Vector Spaces for Cross-Language NLP Applications

Rafael E. Banchs

Human Language Technology Department,

Institute for Infocomm Research, Singapore

INTERSPEECH 2015

September 6 - 10

Dresden, Germany

Tutorial Outline

Section 1: Basic Concepts and Theoretical Framework

Section 2: Vector Spaces in Monolingual NLP

Section 3: Vector Spaces in Cross-language NLP

Section 4: Future Research and Applications

Motivation

- The mathematical metaphor offered by the geometric concept of distance in **vector spaces** with respect to **semantics** and **meaning** has been proven to be useful in monolingual NLP applications.
- There is some recent evidence that this paradigm can also be useful for **cross-language** NLP applications.

Objectives

The main objectives of this tutorial are as follows:

- To introduce the basic concepts related to distributional and cognitive semantics
- To review some classical examples on the use of vector space models in monolingual NLP applications
- To present some novel examples on the use of vector space models in the cross-language NLP applications

Section 1

Basic Concepts and Theoretical Framework

- **The Distributional Hypothesis**
- Vector Space Models and the Term-Document Matrix
- Association Scores and Similarity Metrics
- The Curse of Dimensionality and Dimensionality Reduction
- Semantic Cognition, Conceptualization and Abstraction

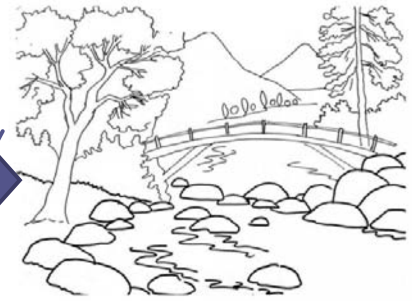
Distributional Hypothesis

"a word is characterized for the company it keeps" *

(meaning is mainly determined by the context rather than from individual language units)

- Please cash the cheque at the **bank**

- Please check for rocks along the **bank**



***** *Firth, J.R. (1957) A synopsis of linguistic theory 1930-1955, in Studies in linguistic analysis, 51: 1-31*

Distributional Structure

Meaning as a result of language's Distributional Structure … or vice versa ?

"… if we consider words or morphemes A and B to be more different in meaning than A and C, then we will often find that the distributions of A and B are more different than the distributions of A and C." *

"In the language itself, there are only differences" **

***** *Harris, Z. (1970) Distributional Structure, in Papers in structural and transformational linguistics*

** *Saussure, F. (1916) Course in General Linguistics*

Not everyone is happy... ☹

Argument against...

- Meaning involves more than language:
  - Images and experiences that are beyond language
  - Objects, ideas and concepts in the minds of the speaker and the listener

Counterargument...

- "if extralingusitc factors *do* influence linguistic events, there will always be a distributional correlate to the event that will suffice as explanatory principle" *

* *Sahlgren, M. (2006) The distributional hypothesis*

Not everyone is happy… ☹

Argument against…

- The concept of semantic difference (or similarity) is too broad to be useful !!!

Counterargument …

➢ Semantic relations "are not axiomatic, and the broad notion of semantic similarity seems perfectly plausible" *

* *Sahlgren, M. (2006) The distributional hypothesis*

Functional Differences

- Functional differences across words are fundamental for defining the notion of meaning
- Two different types of functional differences between words can be distinguished: *
  - Syntagmatic relations:
    Explain how words are combined (co-occurrences)
  - Paradigmatic relations:
    Explain how words exclude each other (substitutions)

* *Saussure, F. (1916) Course in General Linguistics*

Orthogonal Dimensions

Paradigmatic

some	scientists	look	smart
few	people	feel	dumb
most	citizens	seem	gifted
many	lawyers	are	savvy

Syntagmatic

The Term-context Matrix

dogs are animals

cats are animals

orchids are plants

roses are plants

	Animals	Are	Cats	Dogs	Orchids	Plants	Roses
Animals		X	X	X			
Are	X		X	X	X	X	X
Cats	X	X					
Dogs	X	X					
Orchids		X				X	
Plants		X			X		X
Roses		X				X	

Paradigmatic Relation Matrix

Top Paradigmatic Pairs

(dogs, cats)

(orchids, roses)

	Animals	Are	Cats	Dogs	Orchids	Plants	Roses
Animals		X	X	X			
Are	X		X	X	X	X	X
Cats	X	X					
Dogs	X	X					
Orchids		X				X	
Plants		X			X		X
Roses		X				X	

The Term-document Matrix

D1: dogs are animals

D2: cats are animals

D3: orchids are plants

D4: roses are plants

	D1	D2	D3	D4
Animals	X	X		
Are	X	X	X	X
Cats		X		
Dogs	X			
Orchids			X	
Plants			X	X
Roses				X

Syntagmatic Relation Matrix

Top Syntagmatic Pairs

(animals, cats)

(animals, dogs)

(orchids, plants)

(plants, roses)

	D1	D2	D3	D4
Animals	X	X		
Are	X	X	X	X
Cats		X		
Dogs	X			
Orchids			X	
Plants			X	X
Roses				X

Section 1

Basic Concepts and Theoretical Framework

- The Distributional Hypothesis
- **Vector Space Models and the Term-Document Matrix**
- Association Scores and Similarity Metrics
- The Curse of Dimensionality and Dimensionality Reduction
- Semantic Cognition, Conceptualization and Abstraction

Vector Space Models (VSMs)

- Vector Space Models have been extensively used in Artificial Intelligence and Machine Learning applications
- Vector Space Models for language applications were introduced by Gerard Salton* within the context of Information Retrieval
- Vector Spaces allow for simultaneously modeling words and the contexts in which they occur

***** *Salton G. (1971) The SMART retrieval system: Experiments in automatic document processing*

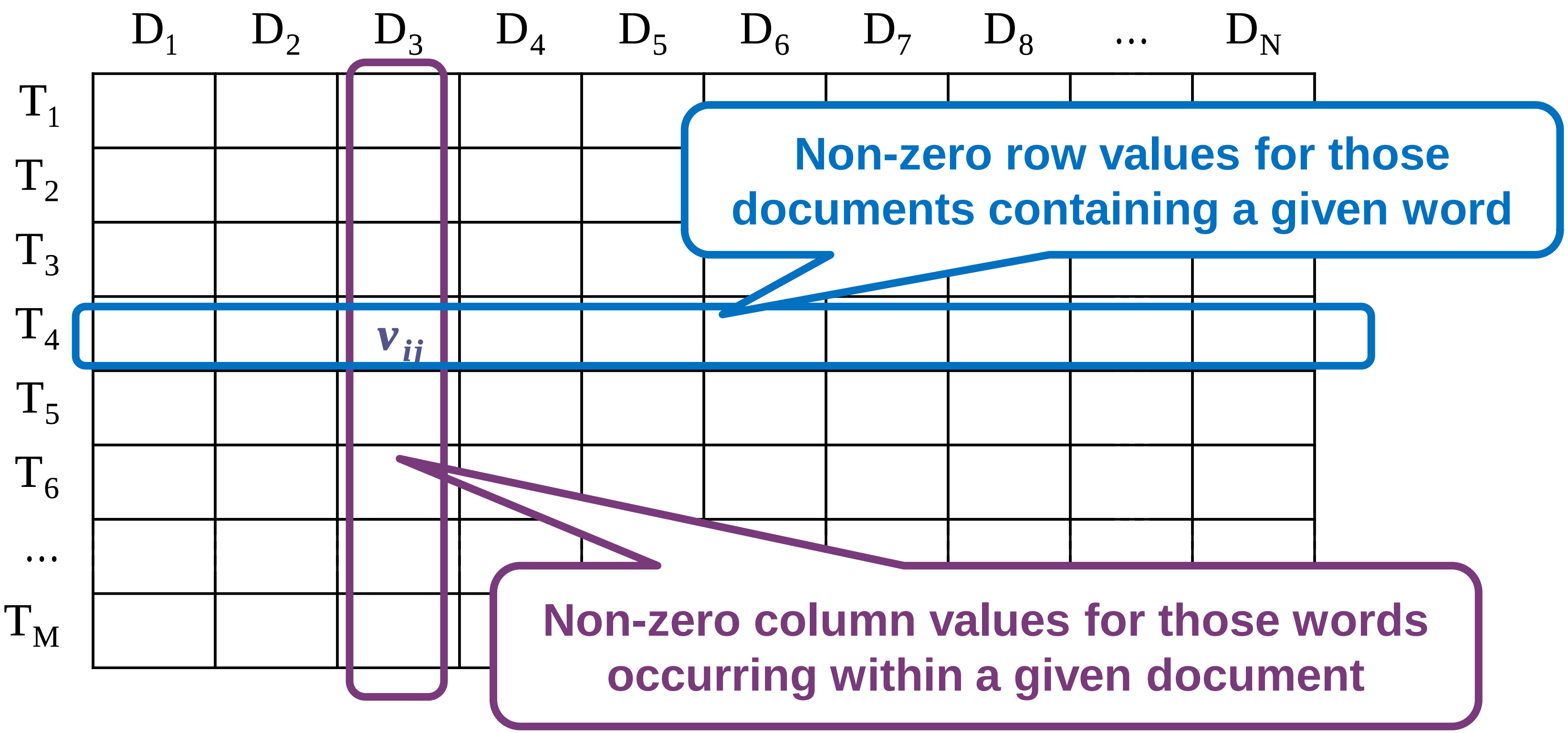
Three Main VSM Constructs*

- The term-document matrix
  - Similarity of documents
  - Similarity of words (Syntagmatic Relations)
- The word-context matrix
  - Similarity of words (Paradigmatic Relations)
- The pair-pattern matrix
  - Similarity of relations

* *Turney P.D., Pantel P. (2010) From frequency to meaning: vector space models of semantics, Journal of Artificial Intelligence Research, 37: 141-188*

The Term-Document Matrix

- A model representing joint distributions between words and documents



The Term-Document Matrix

- Each row of the matrix represents a unique vocabulary word in the data collection
- Each column of the matrix represents a unique document in the data collection
- Represents joint distributions between words and documents
- It is a bag-of-words kind of representation
- A real-valued weighting strategy is typically used to improve discriminative capabilities

A bag-of-words Type of Model

Document collection

Document *x*

response said covering candidate picture

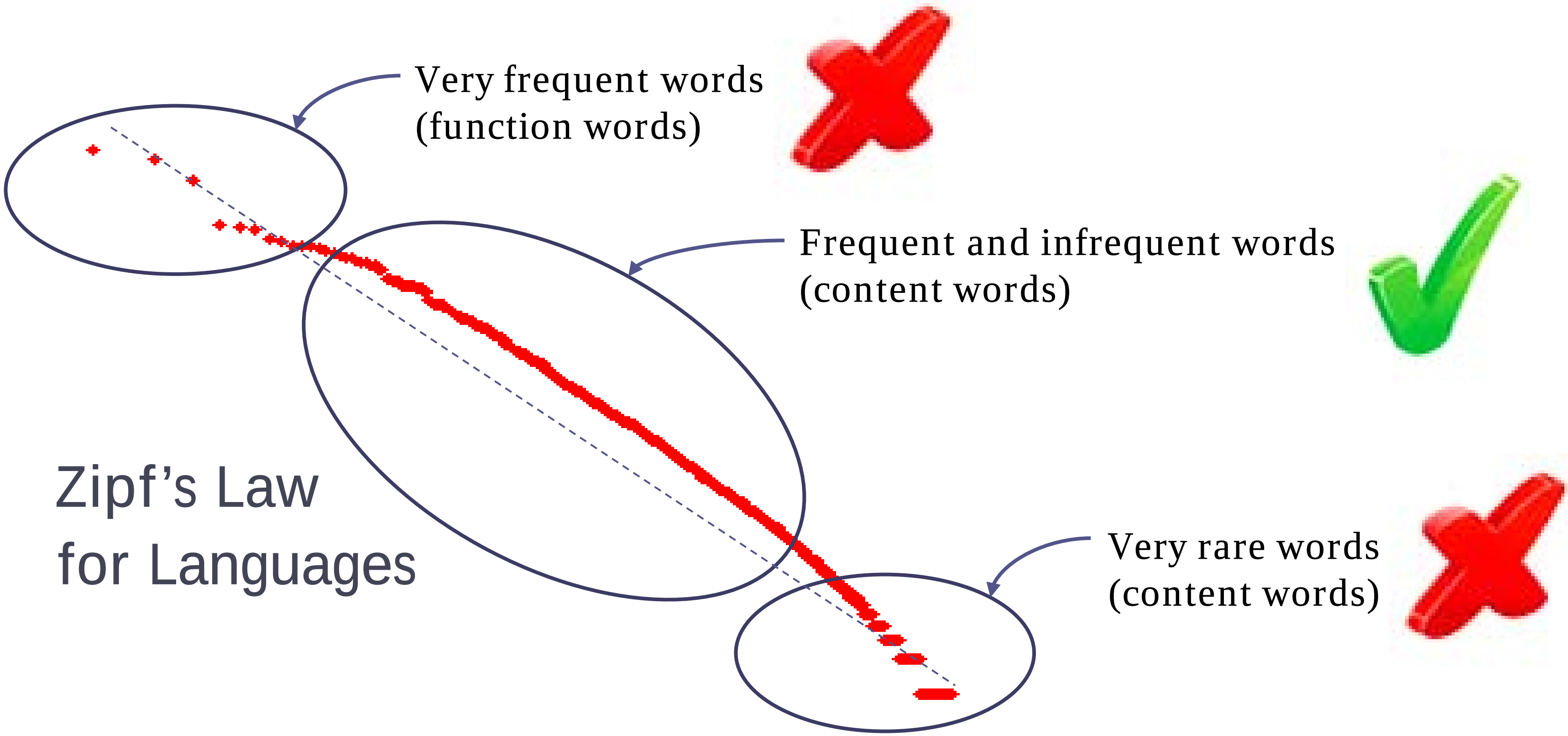
Document *z*

animals rain feeding environment response

- Relative word orderings within the documents are not taken into account

Weighting Strategies

- More discriminative words are more important !



TF-IDF Weighting Scheme*

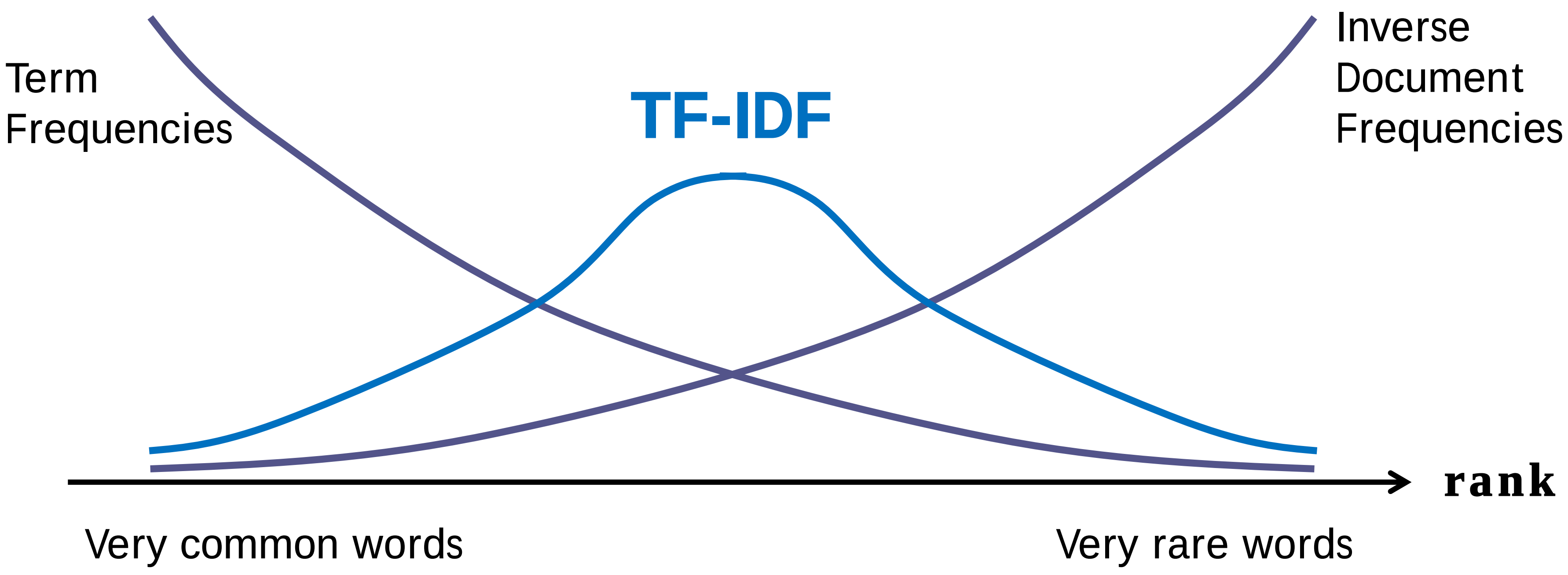
We want to favor words that are:

- Common within documents
  - Term-Frequency Weight (TF): it counts how many times a word occurs within a document
- Uncommon across documents
  - Inverse-Document-Frequency (IDF): it inversely accounts for the number of documents that contain a given word

** Spärck Jones, K. (1972), A statistical interpretation of term specificity and its application in retrieval, Journal of Documentation, 28(1), 11-21*

TF-IDF Weighting Effects

Higher weights are given to those words that are frequent within but infrequent across documents



TF-IDF Weighting Computation

- Term-Frequency (TF):

$$TF(w_i, d_j) = |w_i \epsilon d_j|$$

- Inverse-Document-Frequency (IDF):

$$IDF(w_i) = log\left(\frac{|D|}{1 + |d \epsilon D : w_i \epsilon d|}\right)$$

- TF-IDF with document length normalization:

$$TF\text{-}IFD(w_i, d_j) = \frac{TF(w_i, d_j)\ \ IFD(w_i)}{\sum_i |w_i \epsilon d_j|}$$

PMI Weighting Scheme*

- Point-wise Mutual Information (PMI)

$$PMI(w_i,d_j) = log\left(\frac{p(w_i,d_j)}{p(w_i)\,p(d_j)}\right)$$

- Positive PMI (PPMI)

$$PPMI(w_i,d_j) = \begin{cases} PMI(w_i,d_j) & if > 0 \\ 0 & otherwise \end{cases}$$

- Discounted PMI (compensates the tendency of PMI to increase the importance of infrequent events)

$$DPMI(w_i,d_j) = \delta_{ij}\,PMI(w_i,d_j)$$

** Church, K., Hanks, P. (1989), Word association norms, mutual information, and lexicography, in Proceedings of the 27th Annual Conference of the Association of Computational Linguistics, pp. 76-83*

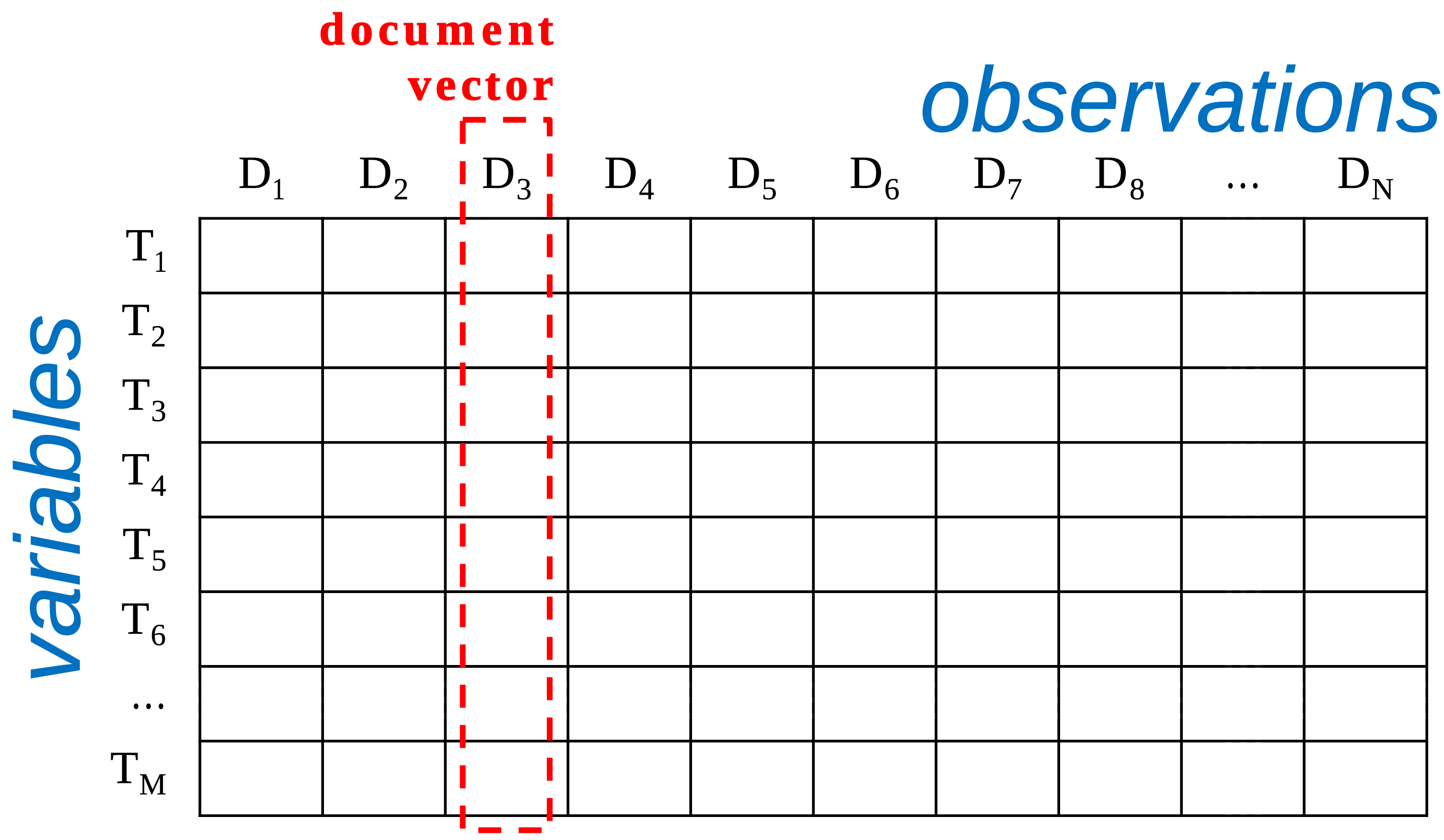
Section 1

Basic Concepts and Theoretical Framework

- The Distributional Hypothesis
- Vector Space Models and the Term-Document Matrix
- **Association Scores and Similarity Metrics**
- The Curse of Dimensionality and Dimensionality Reduction
- Semantic Cognition, Conceptualization and Abstraction

Document Vector Spaces

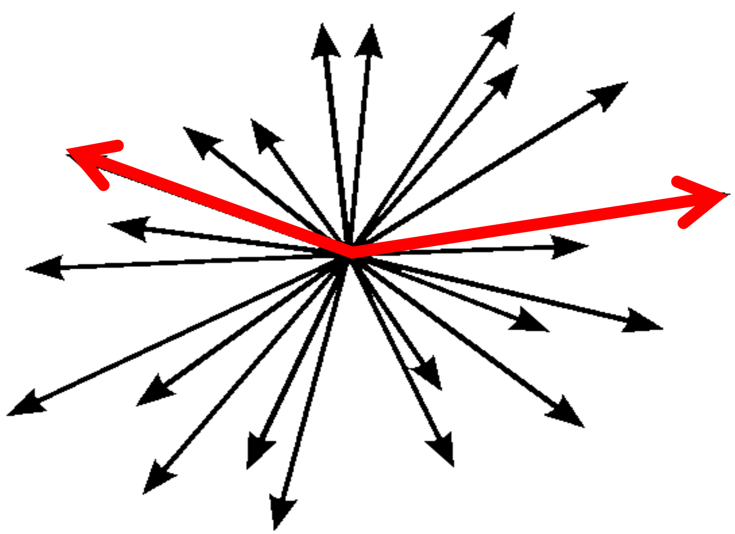
Pay attention to the columns of the term-document matrix



