Monsoon 2020

16 - Distributional Semantics

Information Retrieval

by

Dr. Rajendra Prasath

Indian Institute of Information Technology, Sri City, Chittoor

Sri City – 517 646, Andhra Pradesh, India

Topics Covered So Far

- Term Weighting
- Vector Space Models
- Evaluation Metrics
- Relevance Feedback / Pseudo Relevance Feedback
- Query Expansion appoaches
  - Dictionary Based Approach
  - Co-occurrence Based approach
  - Tag-Cloud Based Approach
  - Pseudo Relevance Based Approaches

Now:

Distributional Semantics

Recap: Overview

- Why Ranked Retrieval?
- Term Frequency
- Term Weighting
- TF-IDF Weighting
- The Vector Space Model
- Relevance Feedback
- Pseudo Relevance Feedback

Query Expansion

Query expansion

- Another Way to increase recall
- Global query expansion

  ⇒ global methods for query reformulation
- In global query expansion, the query is modified based on some global resource, i.e. a resource that is not query-dependent.
- Main information we use: (near-)synonymy
- A publication or database that collects (near-)synonyms is called a thesaurus.
- We will look at two types of thesauri: manually created and automatically created

Query expansion at search engines

- Main source of query expansion at search engines: query logs
- **Example 1:**
  After issuing the query [herbs], users frequently search for [herbal remedies].
  - "herbal remedies" is potential expansion of "herb"
- **Example 2:**
  a) Users searching for [flower pix] frequently click on: photobucket.com/flower
  b) Users searching for [flower clipart] frequently click on the same URL.
  - "flower clipart" and "flower pix" are potential expansions of each other.

Word Space Models

- How do we identify semantically related information that improves documents retrieval better

- User query terms are expanded based on terms with similar word senses that are discovered by the "**associatedness**" of the document context with that of the given query

- Can we capture the term contexts using higher order term associations and assists the effective retrieval of news documents

Motivations

- **Associatedness** - guided by word space models (Kanerva et al 2000)
- The word-space model computes the meaning of terms by implicitly utilizing the contexts of words collected over large text data
- The distributional patterns represent semantic similarity between words in terms of their spatial proximity in the context space
  - Words □ context vectors whose relative directions are assumed to indicate semantic similarity
- **Distributional hypothesis:**
  - words with similar meanings are assumed to have similar contexts
  - word space methodology makes semantics computable
  - Underlying models do not require linguistic or semantic expertise

Interesting Contributions

- Similarity assessment is conjectured to involve higher-order relationships, particularly in the models of analogical reasoning (Gentner and Forbus, 1991)
- Discovered higher-order distributional relations for textual CBR (Chakraborti et al., 2007; Deerwester et al., 1990)
- RI is an alternative to **Latent Semantic Indexing (LSI)** that reduces dimensionality (Kanerva et al., 2000; Sahlgren, 2005)
- Semantic behavior, word order information can be learned in an unsupervised way, using **Holographic Reduced Representations(HRR)** (Plate, 1995; Jones and Mewhort, 2007)
- We follow the evaluation measures proposed for the standard IR systems (Singhal et al., 1996; Raghunathan et al., 2008)

Interesting Contributions (contd)

- Bruninghaus et al. [5] NLP & IE methods to automatically extract relevant factual information without converting into structured documents.
- Harman [12] - Abductive inference reasons upon incomplete or inconsistent information
- Baddeley [3] - find context relationship between documents
- Kanerva [18, 17] - scalable distributional model – meaningful implicit relationships between terms in queries and documents
- Harris [13] - semantically similar terms occur in similar contexts □ semantically similar docs (no topic info)
- Johnson - Lindenstrauss Lemma [16] - a way of encoding textual information in the form of random projections.

