Vector Space Models of Semantics

Siva Reddy
Supervisor: Dr. Suresh Manandhar

Artificial Intelligence Group
Department of Computer Science
University of York

Literature Review Seminar

Outline

Semantics

- Study of meaning
- Humans use language to transfer the meaning
  - Figure out what people mean
  - Herculean task for computers
- Distributional Hypothesis (Harris, 1954)
  - Words that occur in similar contexts tend to have similar meanings
  - e.g. Tree and Plant, Tea and Coffee, Bus and Vehicle
  - *Bag of words hypothesis*: Two documents tend to be similar if they have same distribution of similar words
- You shall know a word by the company it keeps (Firth, 1957)

THE UNIVERSITY of York

Semantics

- Study of meaning
- Humans use language to transfer the meaning
  - Figure out what people mean
  - Herculean task for computers
- Distributional Hypothesis (Harris, 1954)
  - Words that occur in similar contexts tend to have similar meanings
  - e.g. Tree and Plant, Tea and Coffee, Bus and Vehicle
  - *Bag of words hypothesis*: Two documents tend to be similar if they have same distribution of similar words
- You shall know a word by the company it keeps (Firth, 1957)

Vector Space Models (VSMs) of Semantics

- **Interpret semantics using VSM**
  - Backbone: Distributional Hypothesis
- Text entity (we are interested in) as a Vector (point) in dimensional space.
- Context of the entity as dimensions
- VSM are well equipped mathematically
  - Linear Algebra
- Advanced computational techniques
- Widely used in Machine Learning
  - Image, Text, Speech Processing

VSMs of Semantics

- Existing methods represent knowledge in VSMs mainly in three types (Turney and Pantel, 2010)
  - term-document
  - term-context
  - pair-pattern
- Example application: *Information retrieval*
  - Just scratching the surface of human language[1]
  - Immense impact on society and the economy already

[1] Courtesy: (Turney and Pantel, 2010)

Term-Document: (Salton et al., 1975)

Create a word-by-document matrix

	d1	d2	d3	d4	d5	d6	d7	d8	d9
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	0	0	0	0
graph	0	0	0	0	0	0	1	1	1
minors	0	0	0	0	0	0	0	1	1

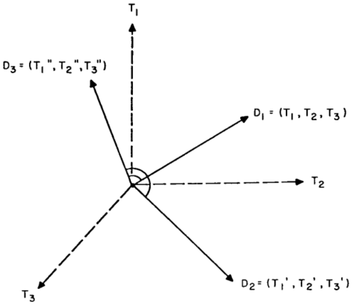
[2]

d1: **Human** machine **interface** for Lab ABC **computer** applications

[2] Image courtesy: (Landauer et al., 1998)

THE UNIVERSITY of York

Term-Document: (Salton et al., 1975)



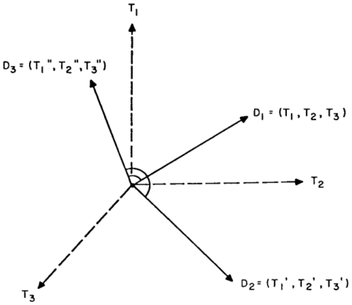
[3]

Document similarity can be found using Cosine similarity

- $sim(D1, D2) = \frac{D1.D2}{||D1||||D2||}$
- A survey on Similarity Measures (Weeds et al., 2004)

[3] Image courtesy: (Salton et al., 1975)

Term-Document: (Salton et al., 1975)



[3]

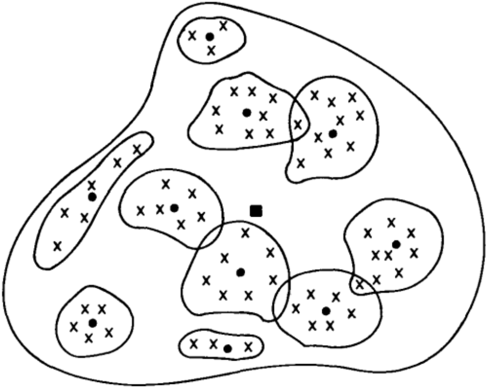
Document similarity can be found using Cosine similarity