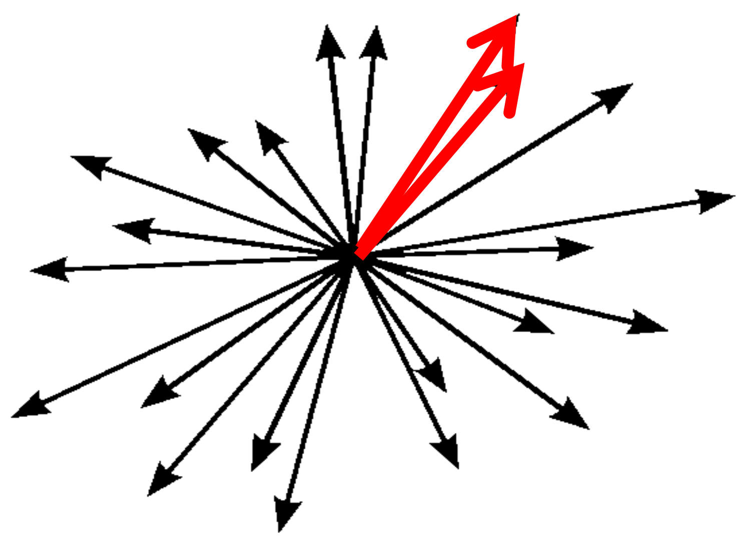
Document Vector Spaces

Association scores and similarity metrics can be used to assess the degree of semantic relatedness among documents

DISSIMILAR DOCUMENTS

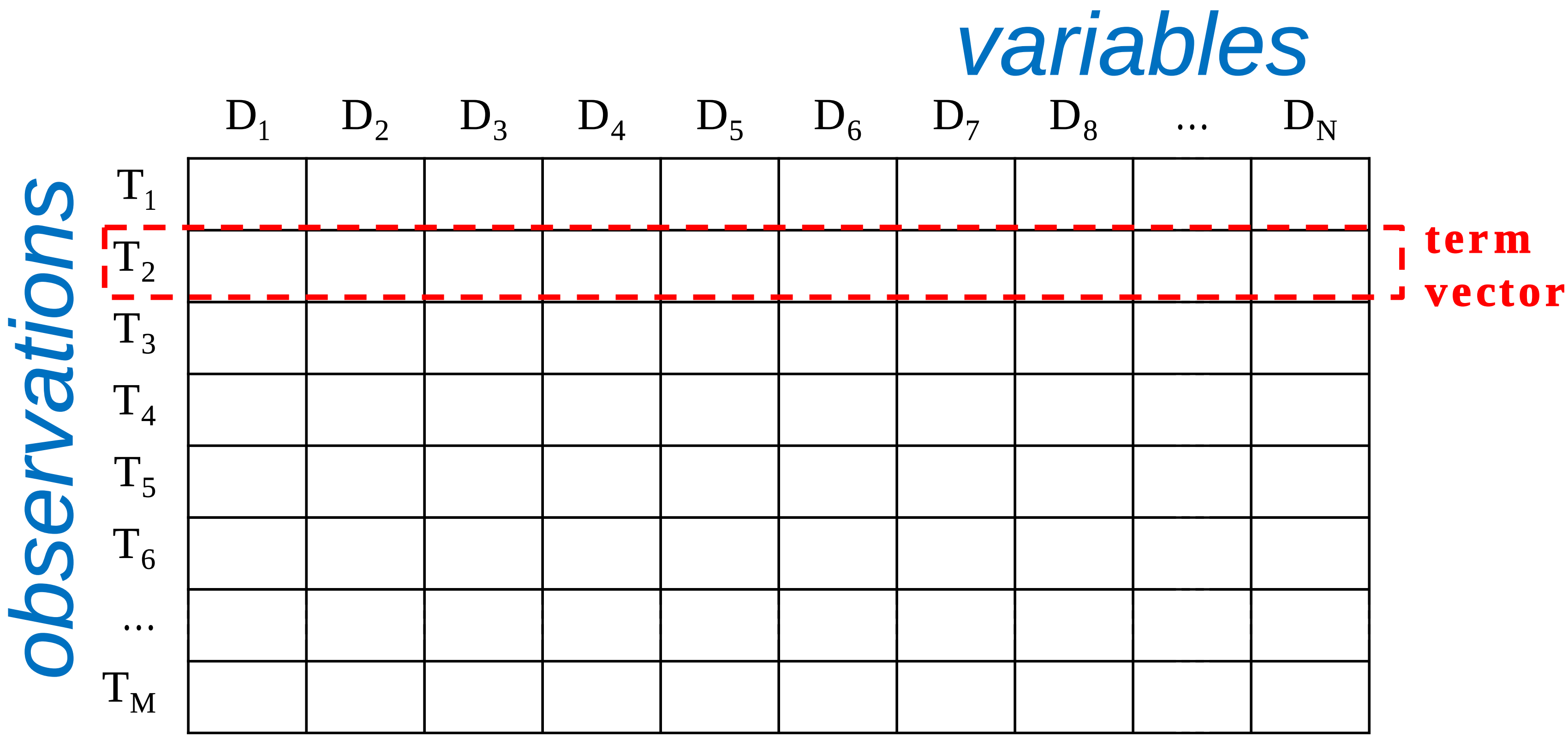


SIMILAR DOCUMENTS



Word Vector Spaces

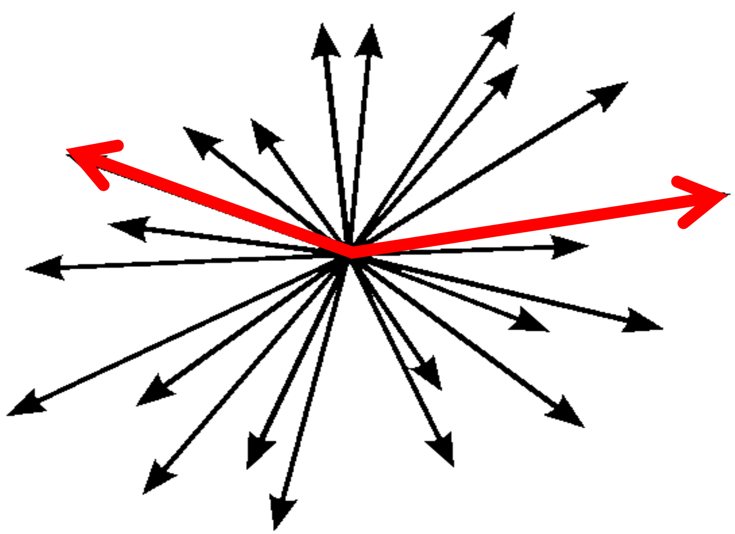
Pay attention to the rows of the term-document matrix



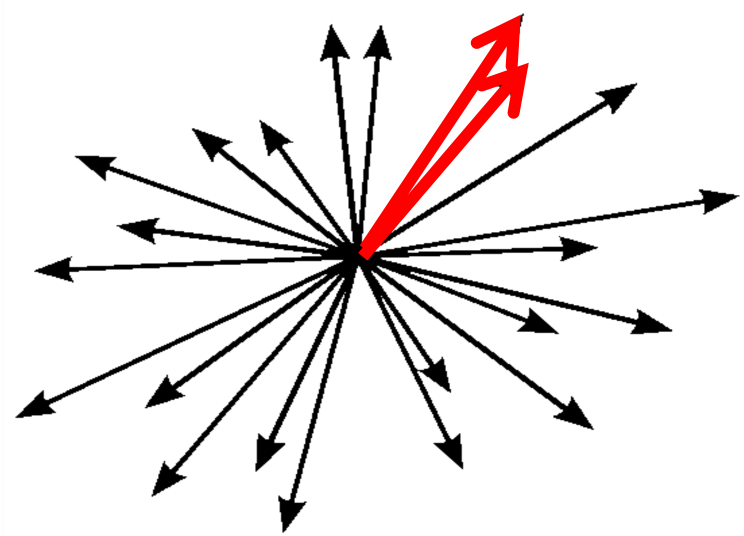
Word Vector Spaces

Association scores and similarity metrics can be used to assess the degree of semantic relatedness among words

DISSIMILAR TERMS



SIMILAR TERMS



Assessing Vector Similarities

- Association scores provide a means for measuring vector similarity
- Distances, on the other hand, provide a means for measuring vector dissimilarities
- Similarities and dissimilarities are in essence opposite measurements, and can be easily converted from one to another

Association Scores

- Dice:

$$dice(V1,V2) = \frac{2\,|N1 \cap N2|}{|N1| + |N2|}$$

- Jaccard:

$$jacc(V1,V2) = \frac{|N1 \cap N2|}{|N1 \cup N2|}$$

- cosine:

$$cos(V1,V2) = \frac{\langle V1,V2 \rangle}{||V1||\;||V2||}$$

Distance Metrics

- Hamming: $hm(V1,V2) = |N1 \cap Z2| + |Z1 \cap N2|$
- Euclidean: $d(V1,V2) = ||V1 - V2||$
- citiblock: $cb(V1,V2) = ||V1 - V2||_1$
- cosine: $dcos(V1,V2) = 1 - cos(V1,V2)$

Section 1

Basic Concepts and Theoretical Framework

- The Distributional Hypothesis
- Vector Space Models and the Term-Document Matrix
- Association Scores and Similarity Metrics
- **The Curse of Dimensionality and Dimensionality Reduction**
- Semantic Cognition, Conceptualization and Abstraction

The Curse of Dimensionality*

- Refers to the data sparseness problem that is intrinsic to high-dimensional spaces
- The problem results from the disproportionate increase of space volume with respect to the amount of available data
- If the statistical significance of results are to be maintained, then the amount of required data will grow exponentially with dimensionality

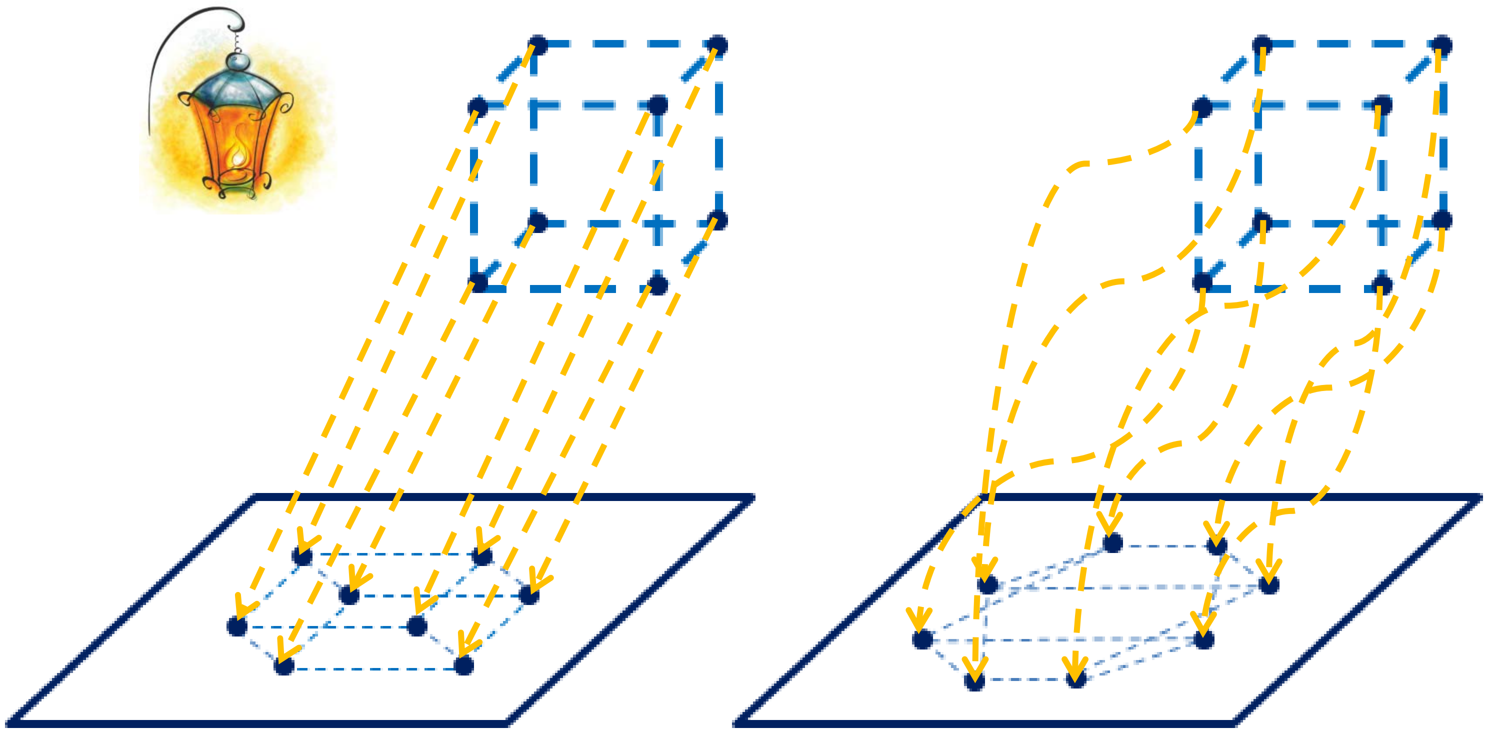
* *Bellman, R.E. (1957), Dynamic programming, Princeton University Press*

Dimensionality Reduction

- Deals with the “curse of dimensionality” problem
- Intends to explain the observations with less variables
- Attempts to find (or construct) the most informative variables

Provides a mathematical metaphor to the cognitive processes of Generalization and Abstraction !

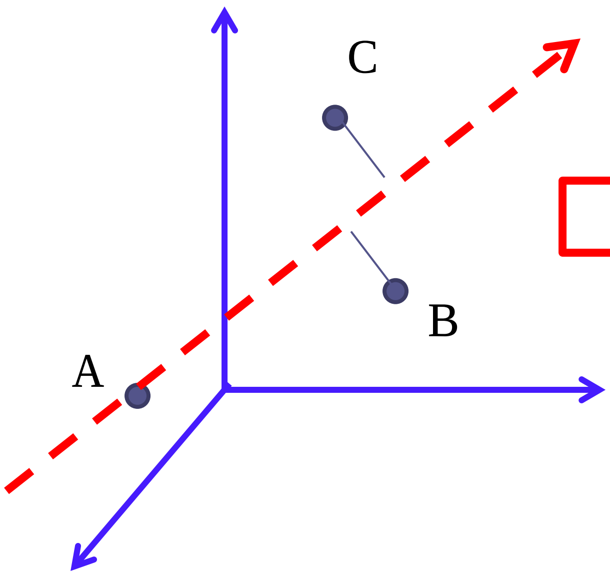
Types of Dimensionality Reduction



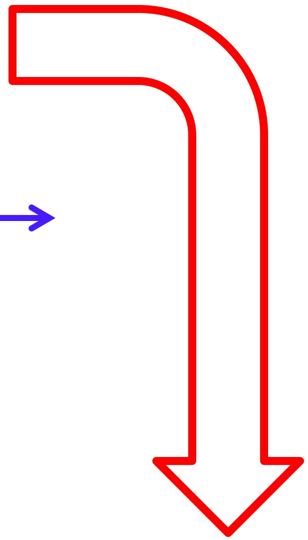
Linear projections are like shadows

Non-linear projections preserve structure

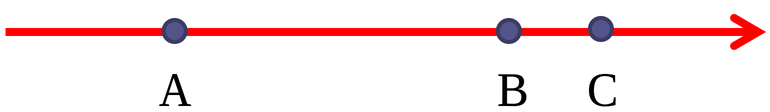
Example of a Linear Projection



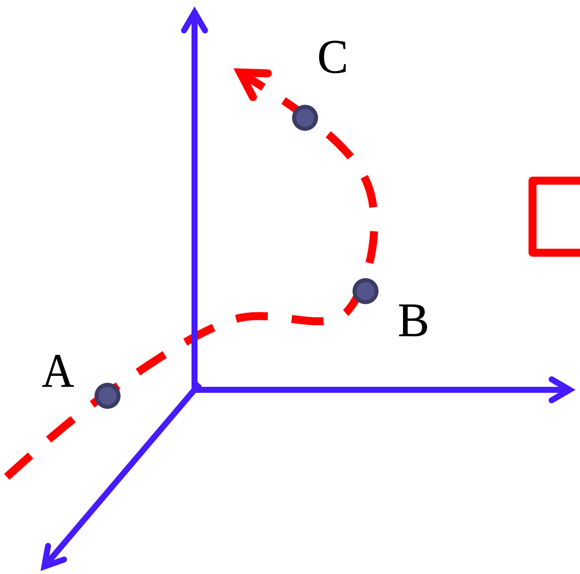
A	B	C
X_A	X_B	X_C
Y_A	Y_B	Y_C
Z_A	Z_B	Z_C



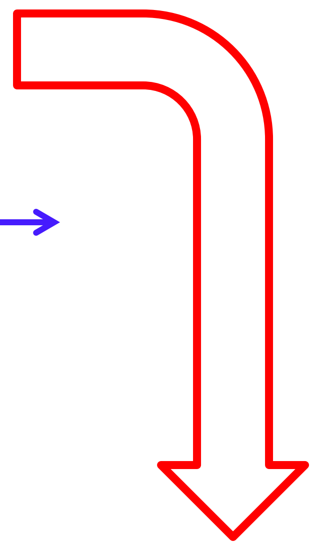
A	B	C
W_A	W_B	W_C



Example of a Non-linear Projection



A	B	C
X_A	X_B	X_C
Y_A	Y_B	Y_C
Z_A	Z_B	Z_C



A	B	C
W_A	W_B	W_C

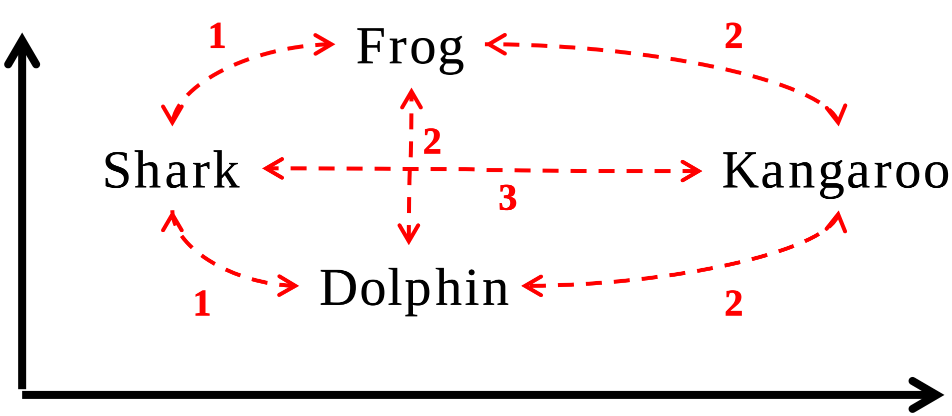
The Case of Categorical Data

Set of Observations

	leaps	swims	eggs
Frog	✓	✓	✓
Dolphin		✓	
Kangaroo	✓		
Shark		✓	✓

Dissimilarity Matrix

	Frog	Dolp.	Kang.	Shark
Frog	**0**	**2**	**2**	**1**
Dolphin	**2**	**0**	**2**	**1**
Kangaroo	**2**	**2**	**0**	**3**
Shark	**1**	**1**	**3**	**0**



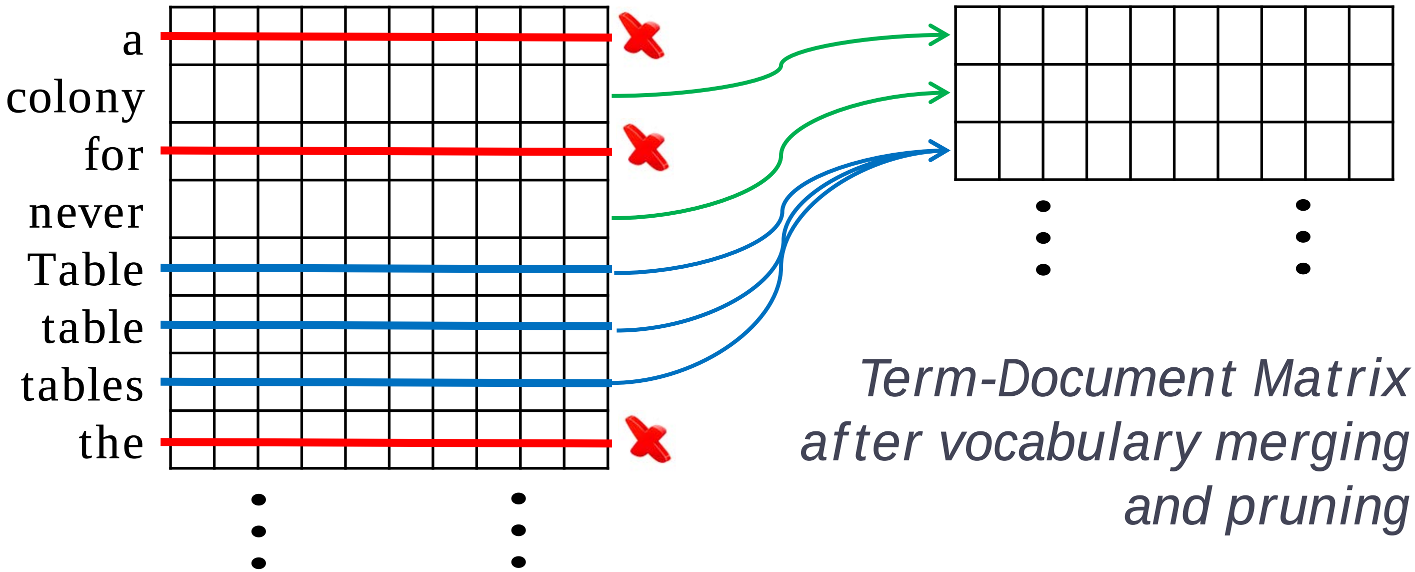
Low-dimensional Embedding

Some Popular Methods

- Variable merging and pruning:
  - Combine correlated variables (merging)
  - Eliminate uninformative variables (pruning)
- Principal Component Analysis (PCA)
  - Maximizes data variance in reduced space
- Multidimensional Scaling (MDS)
  - Preserves data structure as much as possible
- Autoencoders
  - Neural Network approach to Dimensionality Reduction

Variable Merging and Pruning

- Lemmatization and stemming (merging)
- Stop-word-list (pruning)



Term-Document Matrix after vocabulary merging and pruning

Principal Component Analysis (PCA)

- Eigenvalue decomposition of data covariance or correlation matrix (real symmetric matrix)

Diagonal matrix (eigenvalues)

$$\boldsymbol{M}_{N\times N} = \boldsymbol{Q}_{N\times N}\,\boldsymbol{\Lambda}_{N\times N}\,\boldsymbol{Q}^{T}_{N\times N}$$

Orthonormal matrix (eigenvectors)

- Singular value decomposition (SVD) of data matrix

Diagonal matrix (singular values)

