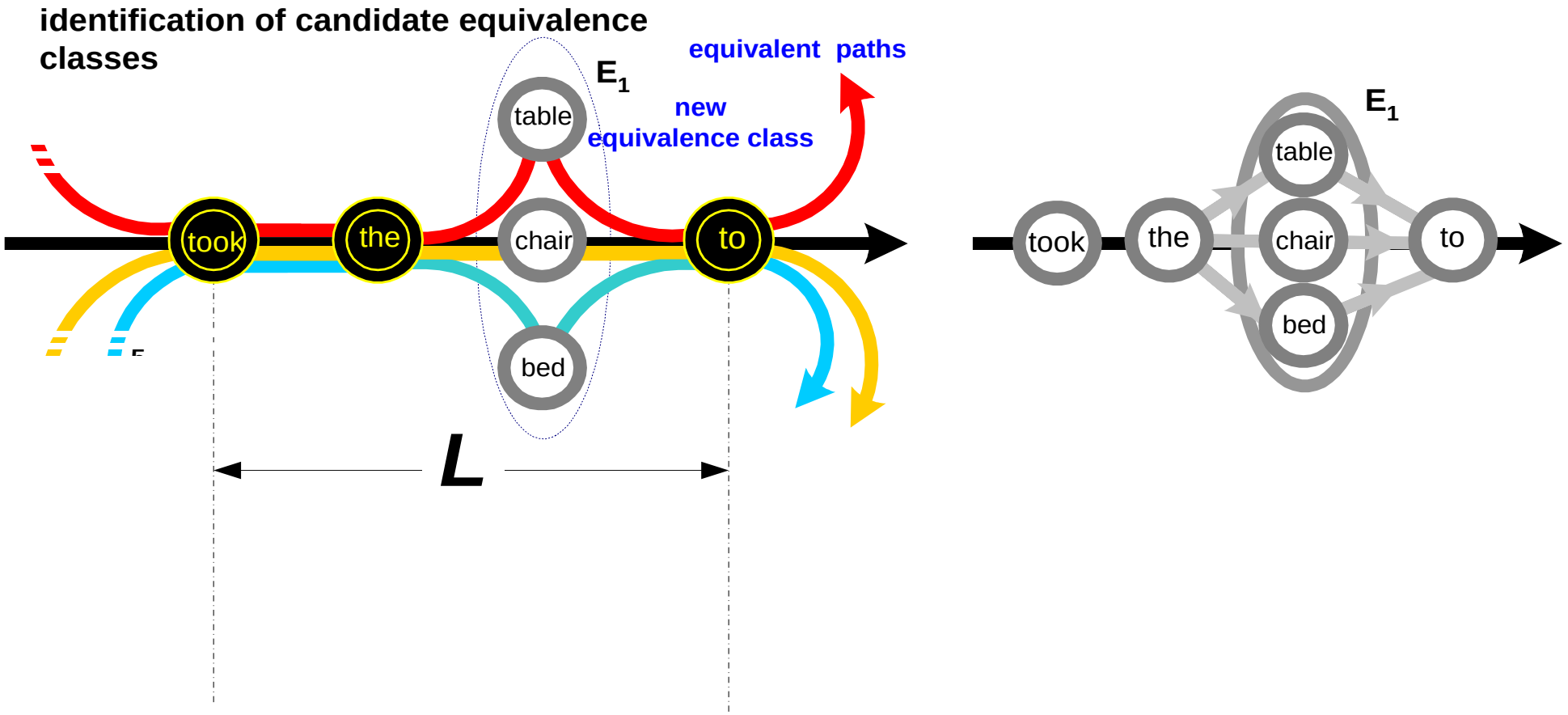
ADIOS



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

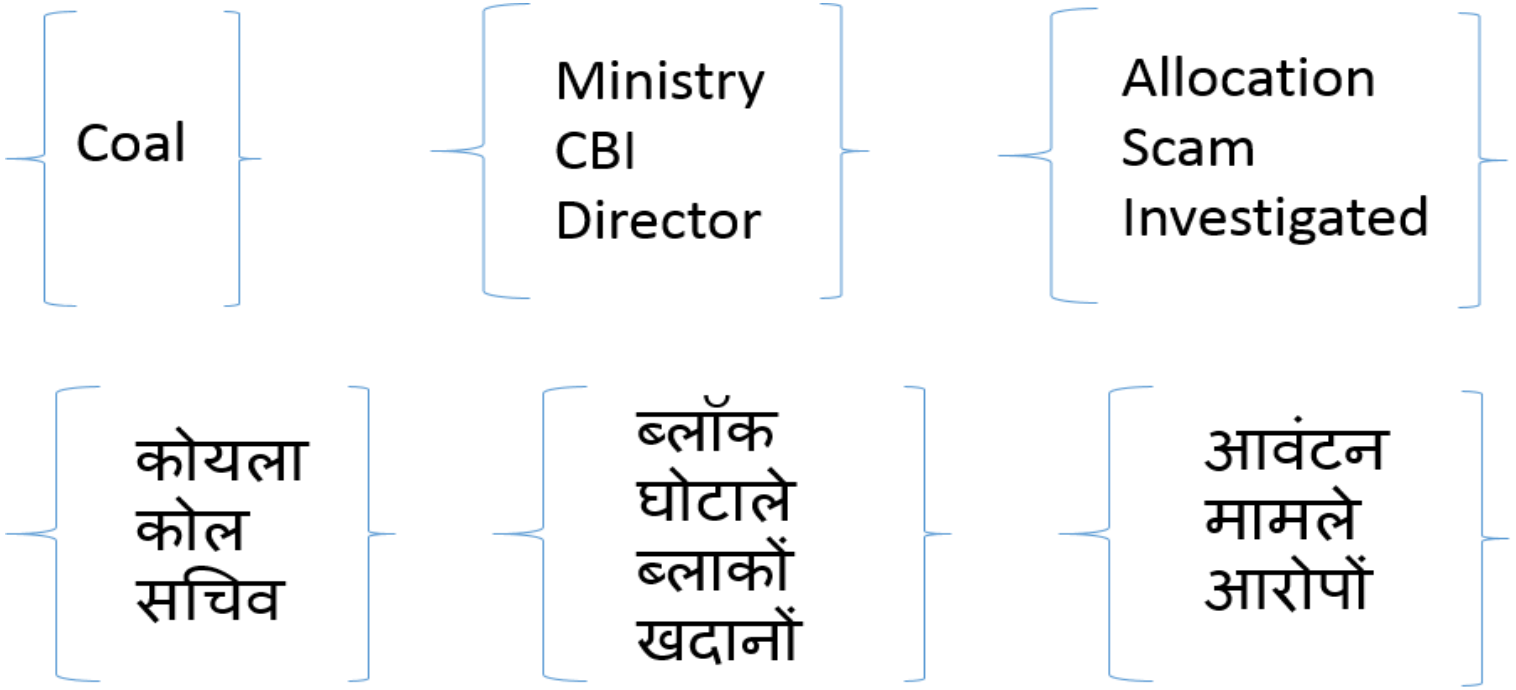

identification of candidate equivalence classes
equivalent paths
E1
new equivalence class
table
took
the
chair
to
bed
L
E1
table
took
the
chair
to
bed

ADIOS ALGORITHM

- Initialization – load data into pseudograph

- Until no more patterns are found do
  - For each path detect all sub-paths that live up to the MEX criterion
  - Pick best pattern, add it to graph and rewire paths

RESULTS(ADIOS)



FUTURE WORK

- Currently looked only at a focused corpus
- To look at documents with multiple topics and then strengthen the alignment in the concerned languages with smaller alignments.
- Look for n-gram clusters in the documents
- Can be used to identify semantic relations between sentences in different languages

FUTURE WORK

- Distributional Features are motivated by the so-called Distributional Hypothesis:

"The degree of semantic similarity between two linguistic expressions A and B is a function of the similarity of the linguistic contexts in which A and B can appear" **

- Harris proposed the method of distributional analysis as a scientific methodology for linguistics:
- introduced for phonology, then methodology for all linguistic levels.