- $sim(D1, D2) = \frac{D1.D2}{\|D1\|\|D2\|}$
- A survey on Similarity Measures (Weeds et al., 2004)

[3] Image courtesy: (Salton et al., 1975)

Term-Document: Applications

- Clustering similar documents



[4]

- SMART information retrieval system
  - Query is represented as a pseudo document

THE UNIVERSITY of York

[4] Image courtesy: (Salton et al., 1975)

Term-Document: Latent Semantic Indexing

- Above matrix representation is sparse and noisy.
- Similar terms are not treated as a single dimension.
- Deerwester et al. (1990) applied Singular Value decomposition (SVD) to the above matrix - Latent Semantic Indexing (LSI)
  - Reduced number of dimensions
  - SVD creates low-dimensional mapping to the given dimensional space
  - Capture concepts instead of words
  - The new document vector discovers latent (hidden) meaning
  - Much higher document similarity precision
- Efficient way of modelling concepts
  - Probabilistic LSI
  - Latent Dirichlet Allocation

THE UNIVERSITY of York

Term-Document: Latent Semantic Indexing

- Above matrix representation is sparse and noisy.
- Similar terms are not treated as a single dimension.
- Deerwester et al. (1990) applied Singular Value decomposition (SVD) to the above matrix - Latent Semantic Indexing (LSI)
  - Reduced number of dimensions
  - SVD creates low-dimensional mapping to the given dimensional space
  - Capture concepts instead of words
  - The new document vector discovers latent (hidden) meaning
  - Much higher document similarity precision
- Efficient way of modelling concepts
  - Probabilistic LSI
  - Latent Dirichlet Allocation

THE UNIVERSITY of York

Term-Document: Latent Semantic Indexing

- Above matrix representation is sparse and noisy.
- Similar terms are not treated as a single dimension.
- Deerwester et al. (1990) applied Singular Value decomposition (SVD) to the above matrix - Latent Semantic Indexing (LSI)
  - Reduced number of dimensions
  - SVD creates low-dimensional mapping to the given dimensional space
  - Capture concepts instead of words
  - The new document vector discovers latent (hidden) meaning
  - Much higher document similarity precision
- Efficient way of modelling concepts
  - Probabilistic LSI
  - Latent Dirichlet Allocation

Term-Context

- Similar to Term-Document but with a focus on "Term"
- Landauer and Dumais (1997) applied LSI to find word similarity
  - Also called Latent Semantic Analysis (Landauer et al., 1998)
  - Extends to Term-Context representation
- Applications:
  - Word Sense Disambiguation (Schütze, 1998)
  - TOEFL Synonym Test (Landauer and Dumais, 1997)
  - Flat and hierarchical word clustering
  - Word Classification
  - Thesaurus building

Term-Context

- Similar to Term-Document but with a focus on “Term”
- Landauer and Dumais (1997) applied LSI to find word similarity
    - Also called Latent Semantic Analysis (Landauer et al., 1998)
    - Extends to Term-Context representation
- Applications:
    - Word Sense Disambiguation (Schütze, 1998)
    - TOEFL Synonym Test (Landauer and Dumais, 1997)
    - Flat and hierarchical word clustering
    - Word Classification
    - Thesaurus building

Pair-Pattern

Extended distributional hypothesis (Lin and Pantel, 2001)

- Patterns that co-occur with similar pairs tend to have similar meanings
- “X cuts Y” and “X works with Y” are similar patterns
  - *mason:stone’*
  - *carpenter:wood*
- Pair-Pattern Matrix

Latent relation hypothesis (Turney, 2008)

- Inverse of Extended distributional hypothesis
- *mason:stone, carpenter:wood, potter:clay* are relationally similar
  - “the X used the Y”
  - “the X shaped the Y into”

Applications: Inference engines, Q/A systems.

Pair-Pattern

Extended distributional hypothesis (Lin and Pantel, 2001)

- Patterns that co-occur with similar pairs tend to have similar meanings
- “X cuts Y” and “X works with Y” are similar patterns
  - *mason:stone’*
  - *carpenter:wood*
- Pair-Pattern Matrix

Latent relation hypothesis (Turney, 2008)

- Inverse of Extended distributional hypothesis
- *mason:stone, carpenter:wood, potter:clay* are relationally similar
  - “the X used the Y”
  - “the X shaped the Y into”

Applications: Inference engines, Q/A systems.