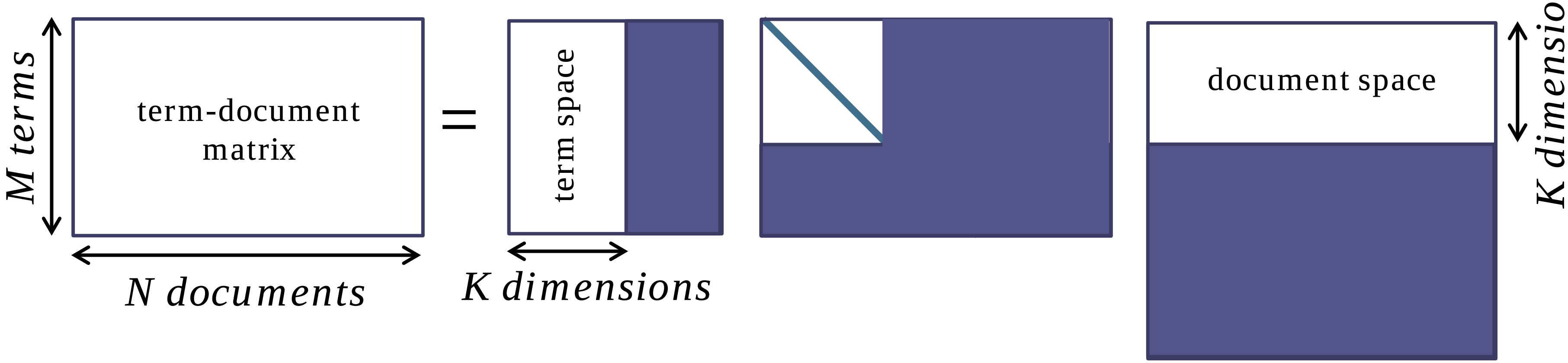
$$\boldsymbol{M}_{M\times N} = \boldsymbol{U}_{M\times M}\,\boldsymbol{\Sigma}_{M\times N}\,\boldsymbol{V}^{T}_{N\times N}$$

Unitary matrices

Latent Semantic Analysis (LSA)*

- Based on the Singular Value Decomposition (SVD) of a term-document matrix

$$\boldsymbol{M}_{M\times N} = \boldsymbol{U}_{M\times M} \boldsymbol{\Sigma}_{M\times N} \boldsymbol{V}^{T}_{N\times N}$$



$$\hat{\boldsymbol{M}}_{M\times N} \approx \boldsymbol{U}_{M\times K} \boldsymbol{\Sigma}_{K\times K} \boldsymbol{V}^{T}_{K\times N}$$

** 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, pp.391-407*

Multidimensional Scaling (MDS)

- Computes a low dimensional embedding by minimizing a “stress” function

Monotonic transformation

Input data dissimilarities

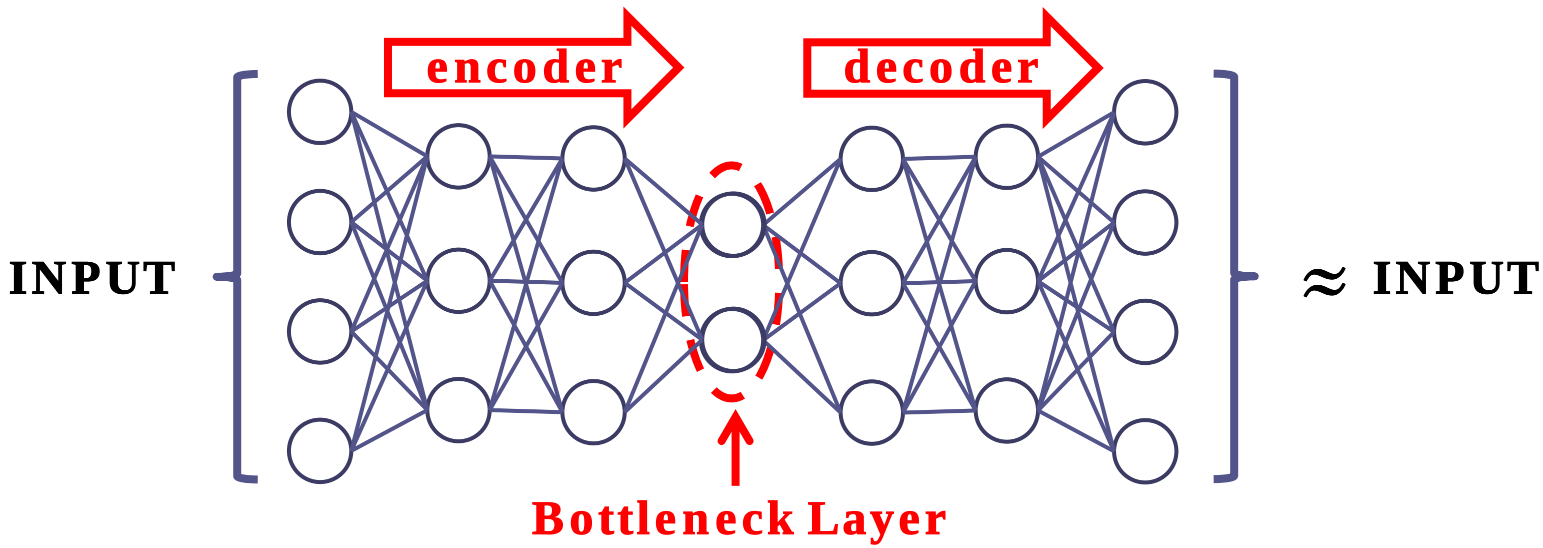
$$\text{Stress function} = \sqrt{\frac{\Sigma\,\Sigma\,(f(x_{ij}) - d_{ij})^2}{\text{Scaling factor}}}$$

Distances among points in the embedding

  - Metric MDS: directly minimizes stress function
  - Non-metric MDS: relaxes the optimization problem by using a monotonic transformation

Autoencoders*

- Symmetric feed-forward non-recurrent neural network
  - Restricted Boltzmann Machine (pre-training)
  - Backpropagation (fine-tuning)



** G. Hinton, R. Salakhutdinov "Reducing the dimensionality of data with neural networks", Science, 313(5786):504-507, 2006*

Section 1

Basic Concepts and Theoretical Framework

- The Distributional Hypothesis
- Vector Space Models and the Term-Document Matrix
- Association Scores and Similarity Metrics
- The Curse of Dimensionality and Dimensionality Reduction
- **Semantic Cognition, Conceptualization and Abstraction**

What is Cognition?

- Cognition is the process by which a sensory input is transformed, reduced, elaborated, stored, recovered, and used*
- Etymology:
  - Latin verb cognosco (“with”+“know”)
  - Greek verb gnósko (“knowledge”)
- It is a faculty that allows for processing information, reasoning and decision making

* *Neisser, U (1967) Cognitive psychology, Appleton-Century-Crofts, New York*

Three Important Concepts

- Memory: is the process in which information is encoded, stored, and retrieved
- Inference: is the process of deriving logical conclusions from premises known or assumed to be true (deduction, induction, abduction)
- Abstraction: is a generalization process by which concepts and rules are derived from a multiplicity of observations

Approaches to Semantic Cognition

- The hierarchical propositional approach*
  - Concepts are organized in a hierarchical fashion
- The parallel distributed processing approach**
  - Concept are stored in a distributed fashion and reconstructed by pattern completion mechanisms

** Quillian M.R. (1968) Semantic Memory, in Semantic Information Processing (ed. Minsky, M.) pp.227-270, MIT Press*

*** McClelland, J.L. and Rogers, T.T. (2003) The Parallel Distributed Processing Approach to Semantic Cognition, Nature Reviews, 4, pp.310-322*

Hierarchical Propositional Model

Example domain of living things

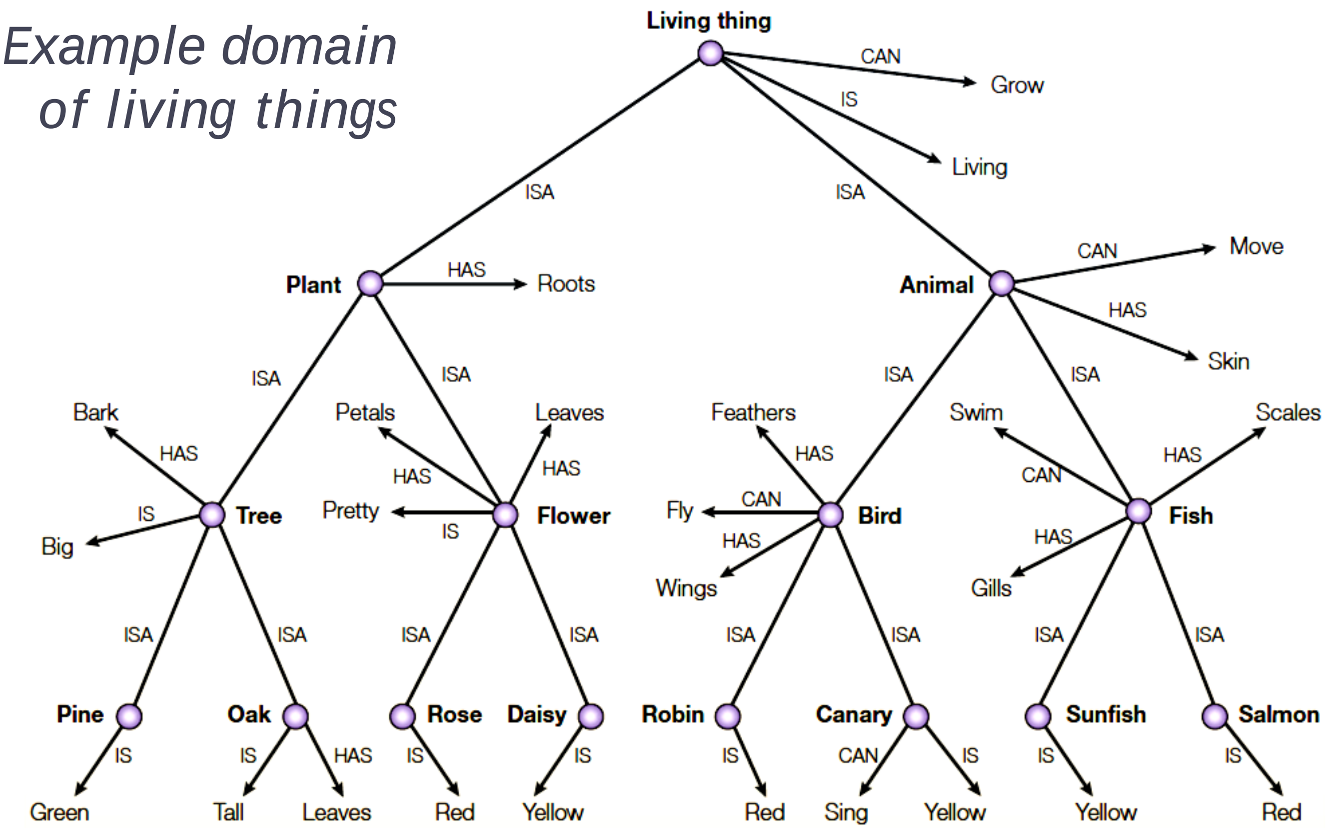


Image taken from: McClelland, J.L. and Rogers, T.T. (2003) The Parallel Distributed Processing Approach to Semantic Cognition, Nature Reviews, 4, pp.310-322

Advantages of Hierarchical Model

- Economy of storage
- Immediate generalization of
  - known propositions to new members
  - new propositions to known members
- Explains cognitive processes of *
  - general-to-specific progression in children
  - progressive deterioration in semantic dementia patients

** Warringtong, E.K. (1975) The Selective Impairment of Semantic Memory, The Quarterly of Journal Experimental Psychology, 27, pp.635-657*

Hierarchical Model Drawback!

There is strong experimental evidence of a graded category membership in human cognition

- Humans are faster verifying the statement *
  - ‘chicken is an animal’ than ‘chicken is a bird’
  - ‘robin is a bird’ than ‘chicken is a bird’
- This is better explained when the verification process is approached by means of assessing similarities across categories and elements

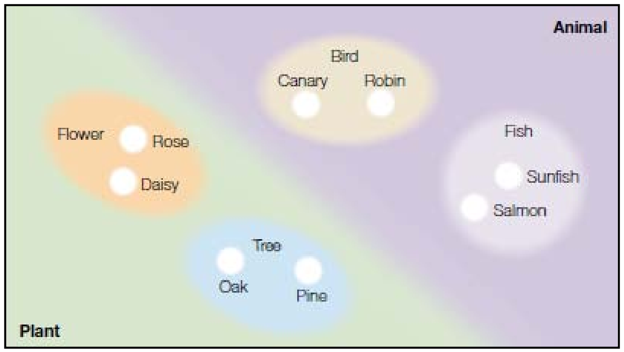
** Rips, L.J., Shoben, E.J. and Smith, E.E. (1973) Semantic distance and the verification of semantic relations, Journal of Verbal Learning and Verbal Behaviour, 12, pp.1-20*

Parallel Distributed Processing*

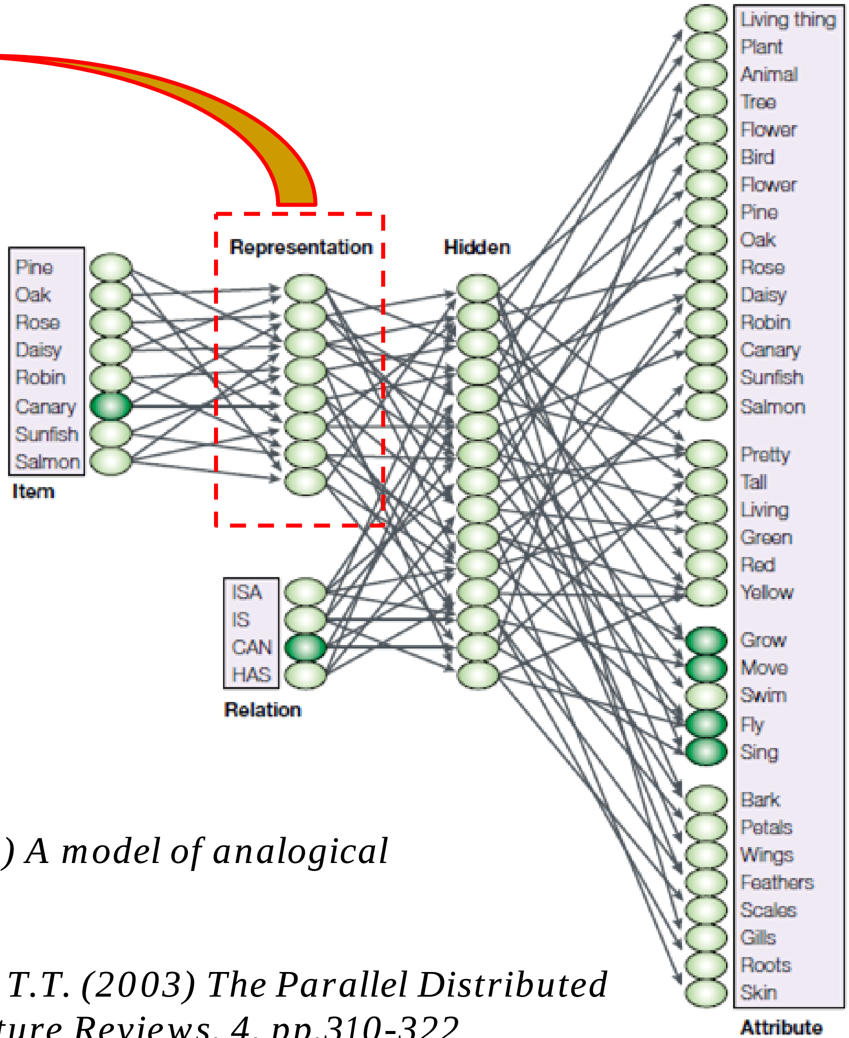
- Semantic information is stored in a distributed manner across the system
- Semantic information is “reconstructed” by means of a pattern completion mechanism
- The reconstruction process is activated as the response to a given stimulus

** McClelland, J.L. and Rogers, T.T. (2003) The Parallel Distributed Processing Approach to Semantic Cognition, Nature Reviews, 4, pp.310-322*

Rumelhart Connectionist Network*



Two-dimensional projection of the representation layer



** Rumelhart, D.E. and Abrahamsonm A.A. (1973) A model of analogical reasoning, Cognitive Psychology, 5, pp.1-28*

Image taken from: McClelland, J.L. and Rogers, T.T. (2003) The Parallel Distributed Processing Approach to Semantic Cognition, Nature Reviews, 4, pp.310-322

Advantages of the PDP Model*

- Also explains both cognitive processes of development and degradation
- Additionally, it can explain the phenomenon of graded category membership:
  - use of intermediate level categories (basic level**)
  - over-generalization of more frequent items

** McClelland, J.L. and Rogers, T.T. (2003) The Parallel Distributed Processing Approach to Semantic Cognition, Nature Reviews, 4, pp.310-322*

*** Rosch E., Mervis C.B., Gray W., Johnson D. and Boyes-Braem, P. (1976) Basic objects in natural categories, Cognitive Psychology, 8, pp.382-439*

PDP, DH and Vector Spaces

- The **Parallel Distributed Processing** (PDP) model explains a good amount of observed cognitive semantic phenomena
- In addition, the connectionist approach has a strong foundation on neurophysiology
- Both PDP and **Distributional Hypothesis** (DH) use differences/similarities over a feature space to model the semantic phenomenon
- **Vector Spaces** constitute a great mathematical framework for this endeavor !!!

Section 1

Main references for this section

- M. Sahlgren, 2006, “The distributional hypothesis”
- P. D. Turney and P. Pantel, 2010, “From frequency to meaning: vector space models of semantics”
- S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman, 1990, “Indexing by latent semantic analysis”
- G. Hinton and R. Salakhutdinov, 2006, “Reducing the dimensionality of data with neural networks”
- J. L. McClelland and T. T. Rogers, 2003, “The Parallel Distributed Processing Approach to Semantic Cognition”

Section 1

Additional references for this section

- Firth, J.R. (1957) A synopsis of linguistic theory 1930-1955, in Studies in linguistic analysis, 51: 1-31
- Harris, Z. (1970) Distributional Structure, in Papers in structural and transformational linguistics
- Saussure, F. (1916) Course in General Linguistics
- Salton G. (1971) The SMART retrieval system: Experiments in automatic document processing
- Spärck Jones, K. (1972), A statistical interpretation of term specificity and its application in retrieval, Journal of Documentation, 28(1), 11-21
- Church, K., Hanks, P. (1989), Word association norms, mutual information, and lexicography, in Proceedings of the 27th Annual Conference of the Association of Computational Linguistics, pp. 76-83

Section 1

Additional references for this section

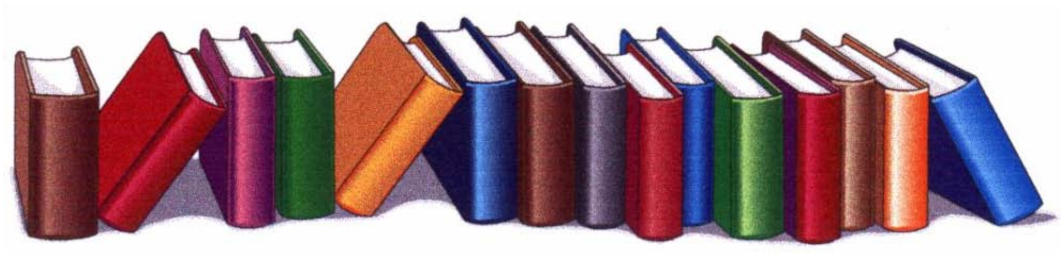
- Bellman, R.E. (1957), Dynamic programming, Princeton University Press
- Neisser, U (1967) Cognitive psychology, Appleton-Century-Crofts, New York
- Quillian M.R. (1968) Semantic Memory, in Semantic Information Processing (ed. Minsky, M.) pp.227-270, MIT Press
- Warringtong, E.K. (1975) The Selective Impairment of Semantic Memory, The Quarterly of Journal Experimental Psychology, 27, pp.635-657
- Rips, L.J., Shoben, E.J. and Smith, E.E. (1973) Semantic distance and the verification of semantic relations, Journal of Verbal Learning and Verbal Behaviour, 12, pp.1-20
- Rumelhart, D.E. and Abrahamsonm A.A. (1973) A model of analogical reasoning, Cognitive Psychology, 5, pp.1-28
- Rosch E., Mervis C.B., Gray W., Johnson D. and Boyes-Braem, P. (1976) Basic objects in natural categories, Cognitive Psychology, 8, pp.382-439

Section 2

Vector Spaces in Monolingual NLP

- **The Semantic Nature of Vector Spaces**
- Information Retrieval and Relevance Ranking
- Word Spaces and Related Word Identification
- Semantic Compositionality in Vector Spaces