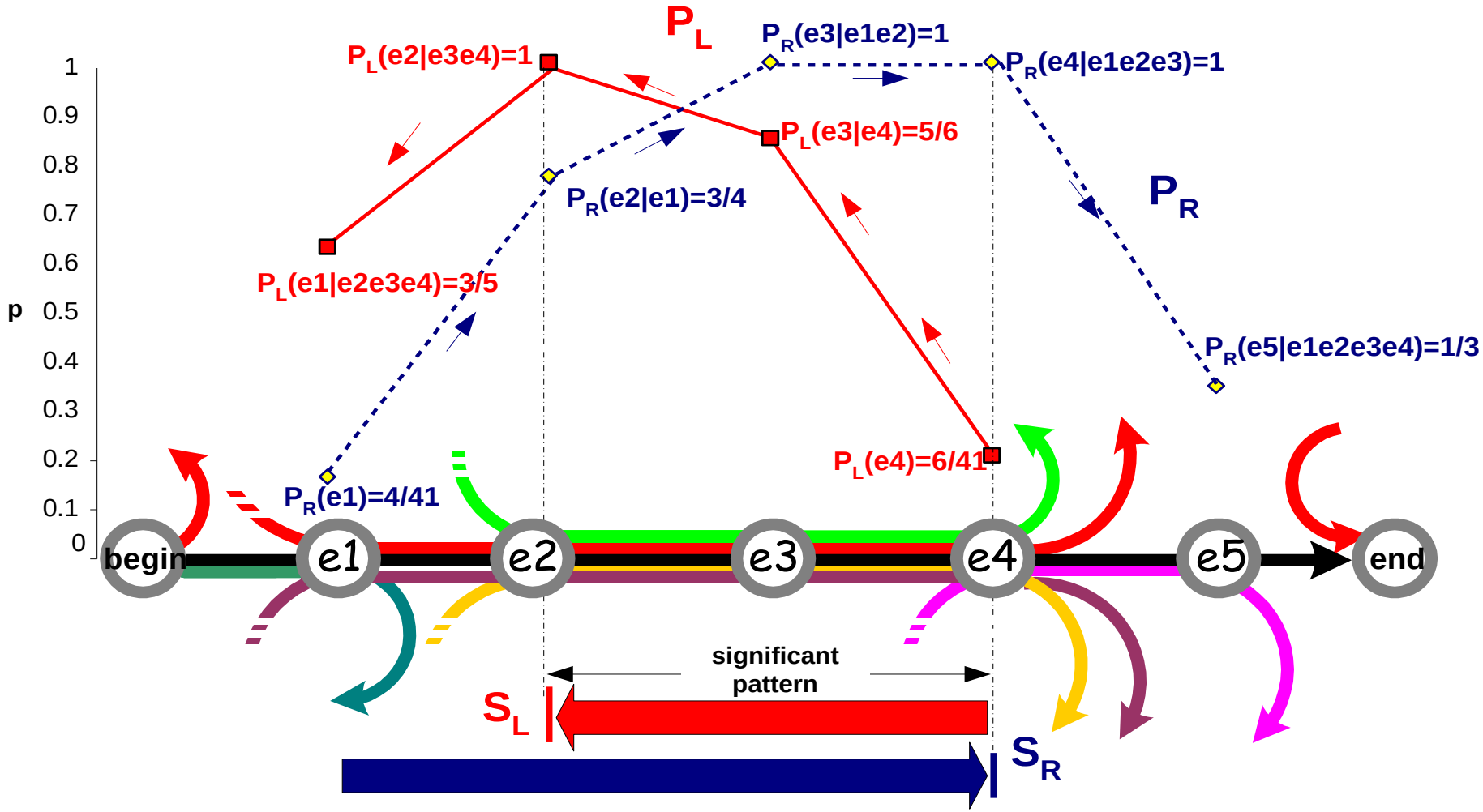
**Z. Harris (1954) Distributional Structure

