Bellcore

Bell Communications Research

Automatic Cross-Language Information Retrieval Using Latent Semantic Indexing (LSI)

Susan T. Dumais
Thomas K. Landauer
Michael L. Littman
Aug. 22, 1996

Sigir96 Multilingual IR Workshop

Overview of Latent Semantic Indexing (LSI)

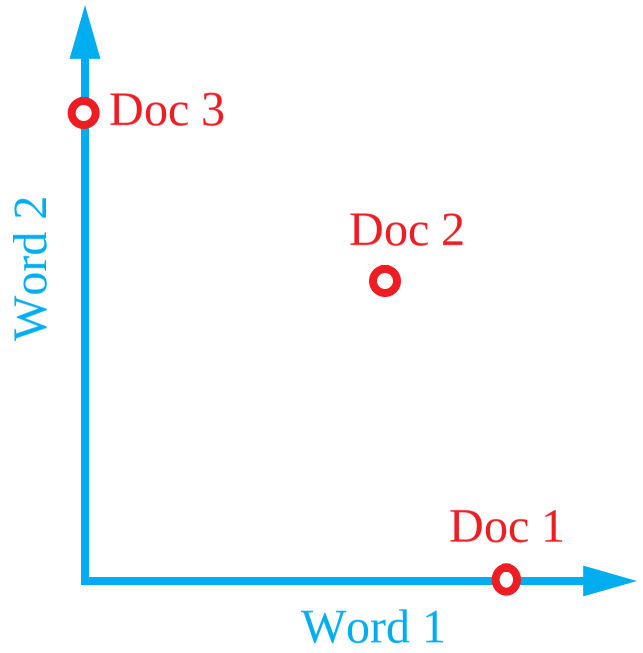
- LSI: modified vector retrieval method; explicitly model inter-item associations

- Begin w/ standard term-document matrix
- “Latent” structure in matrix obscured by noise or variability in word usage
- Use “reduced SVD” to model latent semantic structure
- Use reduced structure for retrieval (~300d)
- Geometric representation

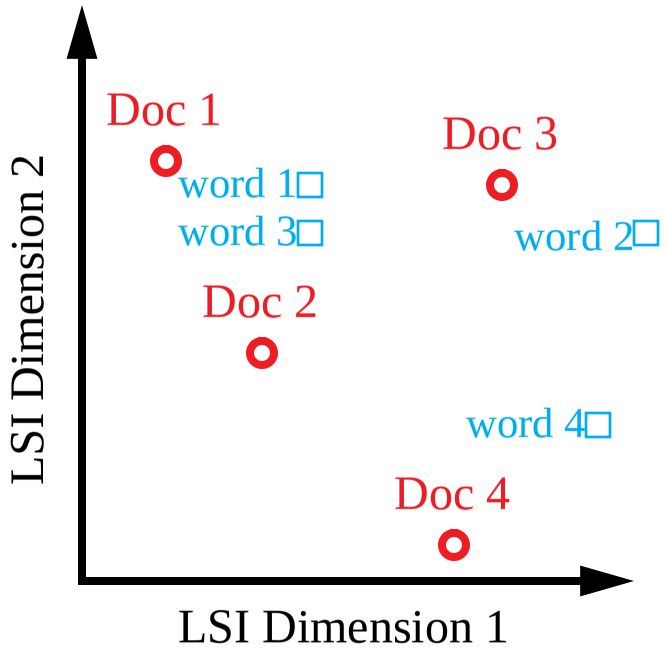
.

Standard Vector vs. Reduced LSI Vector

Standard Vector Space Model
(ndims = nterms)



Reduced LSI Vector Space Model
(ndims << nterms)