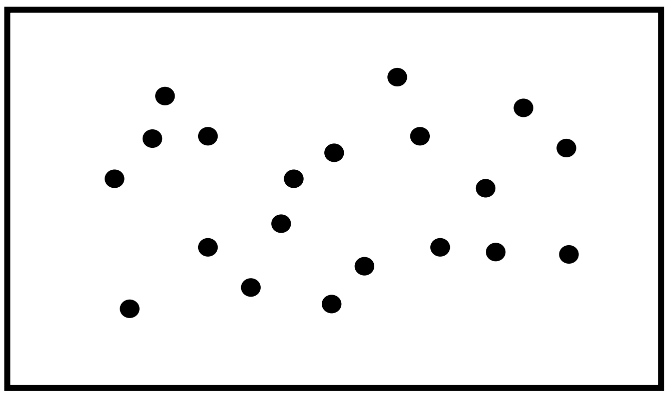
Constructing Semantic Maps



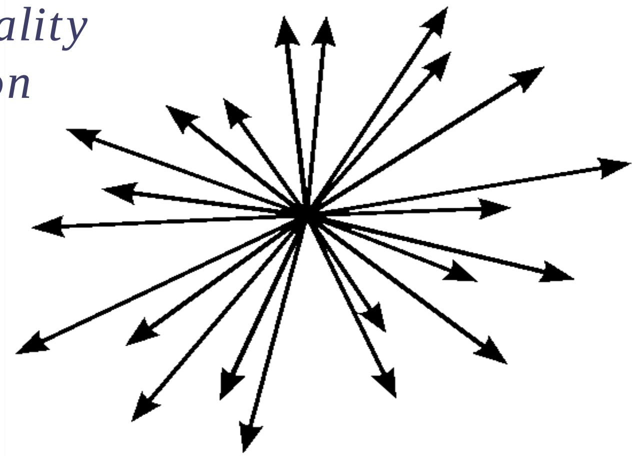
Document collection

TF-IDF Weighting

Dimensionality Reduction



"Semantic Map" of words or documents

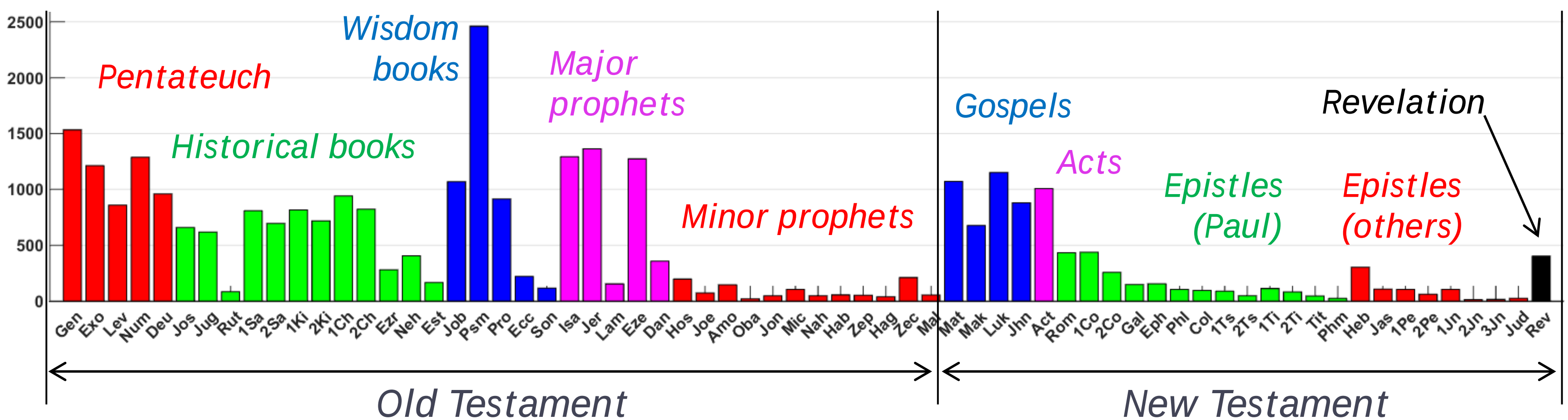


Vector Space of words or documents

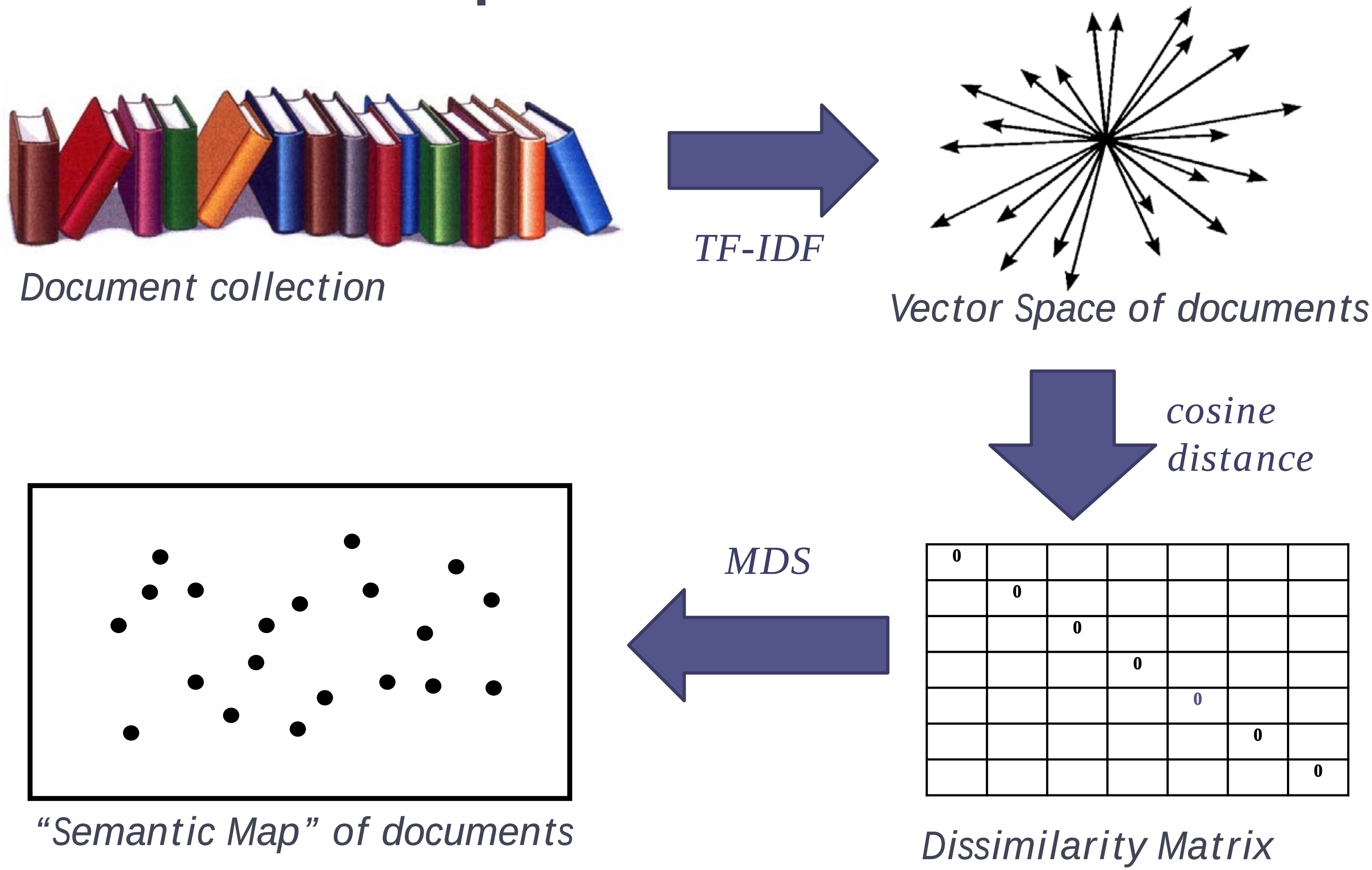
Document Collection

- The Holy Bible
  - 66 books → 1189 chapters → 31103 verses
  - ≈700K running words → ≈12K vocabulary terms

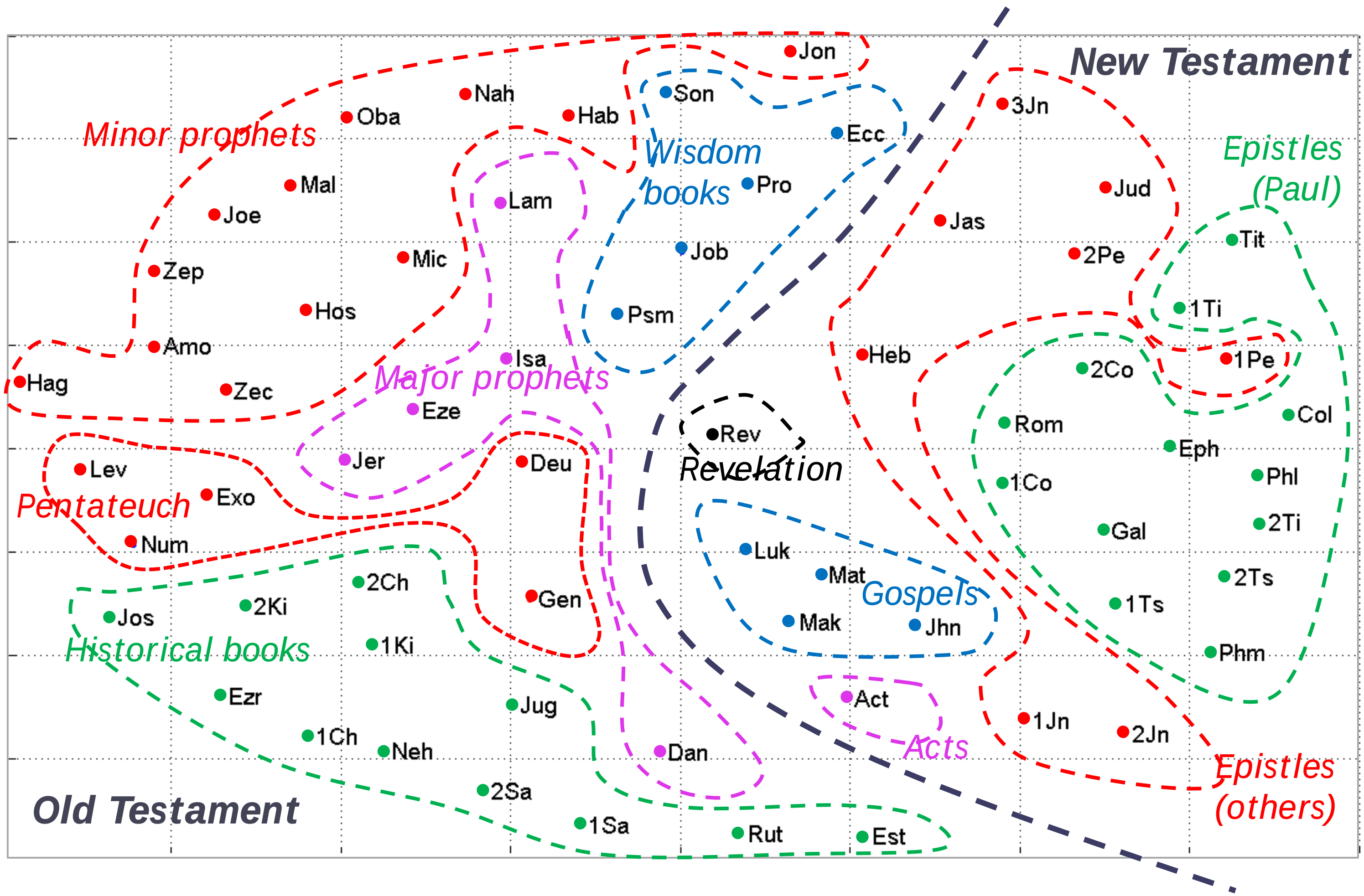
Distribution of verses per book within the collection



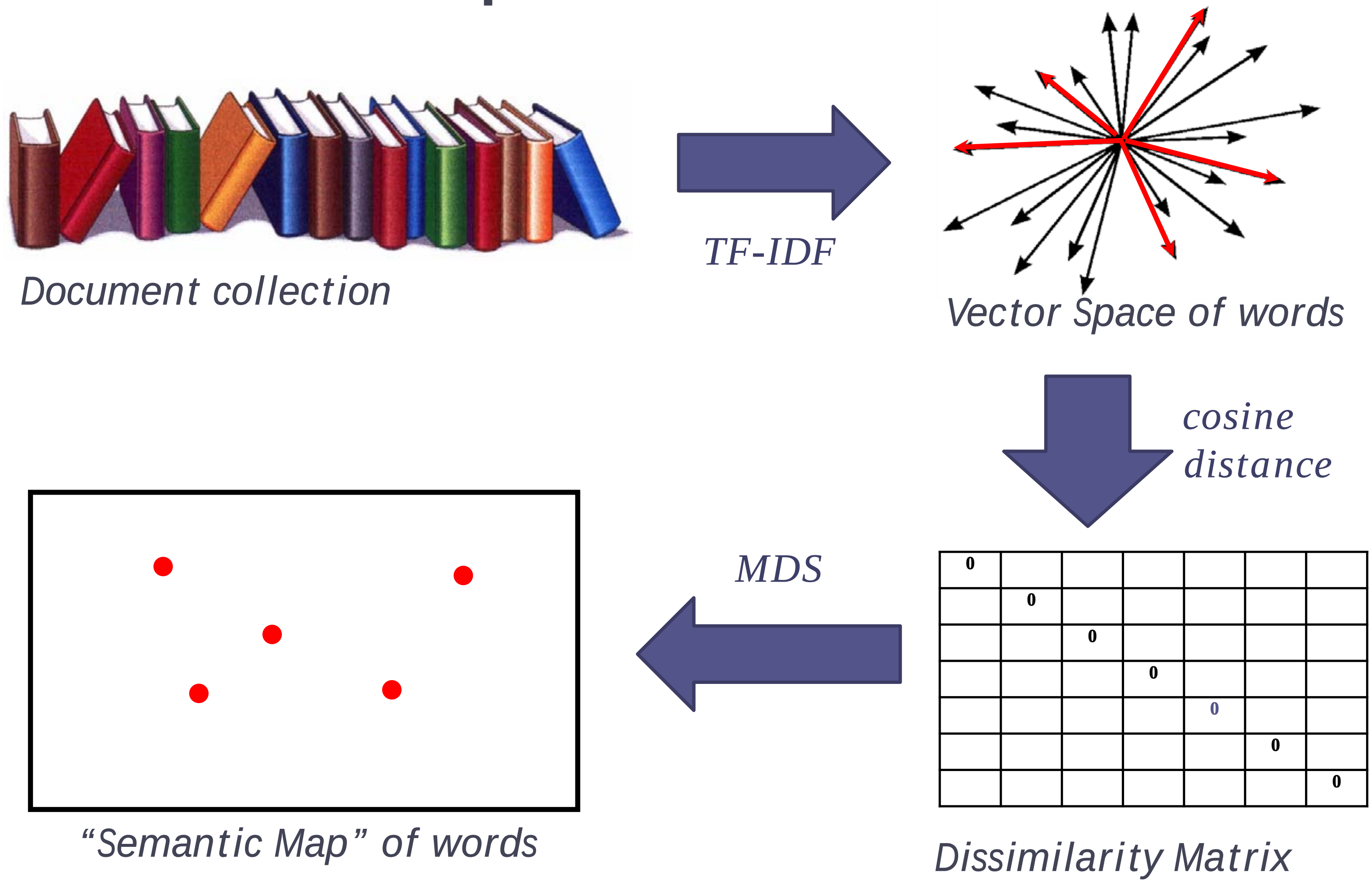
Semantic Maps of Documents



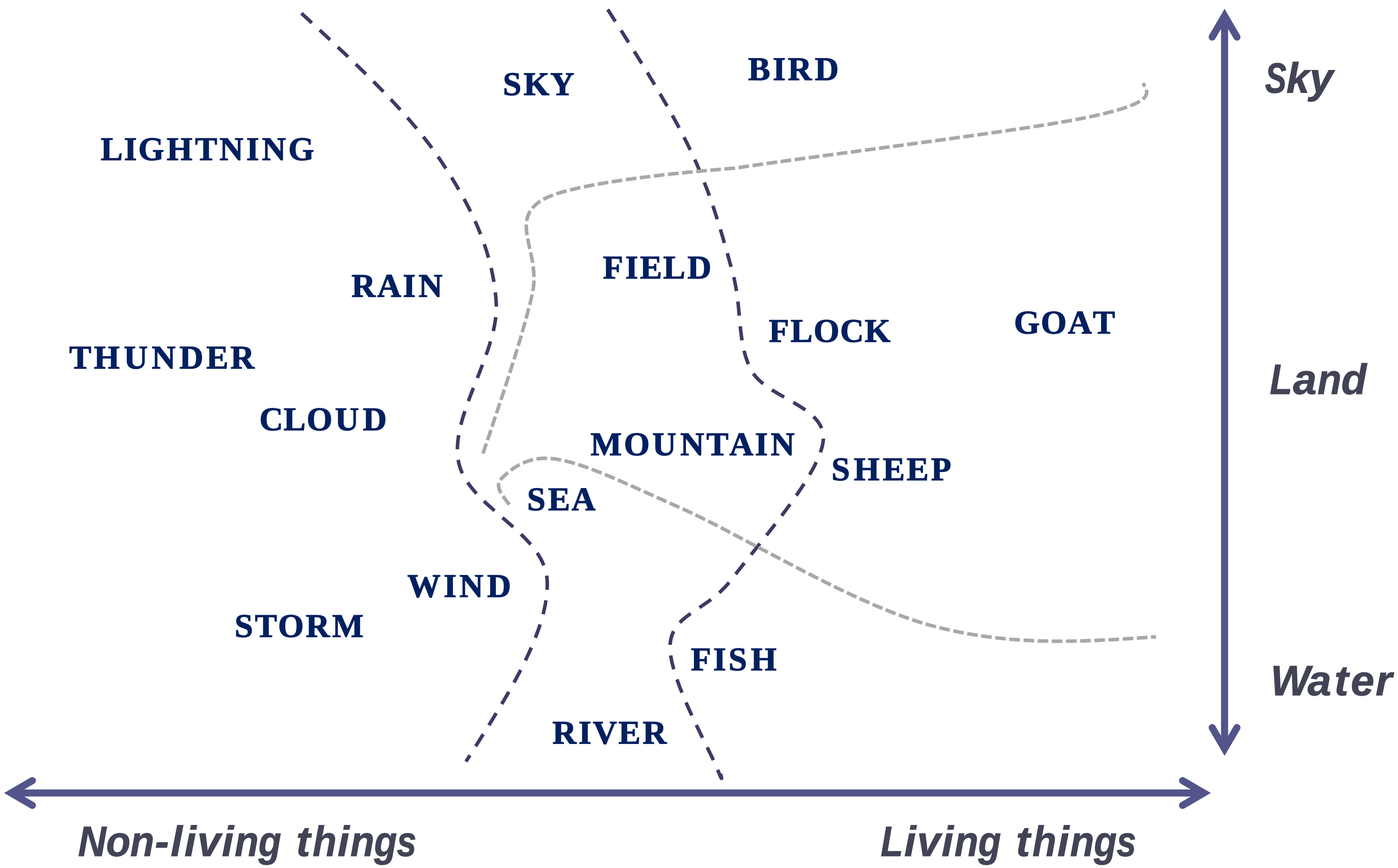
Semantic Maps of Documents



Semantic Maps of Words



Semantic Maps of Words

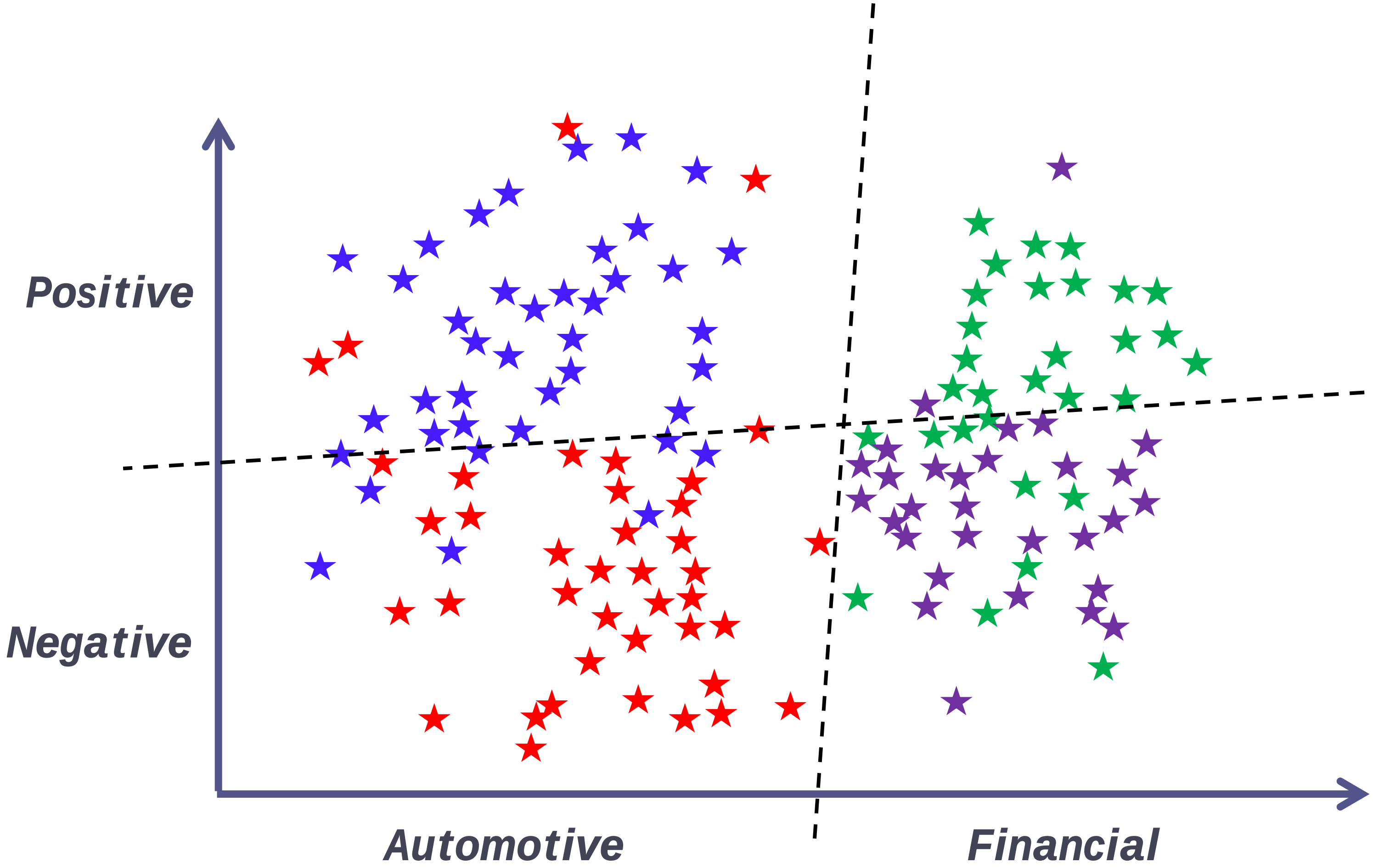


Discriminating Meta-categories

Opinionated content from rating website (Spanish)

- Positive and negative comments gathered from financial and automotive domains:
  - 2 topic categories: automotive and financial
  - 2 polarity categories: positive and negative
- Term-document matrix was constructed using full comments as documents
- A two-dimensional map was obtained by applying MDS to the vector space of documents

Discriminating Meta-categories



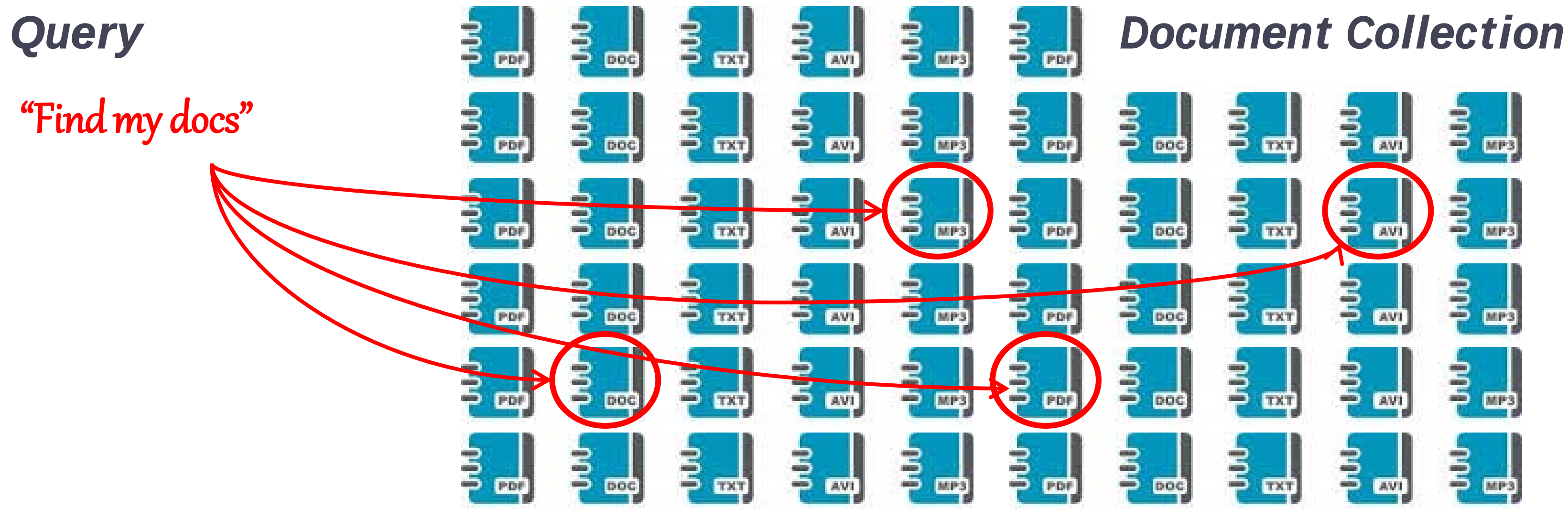
Section 2

Vector Spaces in Monolingual NLP

- The Semantic Nature of Vector Spaces
- **Information Retrieval and Relevance Ranking**
- Word Spaces and Related Word Identification
- Semantic Compositionality in Vector Spaces

Document Search: the IR Problem

- Given an informational need ("search query")
- and a **very large** collection of documents,
- find those documents that are relevant to it