Pair-Pattern

Extended distributional hypothesis (Lin and Pantel, 2001)

- Patterns that co-occur with similar pairs tend to have similar meanings
- “X cuts Y” and “X works with Y” are similar patterns
  - *mason:stone’*
  - *carpenter:wood*
- Pair-Pattern Matrix

Latent relation hypothesis (Turney, 2008)

- Inverse of Extended distributional hypothesis
- *mason:stone, carpenter:wood, potter:clay* are relationally similar
  - “the X used the Y”
  - “the X shaped the Y into”

Applications: Inference engines, Q/A systems.

Vector Compositions

- How do you compose a vector for a given “query”?
  - Search query: *University of York*
  - How near is the *composed vector* to the *true vector*?
  - What is the dimensional space of the composed vector?
  - Opened a new direction of research
- Vector Compositions
  - Build vectors for larger entities from smaller units
  - Vector of “University of York” from vectors of “University” and “York”
  - Can extend to sentence or document or any higher level

Vector Compositions

- How do you compose a vector for a given “query”?
  - Search query: *University of York*
  - How near is the *composed vector* to the *true vector*?
  - What is the dimensional space of the composed vector?
  - Opened a new direction of research
- Vector Compositions
  - Build vectors for larger entities from smaller units
  - Vector of “University of York” from vectors of “University” and “York”
  - Can extend to sentence or document or any higher level

Compositionality functions

- Compositionality function $V \oplus W$
- Existing compositionality functions (Mitchell and Lapata, 2008; Widdows, 2008)
  - Addition
    - V + W
    - $dim(V \oplus W) = dim(V) = dim(W)$
    - Widely used and works in most information retrieval systems
  - Multiplication
    - Multiply values belonging to the same dimension.
    - $dim(V \oplus W) = dim(V) = dim(W)$
    - Paraphrase detection and synonym test

Compositionality functions

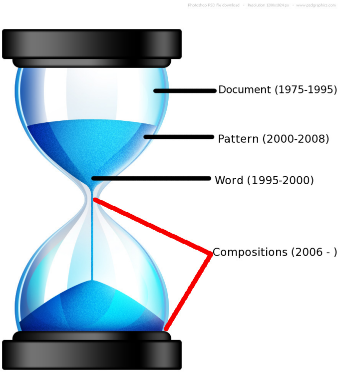
- Complex compositionality functions
  - Can capture hidden relations between vectors
  - $Moscow : X :: London - Britain$
  - Transform into a new dimensional space
  - Direct product: $dim(V \oplus W) = dim(V) + dim(W)$
  - Tensor product: $dim(V \oplus W) = dim(V) \times dim(W)$
- Machine Learning for linear models Z= AV + BW
  - Z is the true vector computed from corpus
  - A and B are the parameters (matrices)
  - Guevara (2010)
  - Zanzotto et al. (2010) (York)

THE UNIVERSITY of York

Compositionality Applications

- Widely used in detecting compositionality of Multi-word
  - Baldwin et al. (2003)
  - Katz and Giesbrecht (2006)
- Paraphrase and synonym detection
  - Erk and Padó (2009)
- Query Expansion
  - Cao et al. (2008)

Trend Shifts



Drawbacks

- A single vector for each entity built from all its instances
  - the entity may be polysemous
  - a need for context aware vectors
- Simple models cannot differentiate word order
  - *house rent* and *rent house*
  - Complex models are highly expensive
- Methods like dimensionality reduction are computationally expensive
  - A problem in scaling
  - Inexpensive models exist and approximate the true values
  - Random Indexing (Sahlgren, 2005)
- Gives similarity value between any two entities - strength and weakness

Future: Context Sensitive Vectors

- Exemplar Model (Erk and Pado, 2010)
  - Store each instance of the entity as an exemplar
- Activate relevant exemplars based on the context
  - e.g. Traffic Light
  - $act(Traffic, light) \oplus act(Light, traffic)$
- Our initial experiments are fruitful

THE UNIVERSITY of York

Summary

- Modelling semantics using vector space models
- Representing entities and constructing the vector space
- Some techniques like LSI
- Applications
- Building semantics from smaller semantic units
- Context Sensitive vectors (Dynamic Vectors)

Bibliography I

Baldwin, T., Bannard, C., Tanaka, T., and Widdows, D. (2003). An empirical model of multiword expression decomposability. In *Proceedings of the ACL 2003 workshop on Multiword expressions: analysis, acquisition and treatment - Volume 18*, pages 89–96, Morristown, NJ, USA. Association for Computational Linguistics.