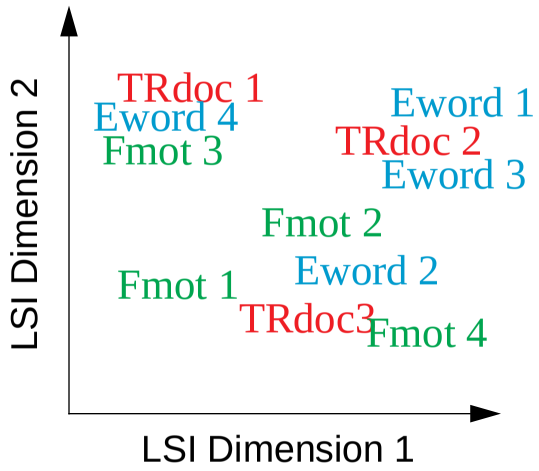
Using LSI for Cross-Language Retrieval

1. *Train* using combined multilingual documents -> derive inter-relationships among terms
2. *Foldin* monolingual documents
3. *Queries* in either language retrieve most similar documents regardless of language

Using LSI for Cross-Language Retrieval
1. Train Combined

- Train: using “combined” documents

Hon. Benoit Bouchard (Secretary of State of Canada): Mr. Speaker, I would like to bring to the attention of the House that today, as Hon. Members are no doubt aware, we are celebrating the anniversary of the proclamation of the Canadian Charter of Rights and Freedoms which took place on April 17, 1982, and also of the coming into effect a year ago of the provisions guaranteeing equality for all members of our society. --- L’hon. Benoit Bouchard (secretaire d’Etat du Canada): Monsieur le President, je voudrais porter a l’attention de la Chambre que nous celebrons aujourd’hui, comme le savent les honorables deputes, l’anniversaire de la proclamation de la Charte canadienne des droits et libertes qui a eu lieu le 17 avril 1982, ainsi que son parachevement, il y a un an, avec l’entree en vigueur des dispositions garantissant l’egalite a tous les membres de notre societe.

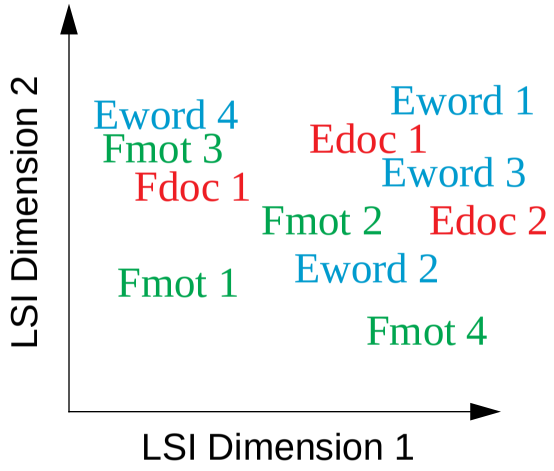


.

Using LSI for Cross-Language Retrieval
2. Foldin Monolingual

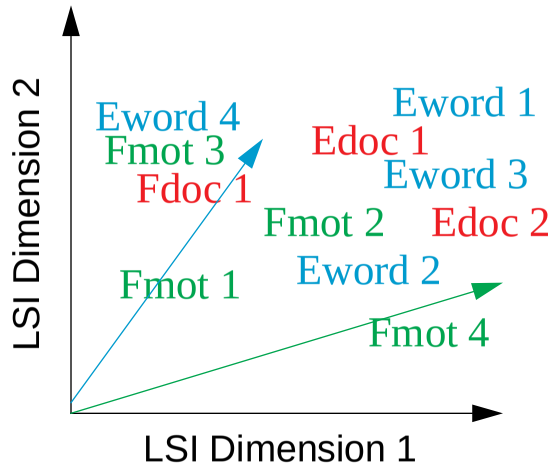
- Foldin: monolingual documents (at vector sum of constituent terms)

Hon. Erik Nielsen (Deputy Prime Minister and Minister of National Defense): Mr. Speaker, we are in constant touch with our consular officials in Libya. We are advised the situation there is stabilizing now. There is no immediate threat to Canadians. Therefore my responses yesterday, which no doubt the Hon. Member has seen, have not altered.



Using LSI for Cross-Language Retrieval
3. Queries

- Queries (at vector sum of constituent words)
- Query in either language retrieves most similar documents regardless of language



Experiments Using Cross-Language LSI

- Landauer & Littman (1990) - French and English
- Landauer, Littman & Stornetta (1992) - Japanese and English
- Young (1994) - Greek and English

- Dumais, Landauer & Littman (1996) - extensions including MT-LSI, no-LSI

Cross-Language LSI (CL-LSI)
Experiment

- Hansard collection
- Train: 982 combined EF documents
  - LSI analysis: 2 minutes on Sparc
- Foldin: 1500 E documents
  1500 F documents
- Several Tests
  - mate retrieval
  - short pseudo queries
  - short human queries

Cross-Language LSI (CL-LSI)
Mate Retrieval

- How well does a test document retrieve its cross-language mate?