Precision and Recall

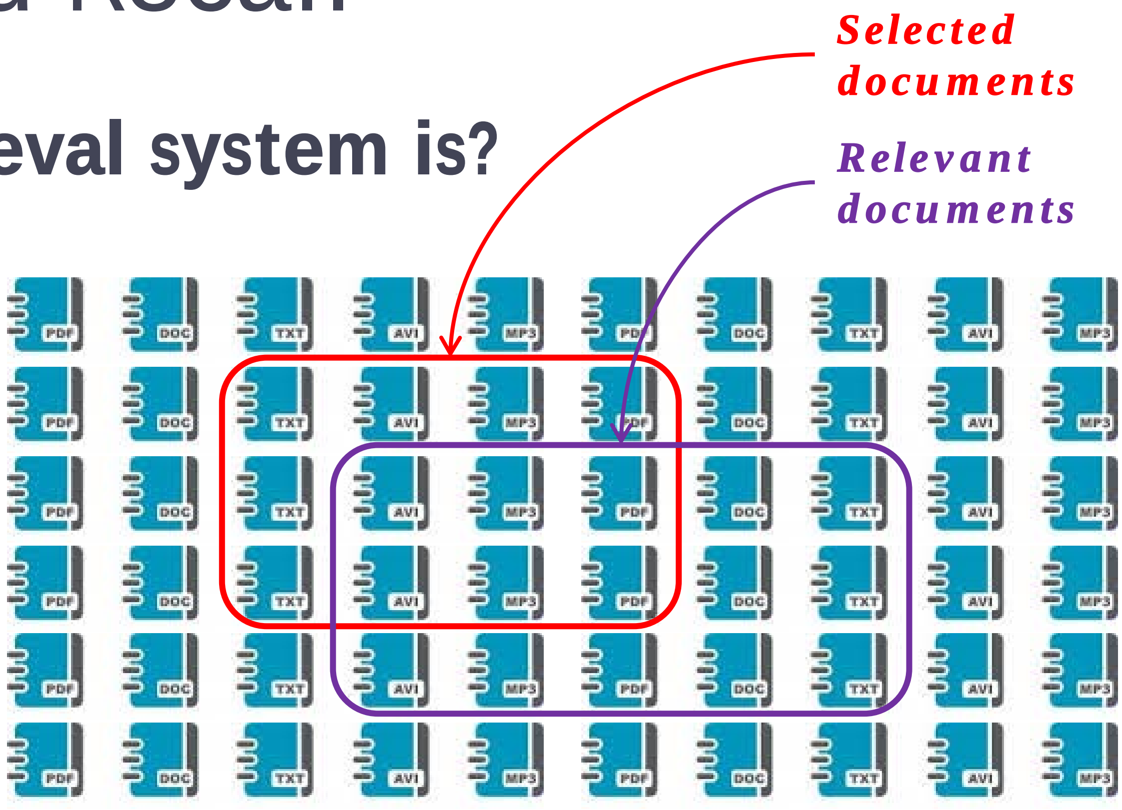
How good a retrieval system is?

$$TP = RD \cap SD$$

$$TN = \neg RD \cap \neg SD$$

$$FP = \neg RD \cap SD$$

$$FN = RD \cap \neg SD$$



$$precision = \frac{TP}{TP + FP} \quad recall = \frac{TP}{TP + FN} \quad F\text{-}score = 2\,\frac{precision \times recall}{precision + recall}$$

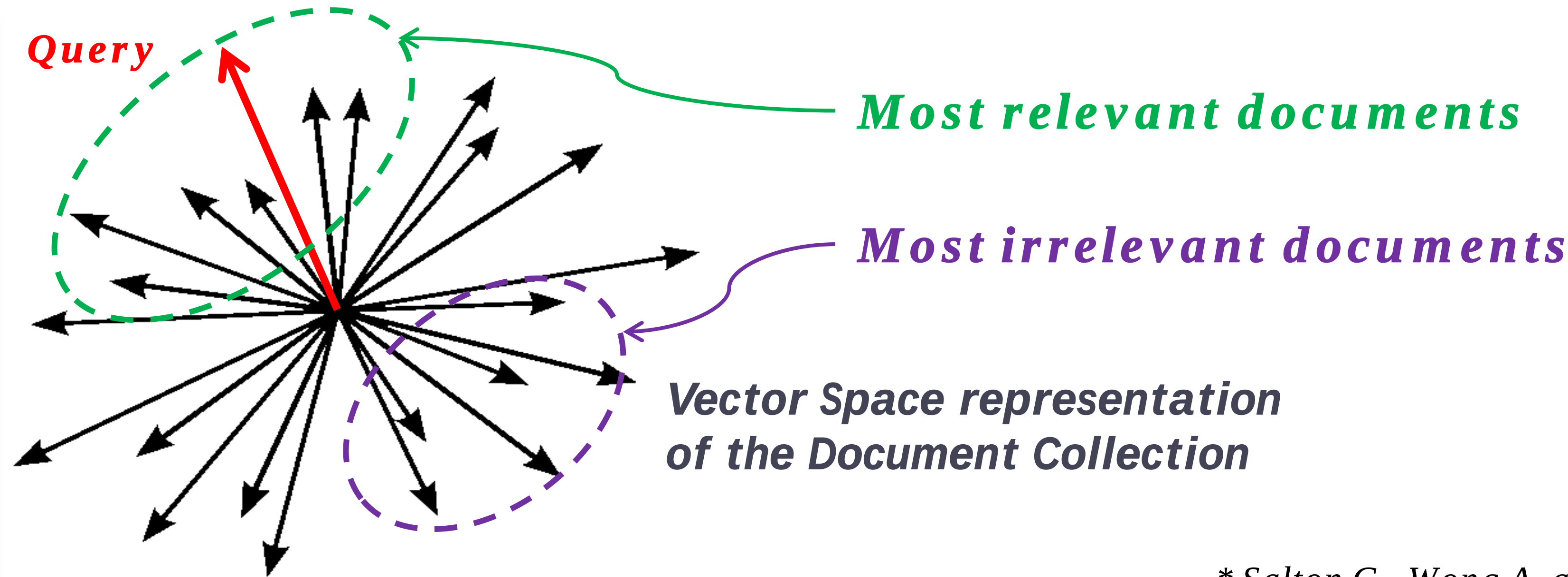
Binary Search*

- Keyword based (query = list of keywords)
  - AND-search: selects documents containing all keywords in the query
  - OR-search: selects documents containing at least one of the keywords in the query
- Documents are either relevant or not relevant (binary relevance criterion)

** Lee, W.C. and Fox, E.A. (1988) Experimental comparison of schemes for interpreting Boolean queries. Technical Report TR-88-27, Computer Science, Virginia Polytechnic Institute and State University*

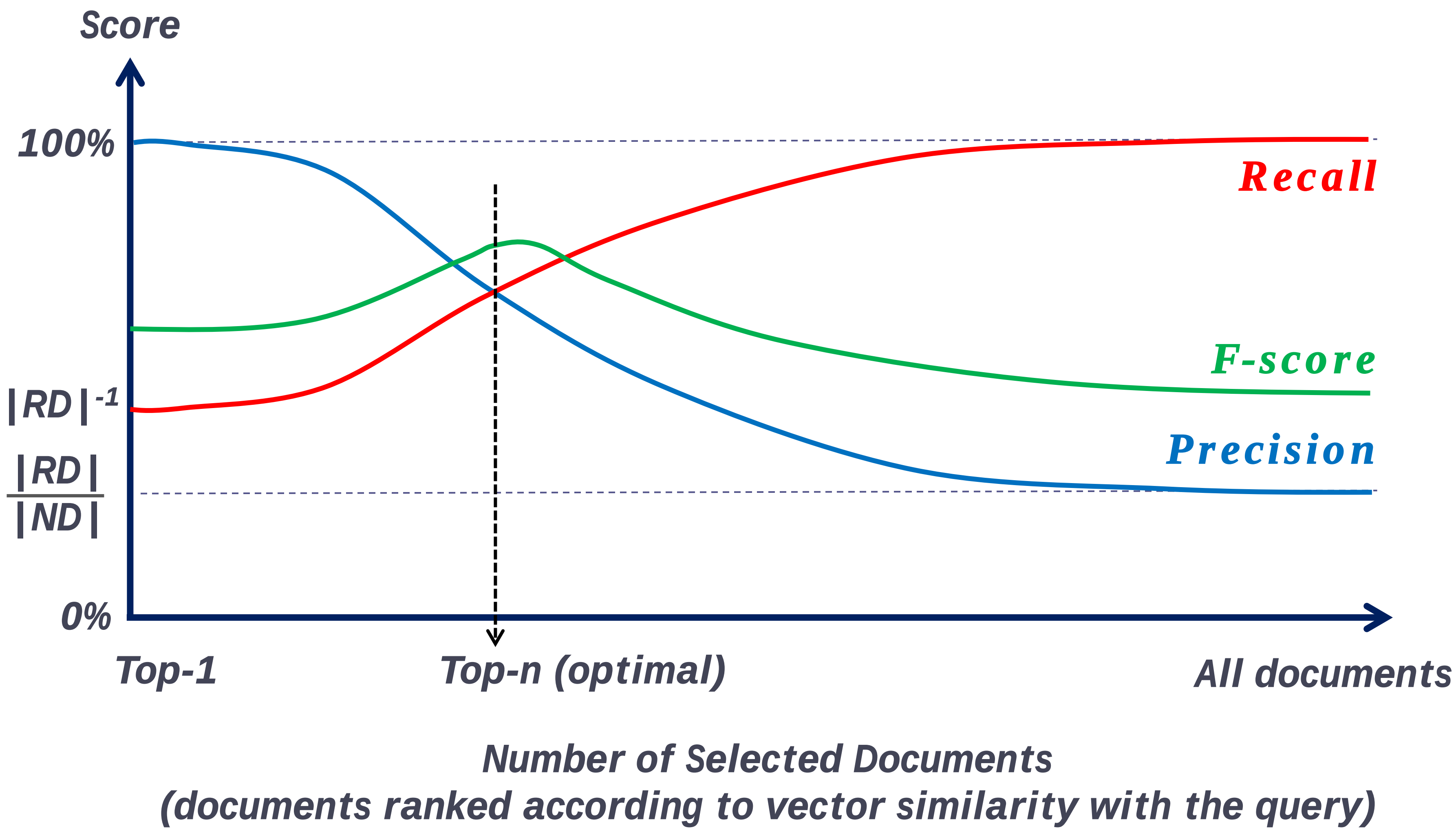
Vector Space Search*

- Keyword based (query = list of keywords)
- Uses vector similarity scores to assess document relevance (a graded relevance criterion)



* Salton G., Wong A. and Yang C.S. (1975) A vector space for automatic indexing. Communications of the ACM, 18(11), pp. 613-620

Precision/Recall Trade-off



Illustrative Example*

Consider a collection of 2349 paragraphs extracted from three different books:

- Oliver Twist by Charles Dickens
  - 840 paragraphs from 53 chapters
- Don Quixote by Miguel de Cervantes
  - 843 paragraphs from 126 chapters
- Pride and Prejudice by Jane Austen
  - 666 paragraphs from 61 chapters

** Banchs R.E. (2013) Text Mining with MATLAB, Springer , chap. 11, pp. 277-311*

Illustrative Example

Distribution of paragraphs per book and chapter

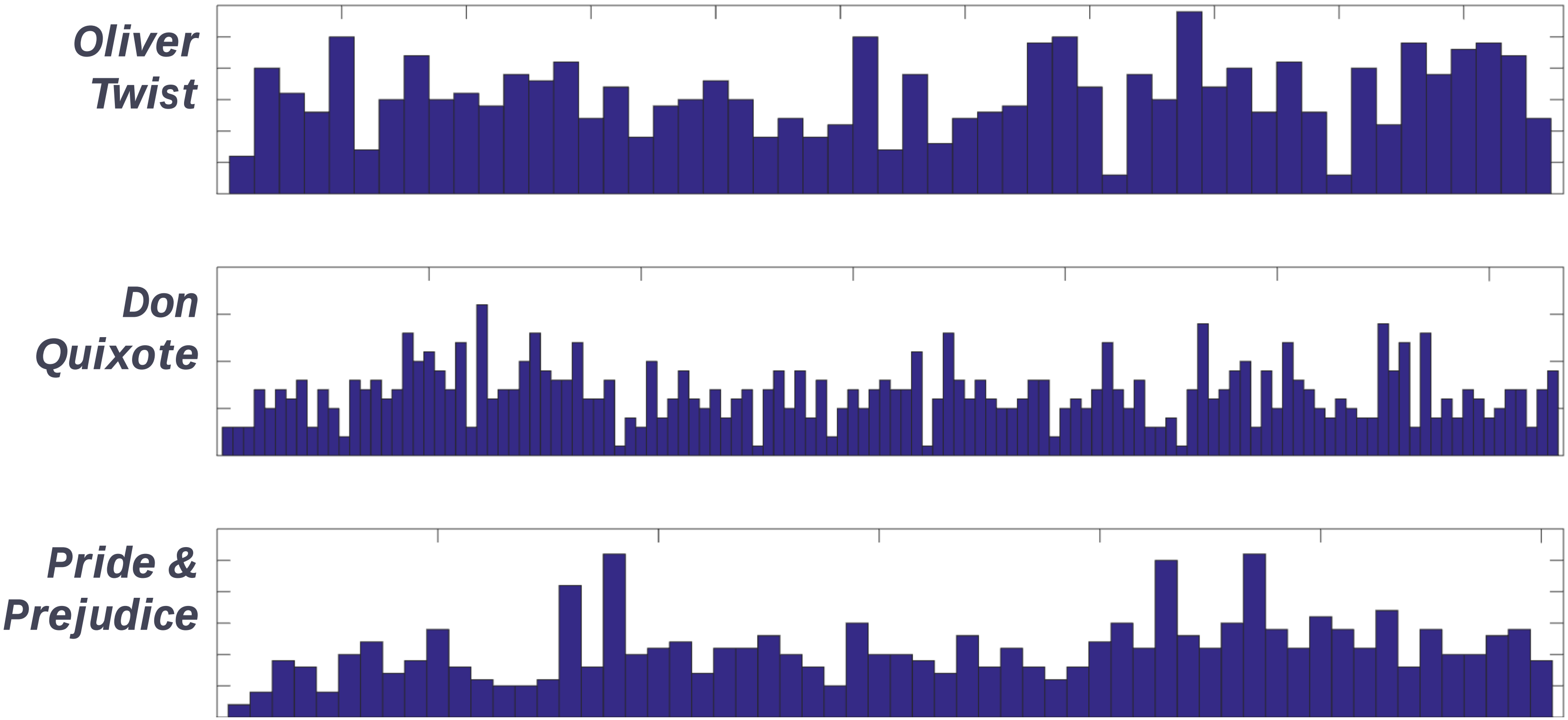


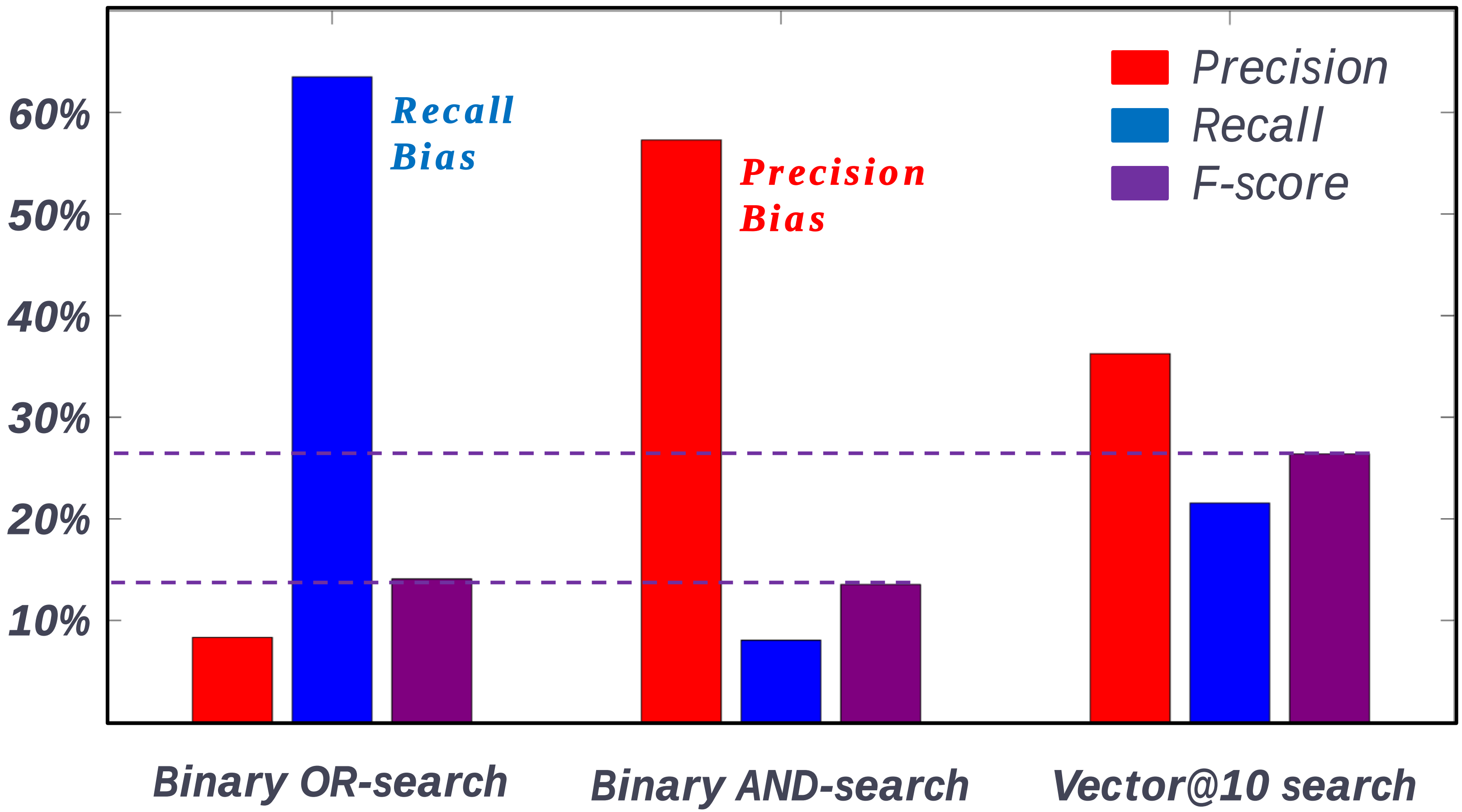
Image taken from Banchs R.E. (2013) Text Mining with MATLAB, Springer , chap. 11, pp. 277-311

Illustrative Example

Consider a set of 8 search queries:

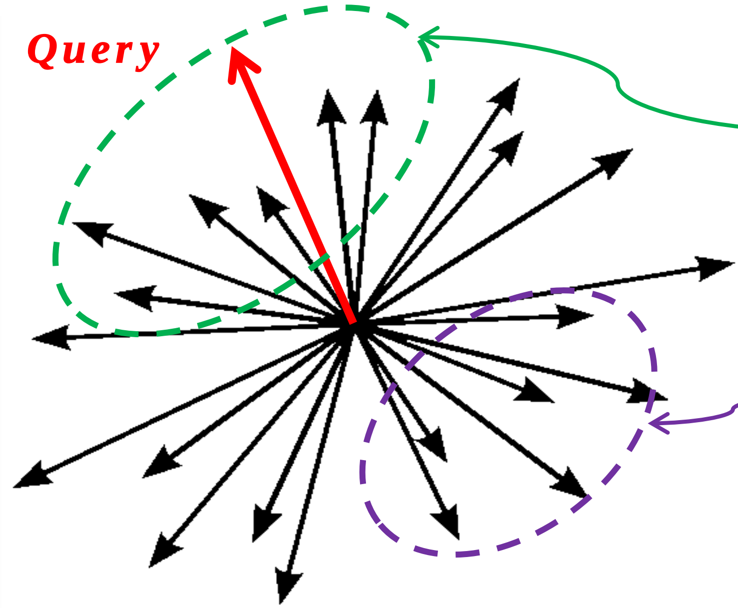
Query	***Relevant Book and Chapter***
oliver, twist, board	Oliver Twist, chapter 2
london, road	Oliver Twist, chapter 8
brownlow, grimwig, oliver	Oliver Twist, chapter 14
curate, barber, niece	Don Quixote, chapter 53
courage, lions	Don Quixote, chapter 69
arrival, clavileno, adventure	Don Quixote, chapter 93
darcy, dance	Pride & Prejudice, chapter 18
gardiner, housekeeper, elizabeth	Pride & Prejudice, chapter 43

Experimental Results



Automatic Relevance Feedback*

Use first search results to improve the search!