LSI vs Random Indexing

- Latent Semantic Indexing (LSI) :
  - Sparse term - document matrix and Singular Value
  - Decomposition (SVD)
  - Not incremental
  - Dimensionality reduction

- Random Indexing (RI): uses distributional statistics for identifying the semantic similarity
  - Random index and context vectors - one for each term /
  - sentence / para - of fixed size
  - Incremental
  - Identifies implicit semantic relations

Distributional Hypothesis

Johnson-Lindenstrauss Lemma [1984]:

- A set of points in a high dimensional vector space can be mapped down into a reduced dimensional space such that the distance between any two points changes but not significantly

- This inherently leads to:
  - Dimensionality Reduction
  - Random Projections

Word Relations

- Associative relations - immediate relations to adjacent words:

  eat → food

- Synonymy relations - second order relations to words that share contexts:

  eat

  ↓

  drink

- word space methodology makes semantics computable and constitutes a purely descriptive approach to semantic modeling

Random Indexing

- **Index Vectors:**
  - each context (e.g. each document or each word) is assigned a unique and randomly generated representation
  - index vectors are sparse, high-dimensional, and ternary
  - their dimensionality (d) is in the order of thousands, and that they consist of a small number of randomly distributed +1s and -1s, with the rest of the elements of the vectors set to 0.
- **Context Vectors:**
  - context vectors are produced by scanning through the text
  - each time a word occurs in a context (within a sliding context window), that context's d-dimensional index vector is added to the context vector for the word in question.
  - Words are thus represented by d-dimensional context vectors that are effectively the sum of the words' contexts

Random Indexing (contd)

Context Vectors:

- For each occurrence of a given feature in all cases, we focus on a fixed window of size (2 × k) + 1 centered at the given feature [suggested window size is 5 (= term + / - k terms)]
- Then feature context vector for featurei is computed using the following equation:

$$C_{feature_i} = C_{feature_i} + \sum_{j=-k;j\neq 0}^{+k} I_{feature_{(i+j)}} \times \frac{1}{d^{|j|}}$$

- where 1/d|j| is the weight proportion w.r.to size j of window (d=2)

- Superposition is used while updating the context vector
- Adding two vectors x and y yields a vector z where z = x + y & the cosine similarities between x & z, and y & z will be high

RI – An Example

New Case: "The fisherman caught a big salmon today"
window size k = 2 , and we are training the feature **big**

The windowed sentence for the feature big looks like this:
The, [fisherman, caught, big, salmon, today].
The feature-context-vector Cbig for big becomes now:

$$C_{big} = C_{big} + (0.25 \times I_{fisherman}) + (0.5 \times I_{caught}) + (0.5 \times I_{salmon}) + (0.25 \times I_{today})$$

Meaning of a case is captured in the collective representation of the constituent features

Case Context Vectors:

- Case context vector = a weighted superposition of context vectors of features that occur in the case, as follows:

$$C_{case} = \sum_{i=1} f_i \times C_{feature_i}$$

where f_i is the number of occurrences of $feature_i$ in case.

Example: Consider 2 Sentences

Two sample sentences:

the weather is **fine** in Hong Kong

the weather is **nice** in Hong Kong

Generate random keys for each word within some context

Window:

weather	{ 0	-1	+1	0 }
is	{ 0	0	+1	-1 }
in	{ +1	0	0	-1 }
hong	{ +1	-1	0	0 }

Collect sums for words of interest:

	d1	d2	d3	d4
weather	{ 0	-1	+1	0 }
is	{ 0	0	+1	-1 }
in	{ +1	0	0	1 }
hong	{ +1	1	0	0 }
				+
fine	{ +2	2	+2	2 }
nice	{ +2	2	+2	2 }

Advantages of RI

- **RI is an incremental method**
  - Similarity computation even with a few examples

- Dimensionality d of the vectors is a parameter

- Random Indexing uses “implicit” dimension reduction
  - [Constant (much lower) dimensionality]

- Random Indexing can be used with any type of context
  - Other word space models typically use either documents or words as contexts

Summary

In this class, we focused on:

(a) Query Expansion

i. Thesaurus based Approach
ii. Co-occurrence Based Approach

(b) Distributional Semantics

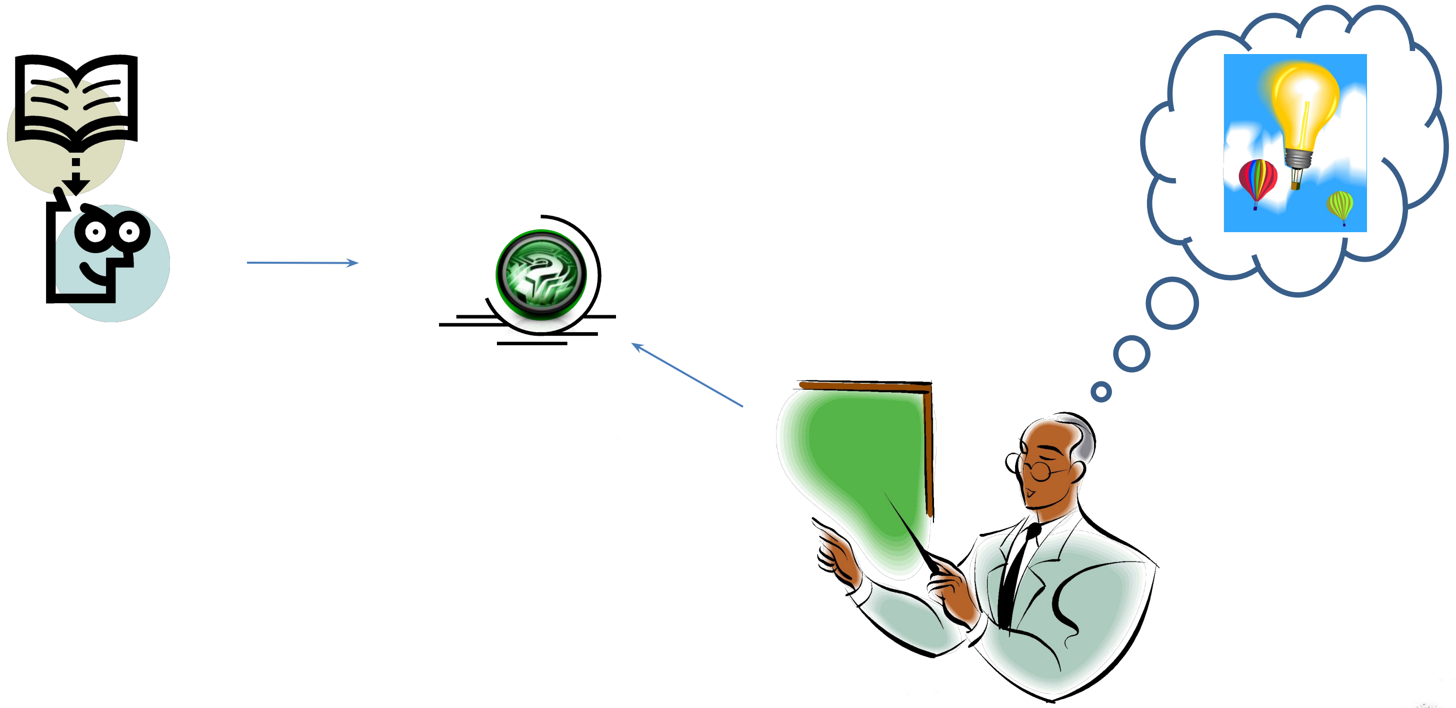
i. Word Space Models
ii. Distributional Hypothesis
iii. Random Indexing
iv. Advantages of RI in IR

Acknowledgements

Thanks to all IR Researchers:

1. Introduction to Information Retrieval Manning, Raghavan and Schutze, Cambridge University Press, 2008.
2. Search Engines Information Retrieval in Practice W. Bruce Croft, D. Metzler, T. Strohman, Pearson, 2009.
3. Information Retrieval Implementing and Evaluating Search Engines Stefan Büttcher, Charles L. A. Clarke and Gordon V. Cormack, MIT Press, 2010.
4. Modern Information Retrieval Baeza-Yates and Ribeiro-Neto, Addison Wesley, 1999.
5. Many Authors who contributed to SIGIR / WWW / KDD / ECIR / CIKM / WSDM and other top tier conferences
6. Prof. Mandar Mitra, Indian Statistical Institute, Kolkatata (https://www.isical.ac.in/~mandar/) for sharing the IR Evaluation Slides

Thanks …



… Questions ???