Query:
Hon. Erik Nielsen (Deputy Prime Minister and Minister of National Defense): Mr. Speaker, we are in constant touch with our consular officials in Libya. We are advised the situation there is stabilizing now. There is no immediate threat to Canadians. Therefore my responses yesterday, which no doubt the Hon. Member has seen, have not altered.

Cross-Language Mate:
L'hon. Erik Nielsen (vice-premier ministre er ministre de la Defense Nationale): Monsieur le President, nous sommes en communication constante avec nos representants consulaire en Libye. D'apres nos informations, la situation est en train de se stabiliser, et les Canadiens ne sont pas immediatement menaces. Par consequent, mes reponses d'hier, dont le representant a du prendre connaissance, n'ont pas change.

- CL-LSI: E->F: 98.3% (1475/1500)
  F->E: 98.5% (1478/1500)

Cross-Language LSI (CL-LSI)
Mate Retrieval

- How much is due to word overlap alone?
- no-LSI: E->F: 47.7%
  F->E: 49.5%

- What if no words overlap?
- no-LSI: E->F: 0%
  F->E: 0%
- CL-LSI: E->F: 98.7%
  F->E: 99.0%

.

Machine Translation LSI (MT-LSI)
Mate Retrieval

- Train: 982 E-only documents
- Foldin: 1500 E documents
  1500 Et (of F) documents
  translated using SYSTRAN

French Mate:

L'hon. Erik Nielsen (vice-premier ministre er ministre de la Defense Nationale): Monsieur le President, nous sommes en communication constante avec nos representants consulaire en Libye. D'apres nos informations, la situation est en train de se stabiliser, et les Canadiens ne sont pas immediatement menaces. Par consequent, mes reponses d'hier, dont le representant a du prendre connaissance, n'ont pas change.

Machine Tranlation of French Mate (Et):
The hon. Erik Nielsen (Deputy Prime Minister and Minister for Defense nationale): Mr. President, we are in constant communication with our representatives consular in Libya. According to our information, the situation is stabilizing itself, and the Canadians are not immediately threatened. Consequently, my answers of yesterday, whose representative had to take note, did not change.

.

Machine Translation LSI (MT-LSI)
Mate Retrieval

- MT-LSI: E->Et: 99.3%
  Et->E: 99.7%
- MT-LSI: F->Ft: 99.1%
  Ft->F: 98.7%

- MT-LSI performance very good (99.2%) ... CL-LSI (98.4%) comparable and it's fully automatic after training

.

Cross-Language LSI (CL-LSI)
Mate Retrieval - Short Pseudo Queries

- Previous expts used very long queries
- Short Pseudo Queries

  e.g. "Nielsen, consular, immediate, threat, inundated"

Hon. Erik *Nielsen* (Deputy Prime Minister and Minister of National Defense): Mr. Speaker, we are in constant touch with our *consular* officials in Libya. We are advised the situation there is stabilizing now. There is no *immediate threat* to Canadians. Therefore my responses yesterday, which no doubt the Hon. Member has seen, have not altered.

- Test: pseudo query E
- CL-LSI: pqE->E: 90.5% pqE->F: 55.4%
  MT-LSI: pqE->E: 91.7% pqE->Et: 62.9%
  no-LSI: pqE->E: 93.3% pqE->F: 18.6%

.

Cross-Language LSI (CL-LSI)
Human Short Queries

- 8 subjects generate English queries
- Return best matching documents according to several methods
- Collect relevance judgements
- Results ... in progress

Conclusions

- Cross-Language LSI (CL-LSI) - practical and effective
  - train with combined multilingual docs
  - foldin monolingual docs
  - query in any language retrieves most similar docs in any language
- CL-LSI better than no-LSI (vector matching)
- CL-LSI comparable to MT-LSI, at lower cost