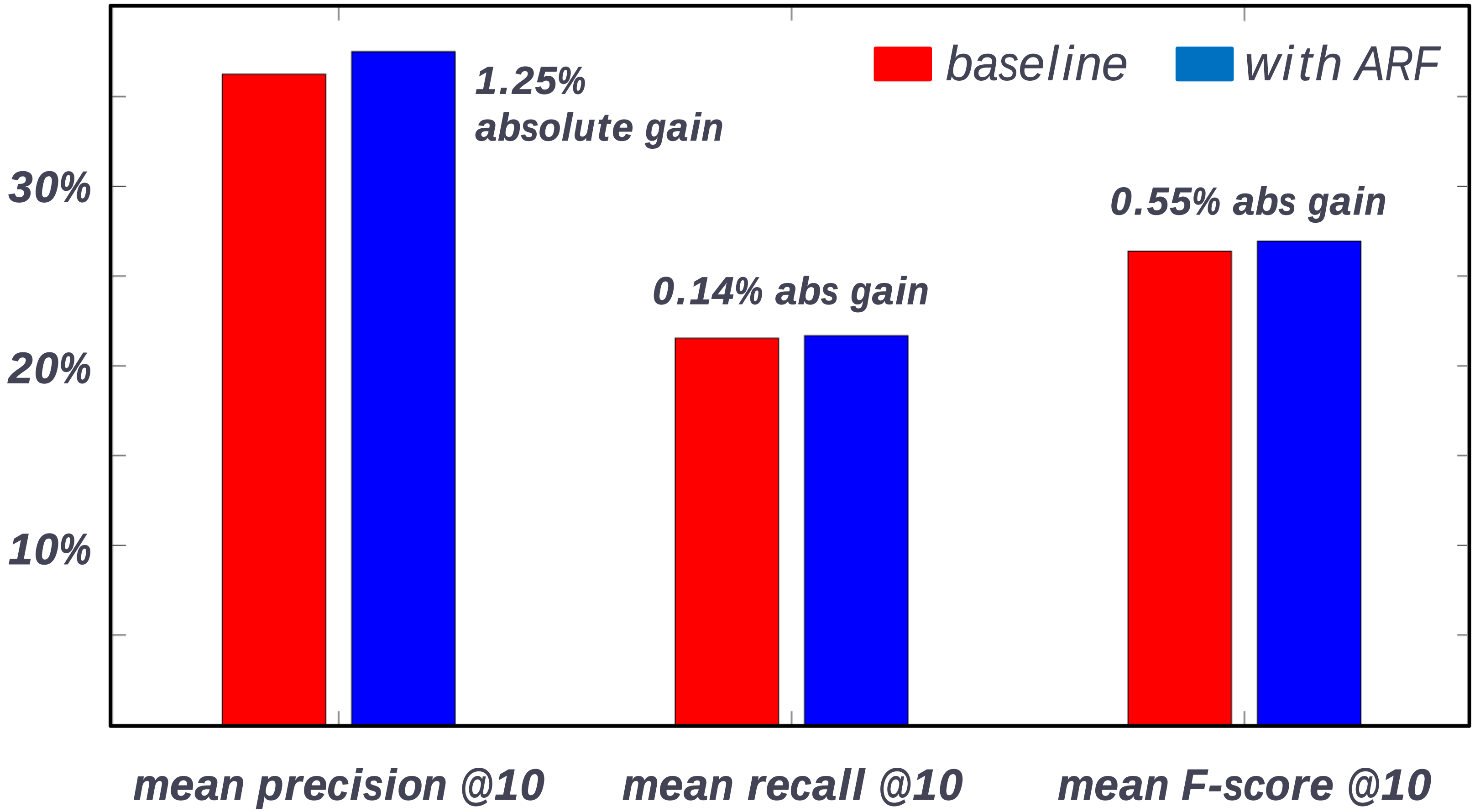
The most relevant documents should contain words that are good additional query keywords

The most irrelevant documents should contain words that are to be avoided as query keywords

$$newQuery = originalQuery + \alpha \frac{1}{|D_R|} \sum D_R - \beta \frac{1}{|D_{NR}|} \sum D_{NR}$$

** Rocchio J.J. (1971) Relevance feedback in information retrieval. In Salton G. (Ed.) The SMART Retrieval System – Experiments in Automatic Document Processing, pp.313-323*

Experimental Results

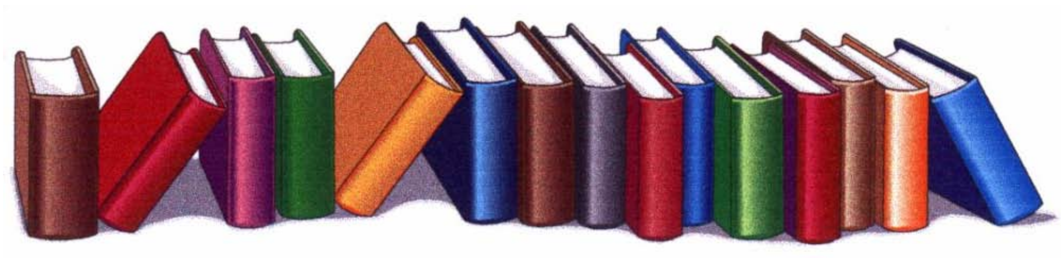


Section 2

Vector Spaces in Monolingual NLP

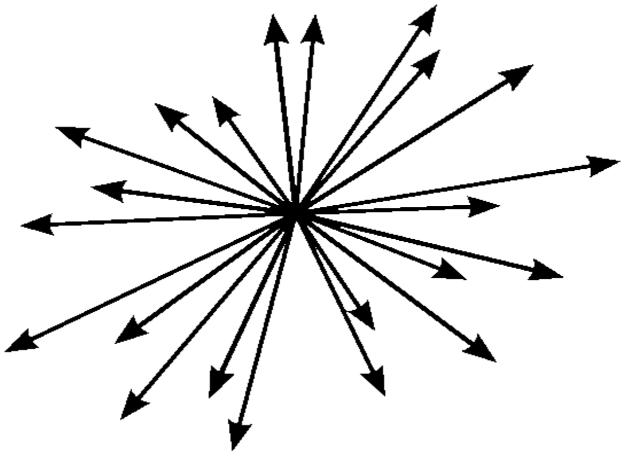
- The Semantic Nature of Vector Spaces
- Information Retrieval and Relevance Ranking
- **Word Spaces and Related Word Identification**
- Semantic Compositionality in Vector Spaces

Latent Semantic Analysis (LSA)



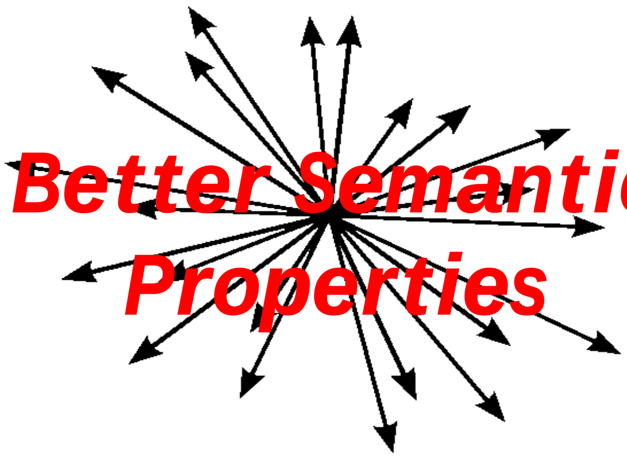
Document collection

TF-IDF Weighting



Vector Space Model

LSA



Reduced-dimensionality Space

Latent Semantic Analysis (LSA)*

SVD: $\boldsymbol{M}_{M\times N} = \boldsymbol{U}_{M\times M}\,\boldsymbol{\Sigma}_{M\times N}\,\boldsymbol{V}^{T}_{N\times N}$

$\boldsymbol{U}^{T}_{M\times M}\,\boldsymbol{M}_{M\times N} = \boldsymbol{D}_{M\times N}$

Documents projected into word space

$\boldsymbol{M}_{M\times N}\,\boldsymbol{V}_{N\times N} = \boldsymbol{W}_{M\times N}$

Words projected into document space

$$\boldsymbol{U}^{T}_{K\times M} = \begin{pmatrix} u_{11} \ldots u_{1k} \ldots u_{m1} \\ u_{21} \ldots u_{2k} \ldots u_{m2} \\ \vdots \quad \vdots \quad \vdots \\ u_{m1} \ldots u_{mk} \ldots u_{mm} \end{pmatrix}^{T}$$

$$\boldsymbol{V}_{N\times K} = \begin{pmatrix} v_{11} \ldots v_{1k} \ldots v_{n1} \\ v_{21} \ldots v_{2k} \ldots v_{n2} \\ \vdots \quad \vdots \quad \vdots \\ v_{n1} \ldots v_{nk} \ldots v_{nn} \end{pmatrix}$$

$\boldsymbol{U}^{T}_{K\times M}\,\boldsymbol{M}_{M\times N} = \boldsymbol{D}_{K\times N}$

Documents projected into reduced word space

$\boldsymbol{M}_{M\times N}\,\boldsymbol{V}_{N\times K} = \boldsymbol{W}_{M\times K}$

Words projected into reduced document space

* 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, pp.391-407

Dataset Under Consideration*

Term definitions from Spanish dictionary used as documents

Collection	Terms	Definitions	Aver. Length
Verbs	4,800	12,414	6.05 words
Adjectives	5,390	8,596	6.05 words
Nouns	20,592	38,689	9.56 words
Others	5,273	9,835	8.01 words
Complete	36,055	69,534	8.32 words

- A document vector space for “verbs” is constructed
- LSA is used to project into a latent semantic space
- MDS is used to create a 2D map for visualization purposes

** Banchs, R.E. (2009), Semantic mapping for related term identification, in Conference on Intelligent Text Processing and Computational Linguistics, CICLing 2009, LNS 5449, pp 111-124*

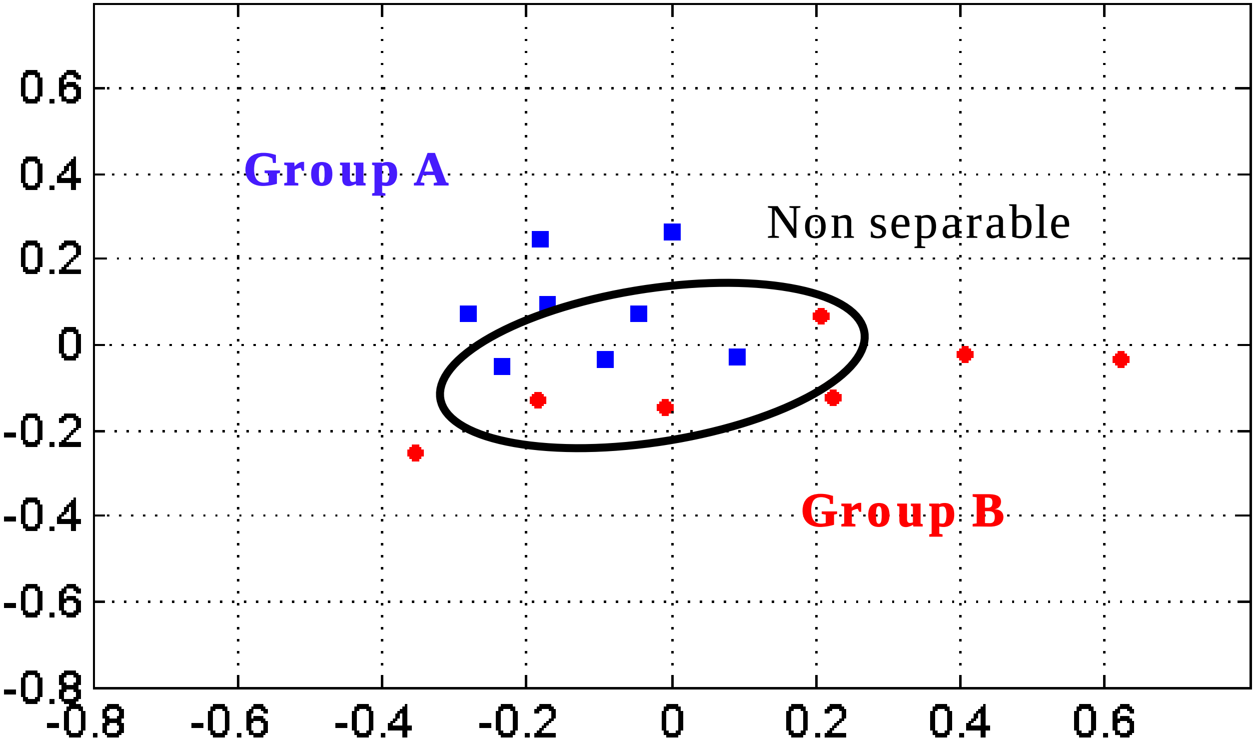
Differentiating Semantic Categories

Two semantic categories of verbs are considered

Group A	Group B
Ayudar (to help)	Agredir (to threaten)
Compartir (to share)	Destruir (to destroy)
Beneficiar (to benefit)	Aniquilar (to eliminate)
Colaborar (to collaborate)	Atacar (to attack)
Salvar (to save)	Arruinar (to ruin)
Apoyar (to support)	Matar (to kill)
Cooperar (to cooperate)	Perjudicar (to perjudice)
Favorecer (to favour)	–

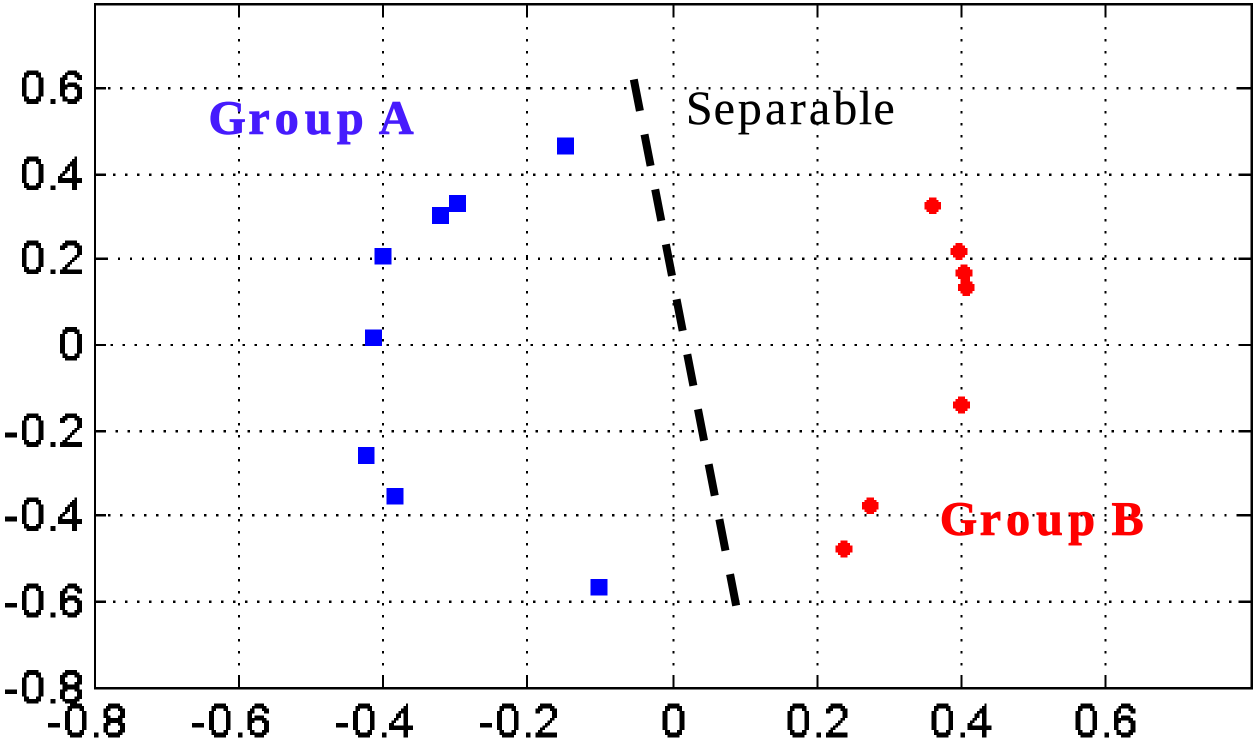
Differentiating Semantic Categories

No LSA applied: original dimensionality maintained



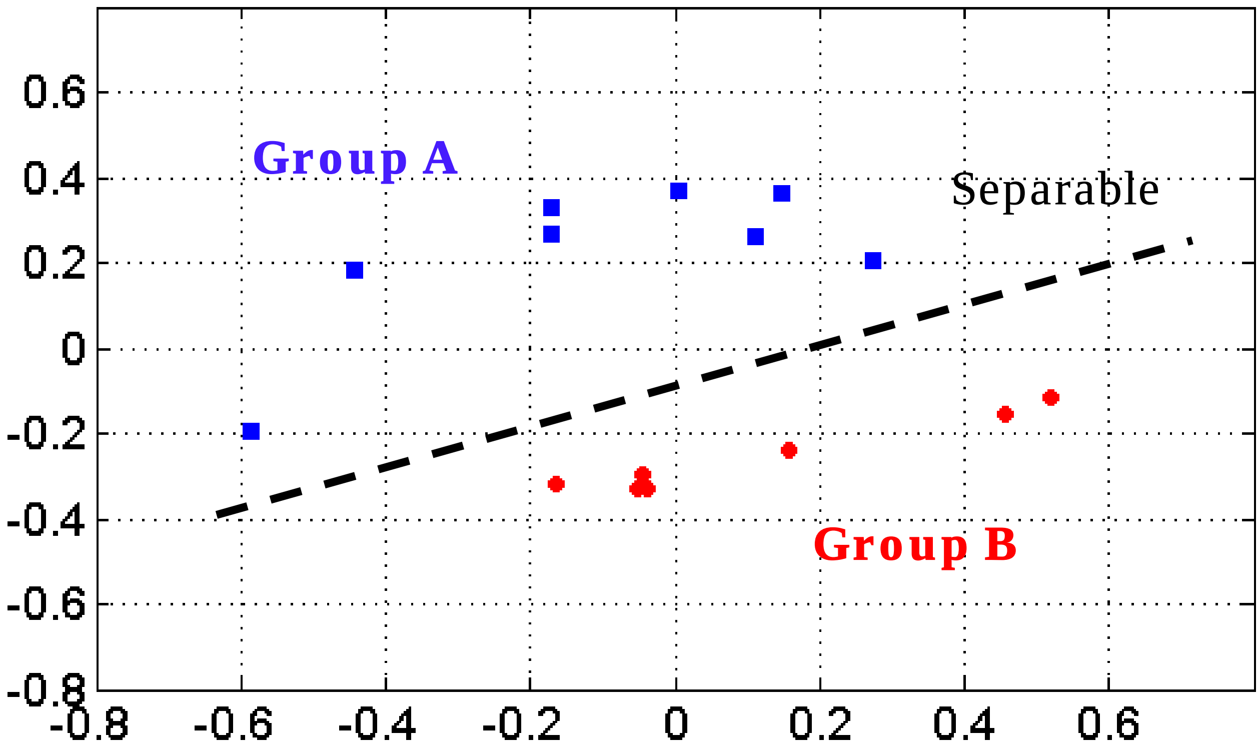
Differentiating Semantic Categories

LSA used to project into latent space of 800 dimensions



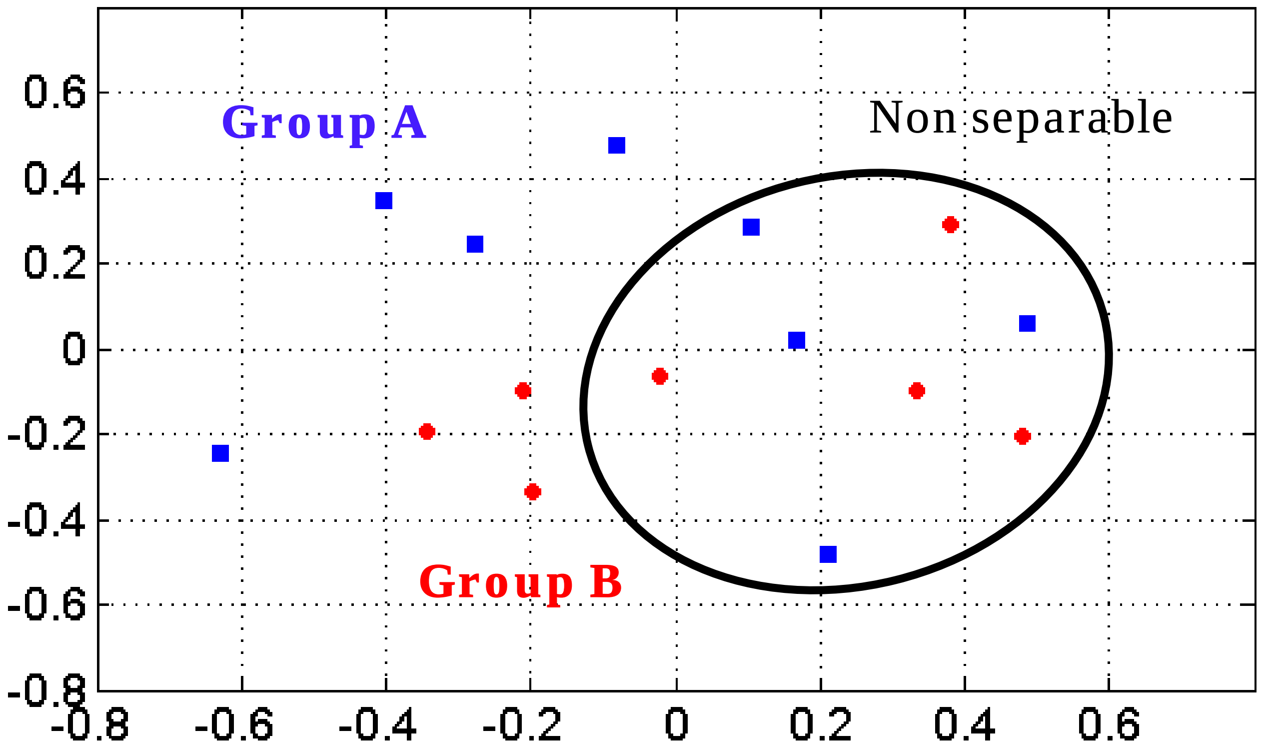
Differentiating Semantic Categories

LSA used to project into latent space of 400 dimensions



Differentiating Semantic Categories

LSA used to project into latent space of 100 dimensions

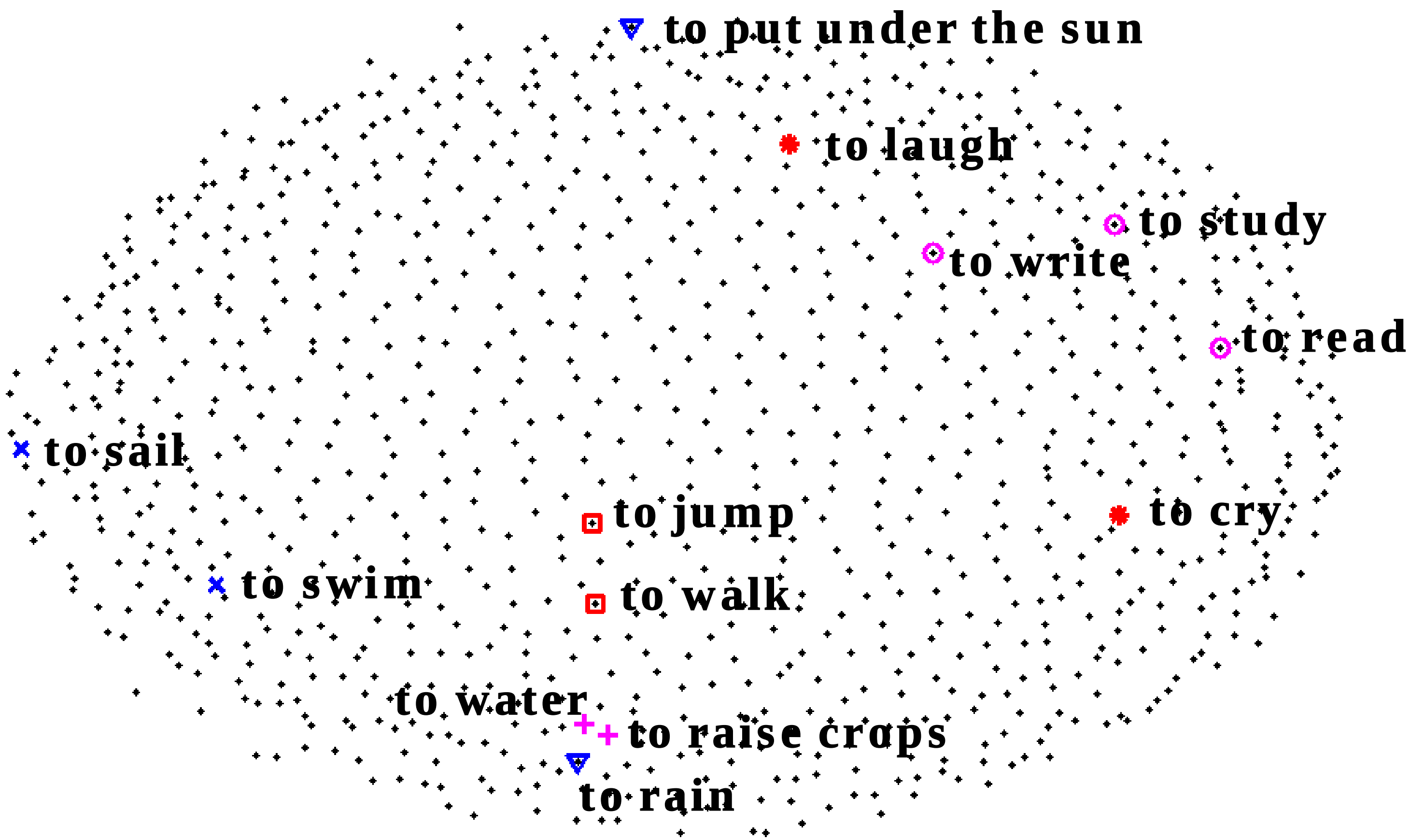


Semantic Similarity of Words

The totality of the 12,414 entries for verbs were considered

- An 800-dimensional latent space representation was generated by applying LSA
- k-means was applied to group the 12,414 entries into 1,000 clusters (minimum size 2, maximum size 36, mean size 12.4, variance 4.7)
- Finally, non-linear dimensionality reduction (MDS) was applied to generate a map