REFERENCES

- 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.
- Oscar Täckström, Ryan McDonald, and Jakob Uszkoreit. Cross-lingual word clusters for directtransfer of linguistic structure. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL HLT ’12, pages 477–487, Stroudsburg, PA, USA, 2012. Association for Computational Linguistics.
- Zach Solan Thesis, Tel Aviv University, 2006

THANK YOU!!!

MOTIF EXTRACTION



MARKOV MATRIX

$$M(e_1 e_2 \cdots e_k) \doteq \begin{pmatrix} p(e_1) & p(e_1|e_2) & p(e_1|e_2e_3) & \dots & p(e_1|e_2e_3\dots e_k) \\ p(e_2|e_1) & p(e_2) & p(e_2|e_3) & \dots & p(e_2|e_3e_4\dots e_k) \\ p(e_3|e_1e_2) & p(e_3|e_2) & p(e_3) & \dots & p(e_3|e_4e_5\dots e_k) \\ \vdots & \vdots & \vdots & & \vdots \\ p(e_k|e_1e_2\dots e_{k-1}) & p(e_k|e_2e_3\dots e_{k-1}) & p(e_k|e_3e_4\dots e_{k-1}) & \dots & p(e_k) \end{pmatrix}$$

- The top right triangle defines the P_L probabilities, bottom left triangle the P_R probabilities
- Matrix is path-dependent

PATTERN SELECTION?

- Obviously, the more significant the pattern the better
- Turns out it helps choosing longer patterns first when segmenting text
  - Lowers the probability for accidentally linking words
- Also turns out it helps to gradually increase ALPHA

CONTEXT SENSITIVE GENERALIZATION

- Slide a context window of size L across current search path
- For each $1 \leq i \leq L$
  - look at all paths that are identical with the search path for $1 \leq k \leq L$, except for $k=i$
  - Define an equivalence class containing the nodes at index i for these paths
  - Replace i^{th} node with equivalence class
  - Find significant patterns using MEX criterion

ADIOS DRAWBACK

- ADIOS is inherently a heuristic and greedy algorithm
  - Once a pattern is created it remains forever – errors conflate
  - Sentence ordering affects outcome

- Running ADIOS with different orderings gives patterns that ‘cover’ different parts of the grammar

PLSA ALGORITHM

- Inputs: term by document matrix X(t,d), t=1:T, d=1:N and the number K of topics sought
- Initialise arrays P1 and P2 randomly with numbers between [0,1] and normalise them to sum to 1 along rows
- Iterate until convergence

  For d=1 to N, For t =1 to T, For k=1:K

$$P1(t,k) \leftarrow P1(t,k)\sum_{d=1}^{N}\frac{X(t,d)}{\sum_{k=1}^{K}P1(t,k)P2(k,d)}P2(k,d); \quad P1(t,k) \leftarrow \frac{P1(t,k)}{\sum_{t=1}^{T}P1(t,k)}$$

$$P2(k,d) \leftarrow P2(k,d)\sum_{t=1}^{T}\frac{x(t,d)}{\sum_{k=1}^{K}P1(t,k)P2(k,d)}P1(t,k); \quad P2(k,d) \leftarrow \frac{P2(k,d)}{\sum_{k=1}^{K}P2(k,d)}$$

- Output: arrays P1 and P2, which hold the estimated parameters P(t|k) and P(k|d) respectively

DISTRIBUTIONAL SEMANTICS

- Harris goes one step farther and claims that *distributions* should be taken as an ***explanans*** *for meaning* itself

- A distributional semantic model (DSM) is a co-occurrence matrix **M** where rows correspond to *target terms* and columns correspond to *context* or *situations* where the target terms appear.