Cao, H., Jiang, D., Pei, J., He, Q., Liao, Z., Chen, E., and Li, H. (2008). Context-aware query suggestion by mining click-through and session data. In *KDD '08: Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 875–883, New York, NY, USA. ACM.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., and Harshman, R. (1990). Indexing by latent semantic analysis. *JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE*, 41(6):391–407.

Bibliography II

Erk, K. and Padó, S. (2009). Paraphrase assessment in structured vector space: exploring parameters and datasets. In *Proceedings of the Workshop on Geometrical Models of Natural Language Semantics*, GEMS '09, pages 57–65, Morristown, NJ, USA. Association for Computational Linguistics.

Erk, K. and Pado, S. (2010). Exemplar-based models for word meaning in context. In *Proceedings of the ACL 2010 Conference Short Papers*, pages 92–97, Uppsala, Sweden. Association for Computational Linguistics.

Firth, J. R. (1957). A synopsis of linguistic theory 1930-55. 1952-59:1–32.

Bibliography III

Guevara, E. (2010). A Regression Model of Adjective-Noun Compositionality in Distributional Semantics. In *Proceedings of the 2010 Workshop on GEometrical Models of Natural Language Semantics*, pages 33–37, Uppsala, Sweden. Association for Computational Linguistics.

Harris, Z. (1954). Distributional structure. *Word*, 10(23):146–162.

Katz, G. and Giesbrecht, E. (2006). Automatic identification of non-compositional multi-word expressions using latent semantic analysis. In *Proceedings of the Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties*, MWE '06, pages 12–19, Morristown, NJ, USA. Association for Computational Linguistics.

Bibliography IV

Landauer, T. K. and Dumais, S. T. (1997). Solution to Plato's Problem: The Latent Semantic Analysis Theory of Acquisition, Induction and Representation of Knowledge. *Psychological Review*, (104).

Landauer, T. K., Foltz, P. W., and Laham, D. (1998). An introduction to latent semantic analysis. *Discourse Processes*, 25:259–284.

Lin, D. and Pantel, P. (2001). Dirt - discovery of inference rules from text. In *In Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 323–328.

Mitchell, J. and Lapata, M. (2008). Vector-based Models of Semantic Composition. In *Proceedings of ACL-08: HLT*, pages 236–244, Columbus, Ohio. Association for Computational Linguistics.

Bibliography V

Sahlgren, M. (2005). An Introduction to Random Indexing. *Methods and Applications of Semantic Indexing Workshop at the 7th International Conference on Terminology and Knowledge Engineering, TKE 2005*.

Salton, G., Wong, A., and Yang, C. S. (1975). A vector space model for automatic indexing. *Commun. ACM*, 18:613–620.

Schütze, H. (1998). Automatic word sense discrimination. *Comput. Linguist.*, 24:97–123.

Turney, P. D. (2008). The latent relation mapping engine: algorithm and experiments. *J. Artif. Int. Res.*, 33:615–655.

Turney, P. D. and Pantel, P. (2010). From frequency to meaning: vector space models of semantics. *J. Artif. Int. Res.*, 37:141–188.

Bibliography VI

Weeds, J., Weir, D., and Mccarthy, D. (2004). Characterising measures of lexical distributional similarity. In *In Proceedings of CoLing 2004*, pages 1015–1021.

Widdows, D. (2008). Semantic vector products: Some initial investigations. In *Proceedings of the Second AAAI Symposium on Quantum Interaction*. AAAI.

Zanzotto, F. M., Korkontzelos, I., Fallucchi, F., and Manandhar, S. (2010). Estimating linear models for compositional distributional semantics. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING)*, pages 1263–1271, Beijing, China. Coling 2010 Organizing Committee.