Semantic Similarity of Words



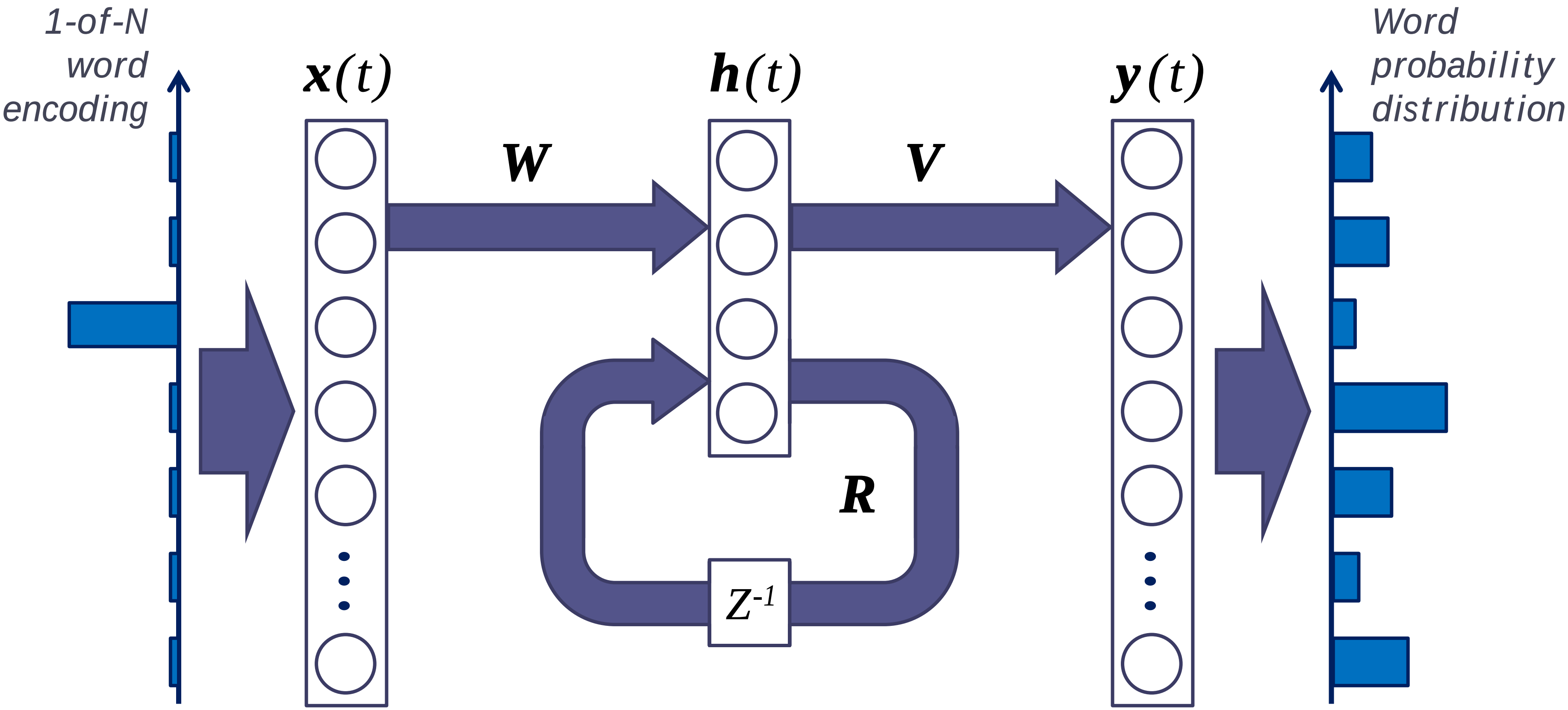
Regularities in Vector Spaces*

Recurrent Neural Network Language Model

- After study internal word representations generated by the model
- Syntactic and semantic regularities were discovered to be mapped into the form of constant vector offsets

** Mikolov T., Yih W.T. and Zweig G. (2013), Linguistic Regularities in Continuous Space Word Representations, NAACL-HLT 2013*

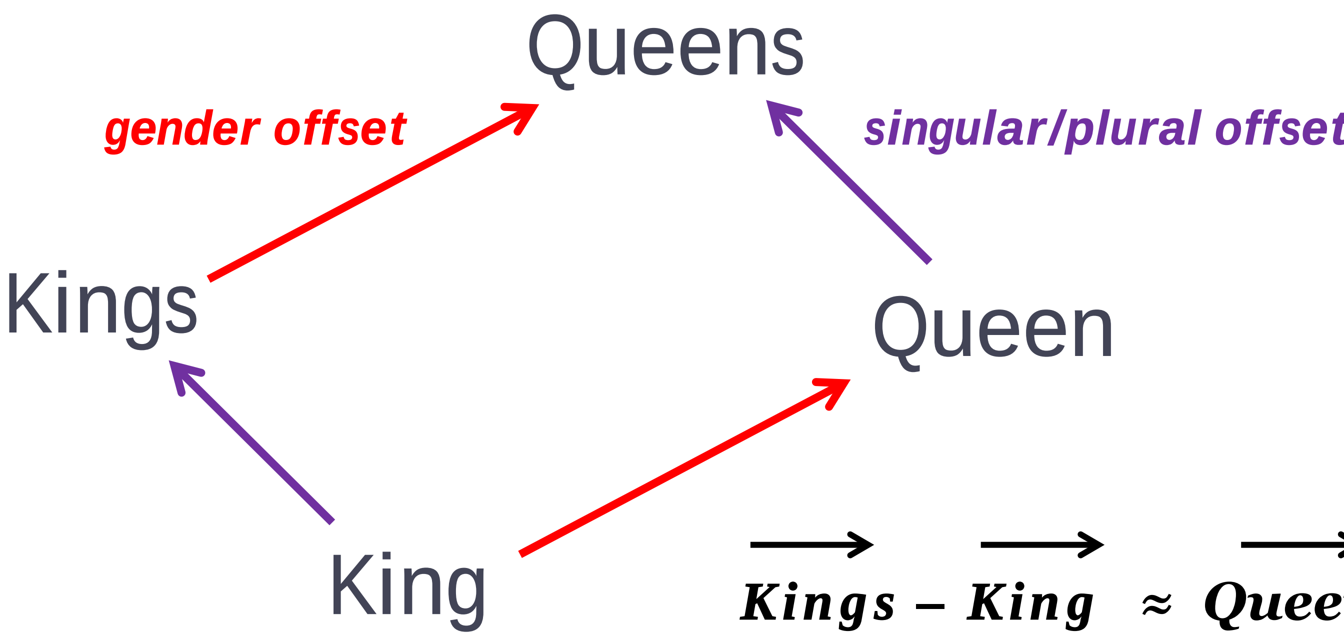
Recurrent Neural Network (RNN)



$$\boldsymbol{h}(t) = Sigmoid(\boldsymbol{W}\ \boldsymbol{x}(t) + \boldsymbol{R}\ \boldsymbol{h}(t\text{-}1))$$

$$\boldsymbol{y}(t) = Softmax(\boldsymbol{V}\ \boldsymbol{h}(t))$$

Regularities as Vector Offsets



$$\overrightarrow{Kings} - \overrightarrow{King} \approx \overrightarrow{Queens} - \overrightarrow{Queen}$$

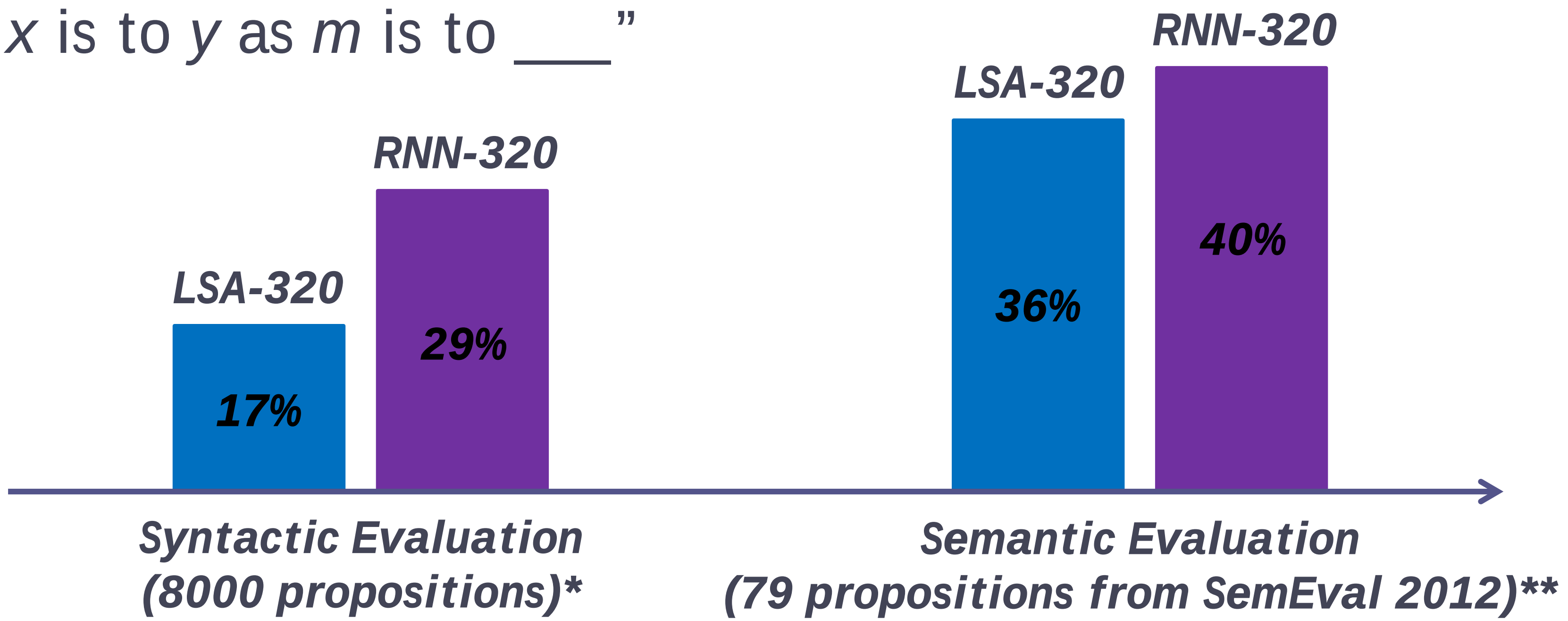
$$\Rightarrow \overrightarrow{Queens} \approx \overrightarrow{Kings} - \overrightarrow{King} + \overrightarrow{Queen}$$

Image taken from Mikolov T., Yih W.T. and Zweig G. (2013), Linguistic Regularities in Continuous Space Word Representations, NAACL-HLT 2013

Comparative Evaluations*

Propositions formulated as analogy questions:

"*x* is to *y* as *m* is to ___"



* *Mikolov T., Yih W.T. and Zweig G. (2013), Linguistic Regularities in Continuous Space Word Representations, NAACL-HLT 2013*

** *Jurgens D., Mohammad S., Turney P. and Holyoak K. (2012), Semeval-2012 task: Measuring degrees of relational similarity, in SemEval 2012, pp. 356-364*

Section 2

Vector Spaces in Monolingual NLP

- The Semantic Nature of Vector Spaces
- Information Retrieval and Relevance Ranking
- Word Spaces and Related Word Identification
- **Semantic Compositionality in Vector Spaces**

Semantic Compositionality

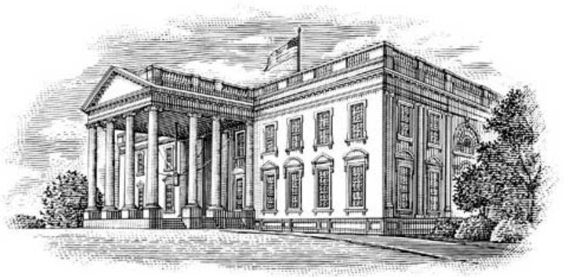
- The ***principle of compositionality*** states that the meaning of a complex expression depends on:
  - the meaning of its constituent expressions
  - the rules used to combine them
- Some idiomatic expressions and named entities constitute typical exceptions to the principle of compositionality in natural language

Compositionality and Exceptions

Consider the adjective-noun constructions

RED CAR

WHITE HOUSE



???

Compositionality in Vector Space

- *Can this principle be modeled in Vector Space representations of language*?

- Two Basic mechanisms can be used to model compositionality in the vector space model framework*
  - Intersection of properties (multiplicative approach)
  - Combination of properties (additive approach)

** Mitchell J. and Lapata M. (2008), Vector-based Models of Semantic Composition, in Proceedings of ACL-HLT 2008, pp. 236-244*

Compositionality Models

- Given two word vector representations $\boldsymbol{x}$ and $\boldsymbol{y}$
- A composition vector $\boldsymbol{z}$ can be computed as:

Multiplicative Models

Tensor product

Linear combination

Additive Models

$$\boldsymbol{z} = \boldsymbol{C}\,\boldsymbol{x}\,\boldsymbol{y}$$

$$\boldsymbol{z} = \boldsymbol{A}\,\boldsymbol{x} + \boldsymbol{B}\,\boldsymbol{y}$$

$$z_i = \sum_j x_j\, y_{i-j}$$

Circular convolution

$$z_i = x_i + y_i$$

Simple additive

$$z_i = x_i\, y_i$$

Simple multiplicative

$$z_i = \alpha\, x_i + \beta\, y_i$$

Weighted additive

$$z_i = \alpha\, x_i + \beta\, y_i + \gamma\, x_i\, y_i$$

Combined model

Additive Compositionality*

- Use unigram and bigram counts to identify phrases
- Uses Skip-gram model to compute word representations
- Compute element-wise additions of word vectors to retrieve associated words:
  - Czech + currency → koruna, Check crown, ...
  - German + airline → airline Lufthansa, Lufthansa, ...
  - Russian + river → Moscow, Volga River, ...

** Mikolov T., Sutskever I., Chen K., Corrado G. and Dean J. (2013), Distributed Representations of Words and Phrases and their Compositionality, arXiv:1310.4546v1*

Adjectives as Linear Maps*

- An adjective-noun composition vector is: $\boldsymbol{z} = \boldsymbol{A}\ \boldsymbol{n}$
- The rows of $\boldsymbol{A}$ are estimated by linear regressions
- Some examples of predicted nearest neighbors:
  - general question ➡ general issue
  - recent request ➡ recent enquiry
  - current dimension ➡ current element
  - special something ➡ special thing

** Baroni M. and Zamparelli R. (2010), Nous are vectors, adjectives are matrices: Representing adjective-noun constructions in semantic space, in EMNLP 2010*

Section 2

Main references for this section

- G. Salton, A. Wong and C. S. Yang, 1975, “A Vector Space for Automatic Indexing”
- R. E. Banchs, 2013, “Text Mining with MATLAB”
- R. E. Banchs, 2009, “Semantic mapping for related term identification”
- T. Mikolov, W. T. Yih and G. Zweig, 2013, “Linguistic Regularities in Continuous Space Word Representations”
- J. Mitchell and M. Lapata, 2008, “Vector-based Models of Semantic Composition”

Section 2

Additional references for this section

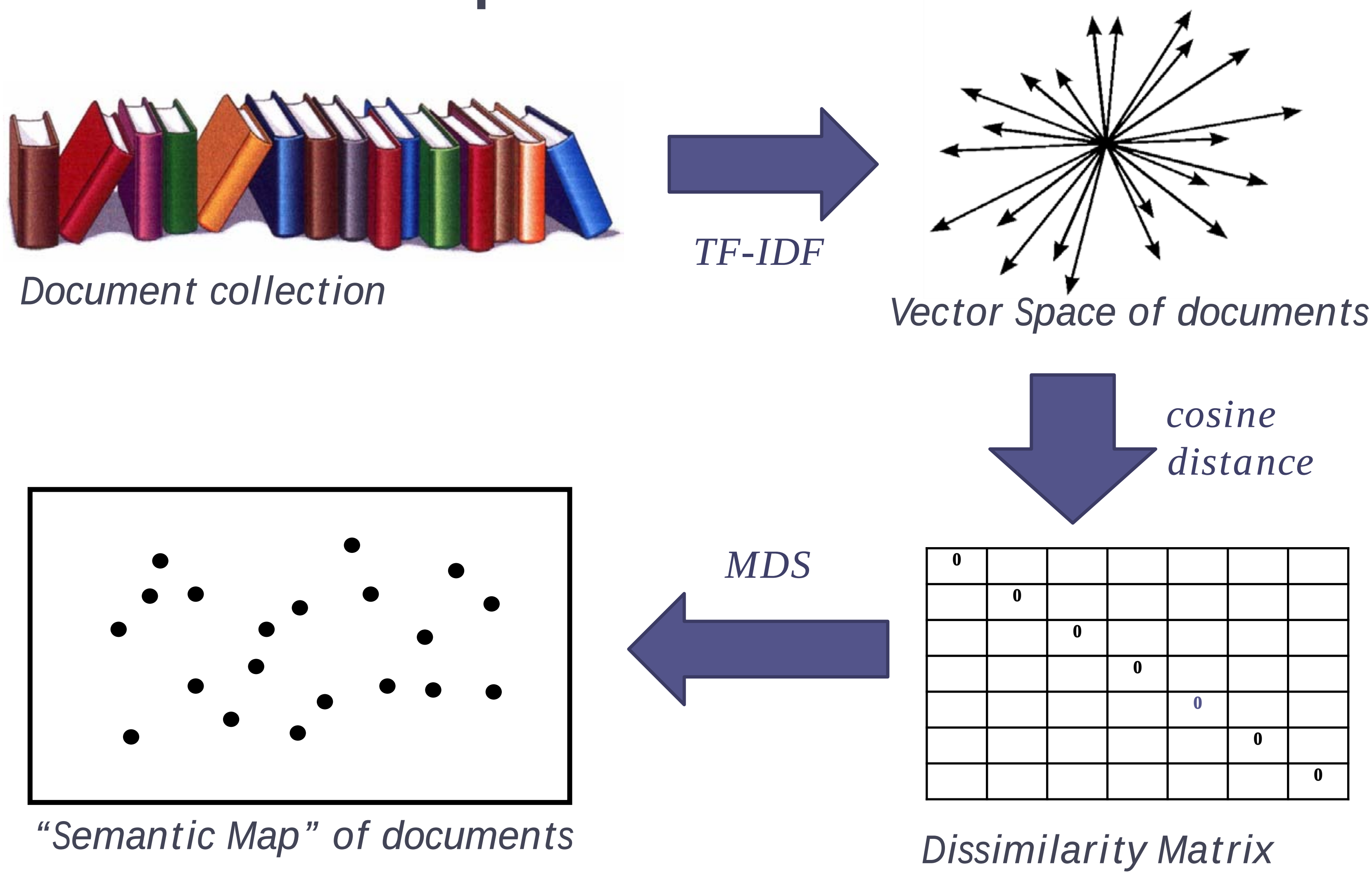
- Lee, W.C. and Fox, E.A. (1988) Experimental comparison of schemes for interpreting Boolean queries. Technical Report TR-88-27, Computer Science, Virginia Polytechnic Institute and State University
- Rocchio J.J. (1971) Relevance feedback in information retrieval. In Salton G. (Ed.) The SMART Retrieval System – Experiments in Automatic Document Processing, pp.313-323
- 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, pp.391-407
- Jurgens D., Mohammad S., Turney P. and Holyoak K. (2012), Semeval-2012 task: Measuring degrees of relational similarity, in SemEval 2012, pp. 356-364
- Mikolov T., Sutskever I., Chen K., Corrado G. and Dean J. (2013), Distributed Representations of Words and Phrases and their Compositionality, arXiv:1310.4546v1
- Baroni M. and Zamparelli R. (2010), Nous are vectors, adjectives are matrices: Representing adjective-noun constructions in semantic space, in EMNLP 2010

Section 3

Vector Spaces in Cross-language NLP

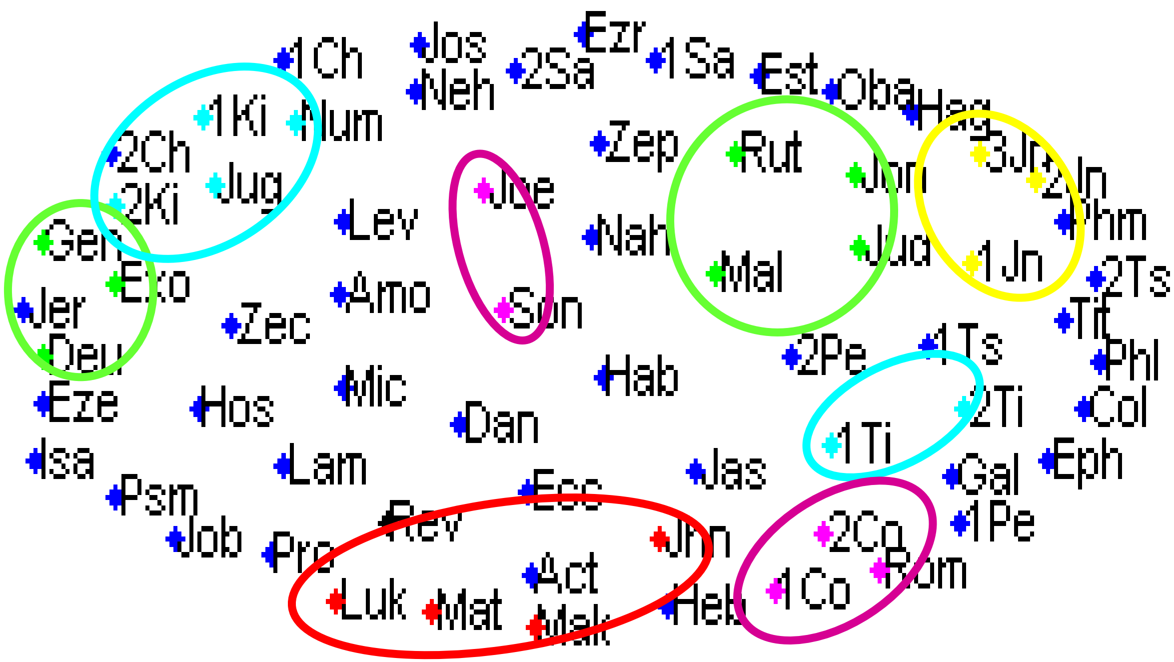
- **Semantic Map Similarities Across Languages**
- Cross-language Information Retrieval in Vector Spaces
- Cross-script Information Retrieval and Transliteration
- Cross-language Sentence Matching and its Applications
- Semantic Context Modelling for Machine Translation
- Bilingual Dictionary and Translation-table Generation
- Evaluating Machine Translation in Vector Space

Semantic Maps Revisited



Multilingual Document Collection

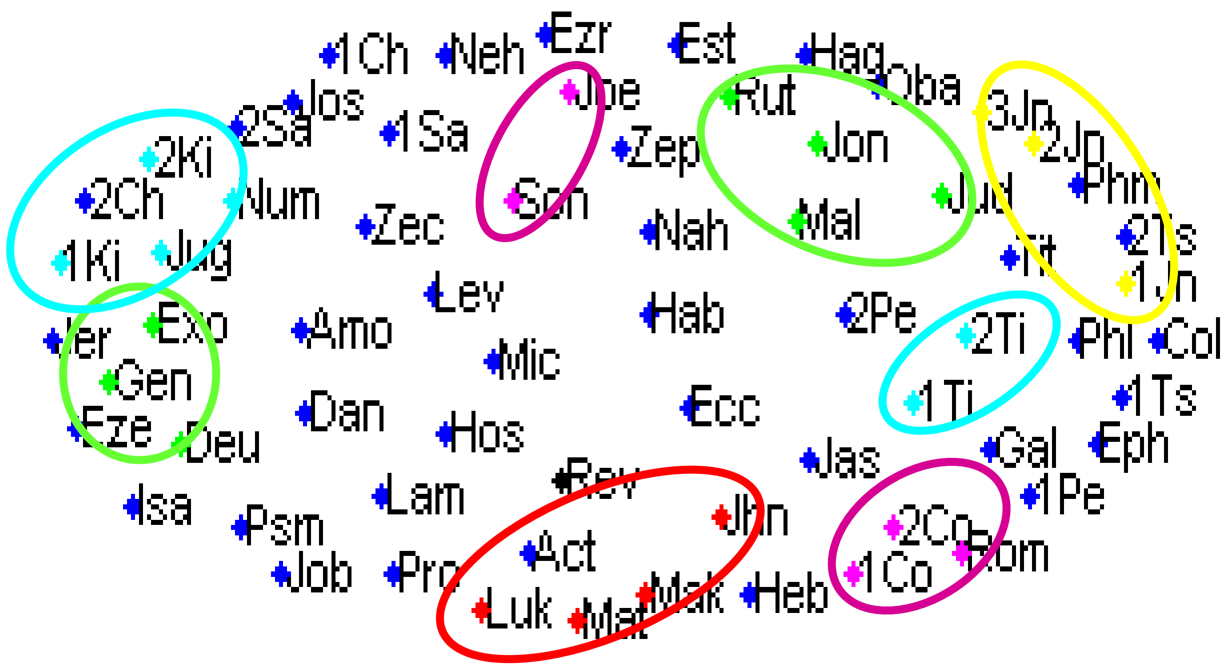
66 Books from The Holy Bible: English version



(vocabulary size: 8121 words)

Multilingual Document Collection

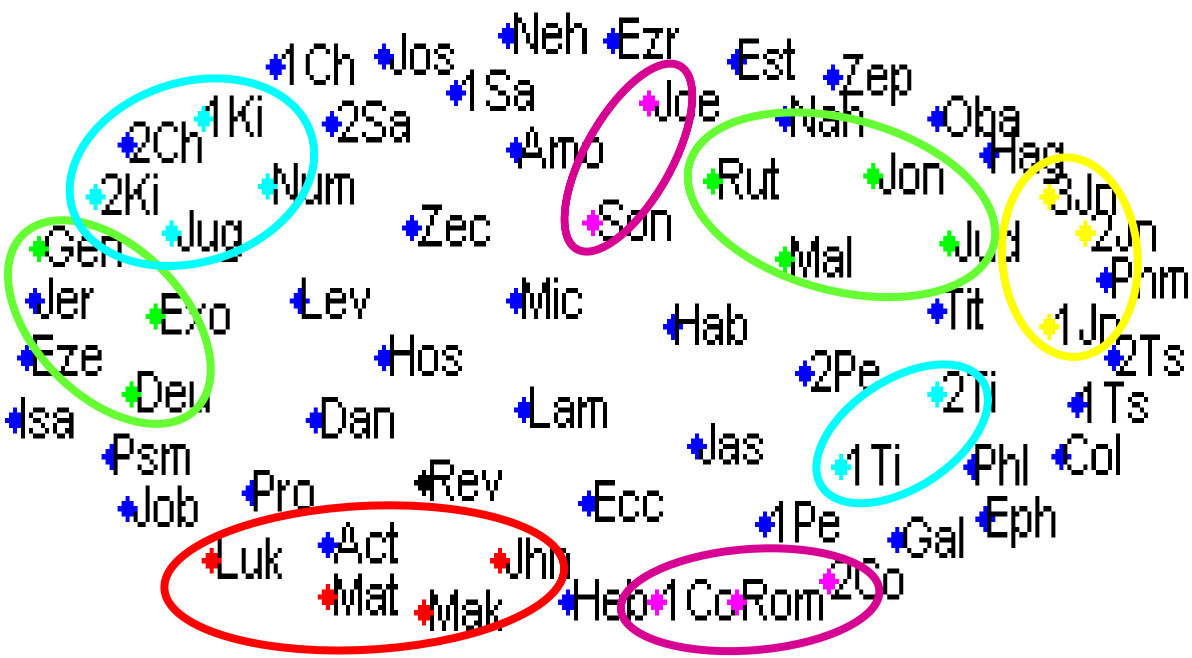
66 Books from The Holy Bible: Chinese version



(vocabulary size: 12952 words)

Multilingual Document Collection

66 Books from The Holy Bible: Spanish version



(vocabulary size: 25385 words)

Cross-language Similarities

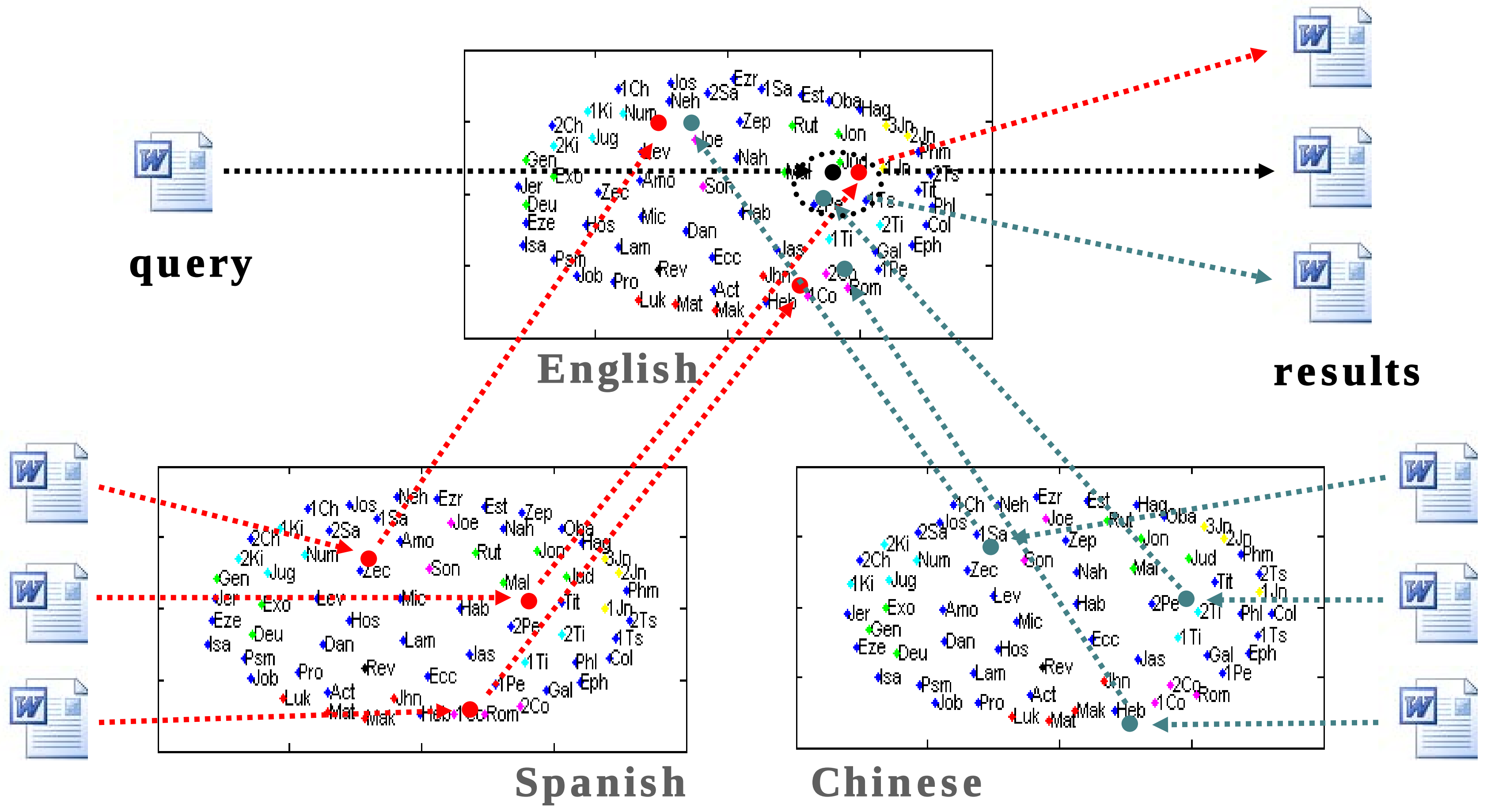
- Each language map has been obtained independently from each other language (monolingual context)
- The similarities among the maps are remarkable
- *Could we exploit these similarities for performing cross-language information retrieval tasks?*

Section 3

Vector Spaces in Cross-language NLP

- Semantic Map Similarities Across Languages
- **Cross-language Information Retrieval in Vector Spaces**
- Cross-script Information Retrieval and Transliteration
- Cross-language Sentence Matching and its Applications
- Semantic Context Modelling for Machine Translation
- Bilingual Dictionary and Translation-table Generation
- Evaluating Machine Translation in Vector Space

Semantic Maps for CLIR

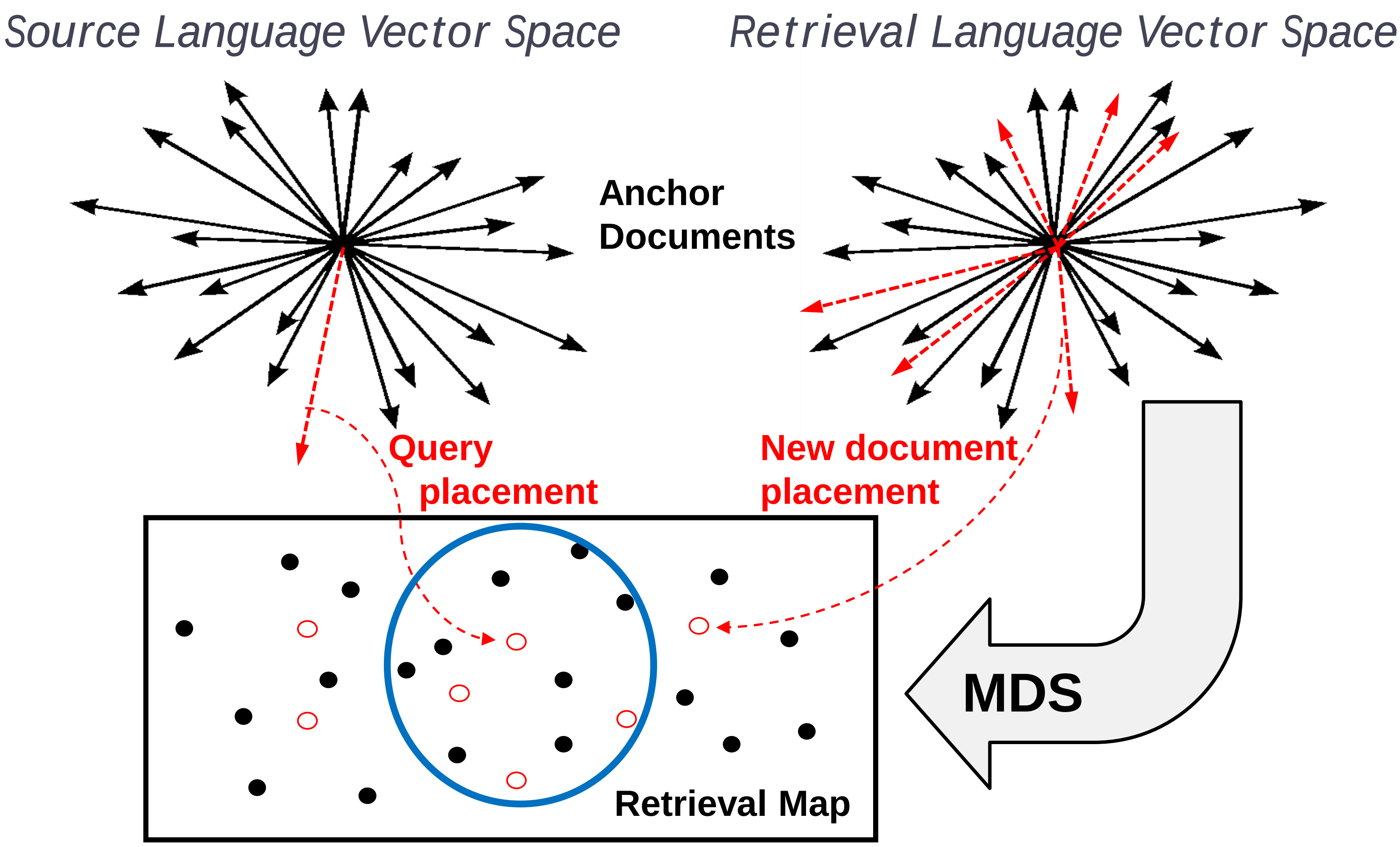


CLIR by Using MDS Projections*

- Start from a multilingual collection of "anchor documents" and construct the retrieval map
- Project new documents and queries from any source language into the retrieval language map
- Retrieve documents over retrieval language map by using a distance metric

** Banchs R.E. and Kaltenbrunner A. (2008), Exploring MDS projections for cross-language information retrieval, in Proceedings of the 31st Annual International ACM SIGIR 2008*

CLIR by Using MDS Projections



Computing a Projection Matrix

A linear transformation from the original high dimensional space into the lower dimensionality map can be inferred from anchor documents

Coordinates of anchor documents in the projected space (KxN)

$$M = T\,D$$

Distances among anchor documents in the original space (NxN)

Transformation Matrix (KxN)

$$\boxed{T = M\,D^{-1}}$$

Projecting Documents and Queries

A probe document or query can be placed into the retrieval map by using the transformation matrix

Coordinates of probe document (or query) in the projected space of retrieval language

$$m = T\,d$$

Transformation Matrix (KxN)

Distances between probe document (or query) and anchor documents in the original language space

Computing a Projection Matrix

Two different variants of the linear projection matrix $\boldsymbol{T}$ can be computed:

- A monolingual projection matrix:*
  - $\boldsymbol{M}$ and $\boldsymbol{D}$ are computed on the retrieval language
- A cross-language projection matrix:**
  - $\boldsymbol{M}$ is computed on the retrieval language, and
  - $\boldsymbol{D}$ is computed on the source language

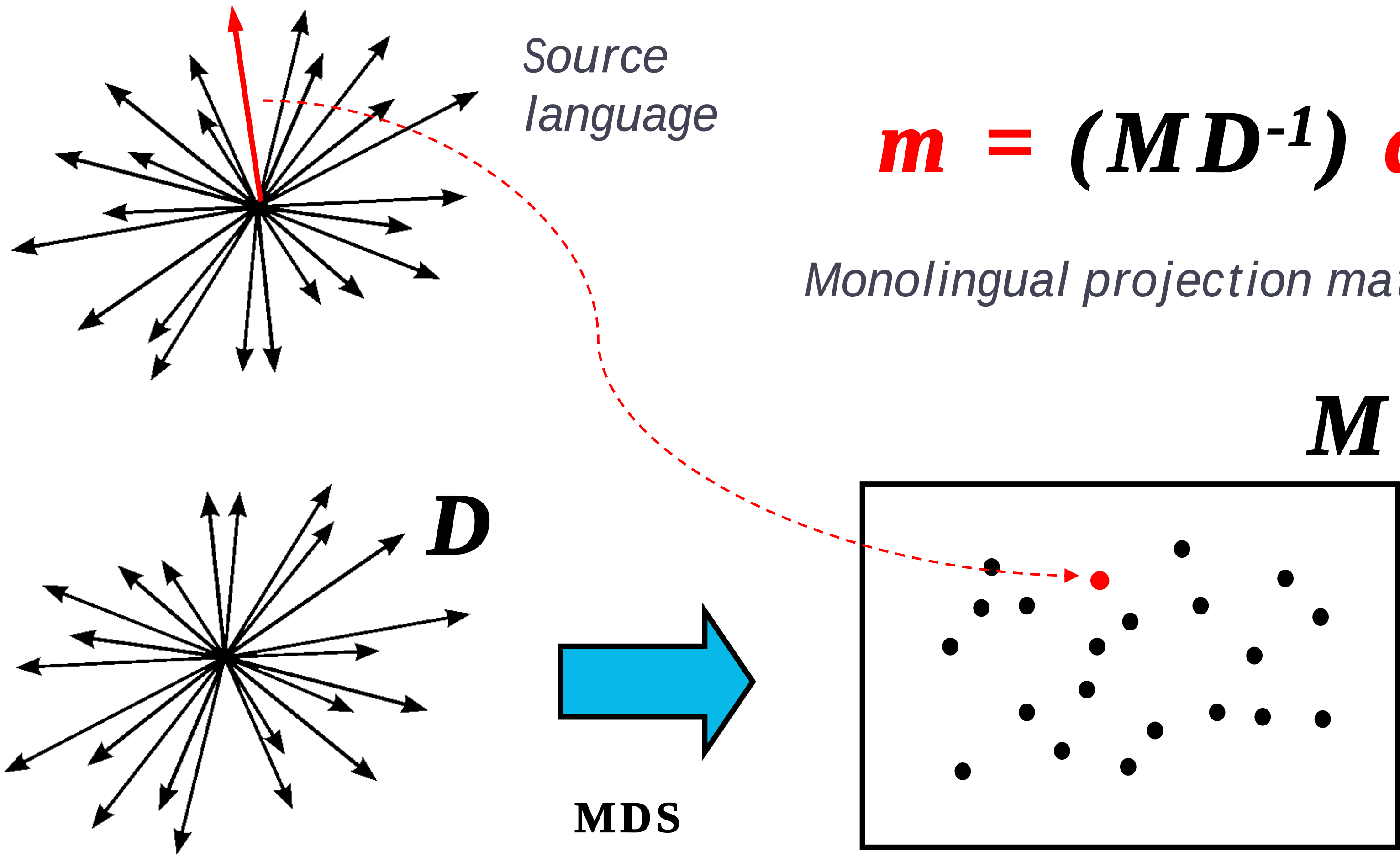
** Banchs R.E. and Kaltenbrunner A. (2008), Exploring MDS projections for cross-language information retrieval, in Proceedings of the 31st Annual International ACM SIGIR 2008*

*** Banchs R.E. and Costa-jussà M.R. (2013), Cross-Language Document Retrieval by using Nonlinear Semantic Mapping, International Journal of Applied Artificial Intelligence, 27(9), pp. 781-802*

Monolingual Projection Method

$$m = (MD^{-1})\ d$$

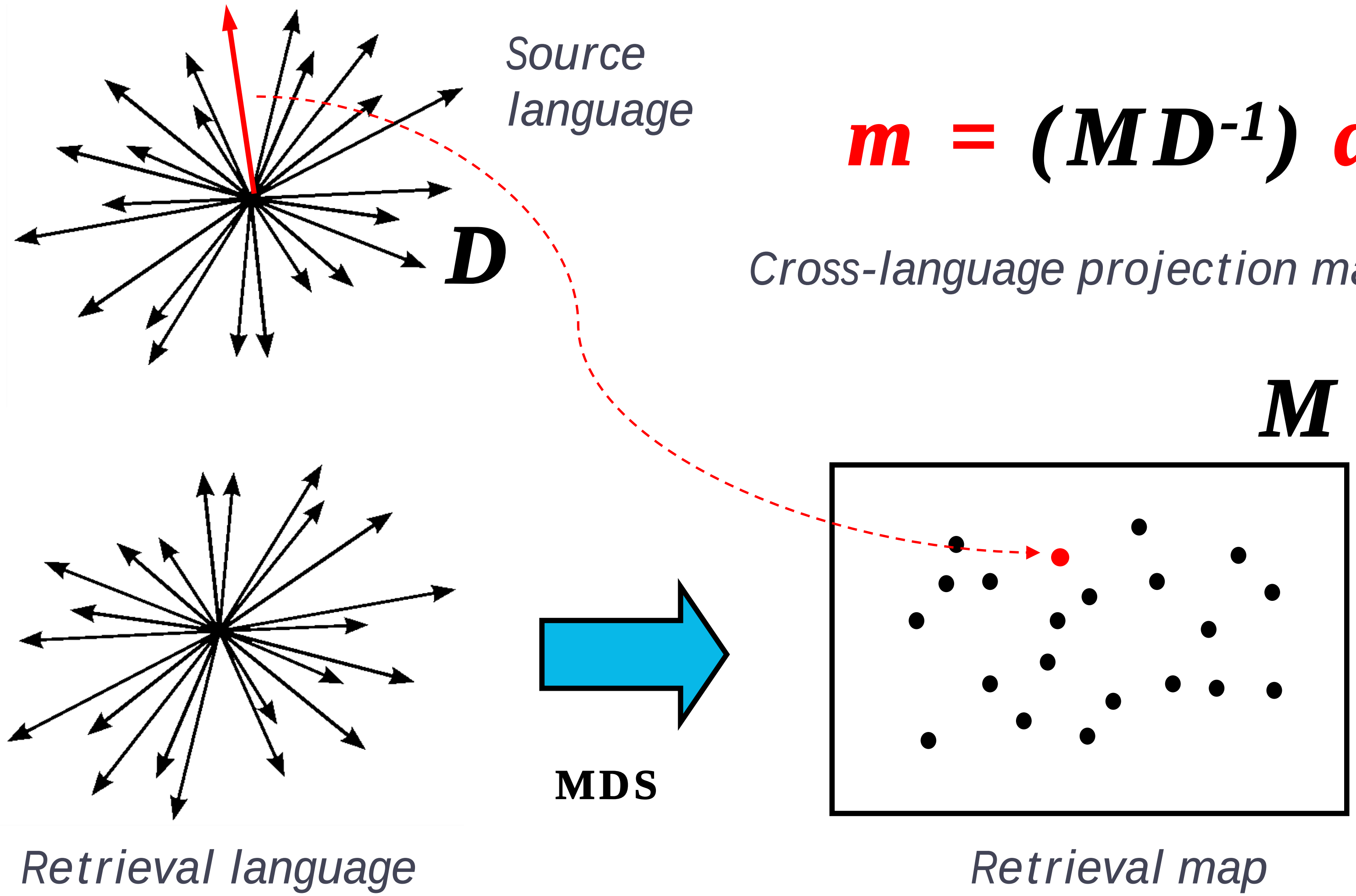
Monolingual projection matrix



Retrieval language

Retrieval map

Cross-language Projection Method



$$\mathbf{m} = (\mathbf{M}\mathbf{D}^{-1})\,\mathbf{d}$$

Cross-language projection matrix

CLIR by Using Cross-language LSI*

- In monolingual LSI, the term-document matrix is decomposed into a set of *K* orthogonal factors by means of Singular Value Decomposition (SVD)
- In cross-language LSI, a **multilingual term-document matrix** is constructed from a multilingual parallel collection and LSI is applied by considering multilingual “extended” representations of query and documents

** Dumais S.T., Letsche T.A., Littman M.L. and Landauer T.K. (1997), Automatic Cross-Language Retrieval Using Latent Semantic Indexing, in AAAI-97 Spring Symposium Series: Cross-Language Text and Speech Retrieval, pp. 18-24*

The Cross-language LSI Method

Multilingual term-document matrix

$$X = \begin{pmatrix} X_a \\ X_b \end{pmatrix}$$

Term-document matrix in language A

Term-document matrix in language B

SVD: $\boldsymbol{X} = \boldsymbol{U}\,\boldsymbol{\Sigma}\,\boldsymbol{V}^T$

Retrieval is based on internal product of the form: $\langle \boldsymbol{U}^T \boldsymbol{d}, \boldsymbol{U}^T \boldsymbol{q} \rangle$

document in language A

query in language A

With: $\boldsymbol{d} = \begin{pmatrix} d_a \\ 0 \end{pmatrix}$ or $\begin{pmatrix} 0 \\ d_b \end{pmatrix}$ $\quad \boldsymbol{q} = \begin{pmatrix} q_a \\ 0 \end{pmatrix}$ or $\begin{pmatrix} 0 \\ q_b \end{pmatrix}$

document in language B

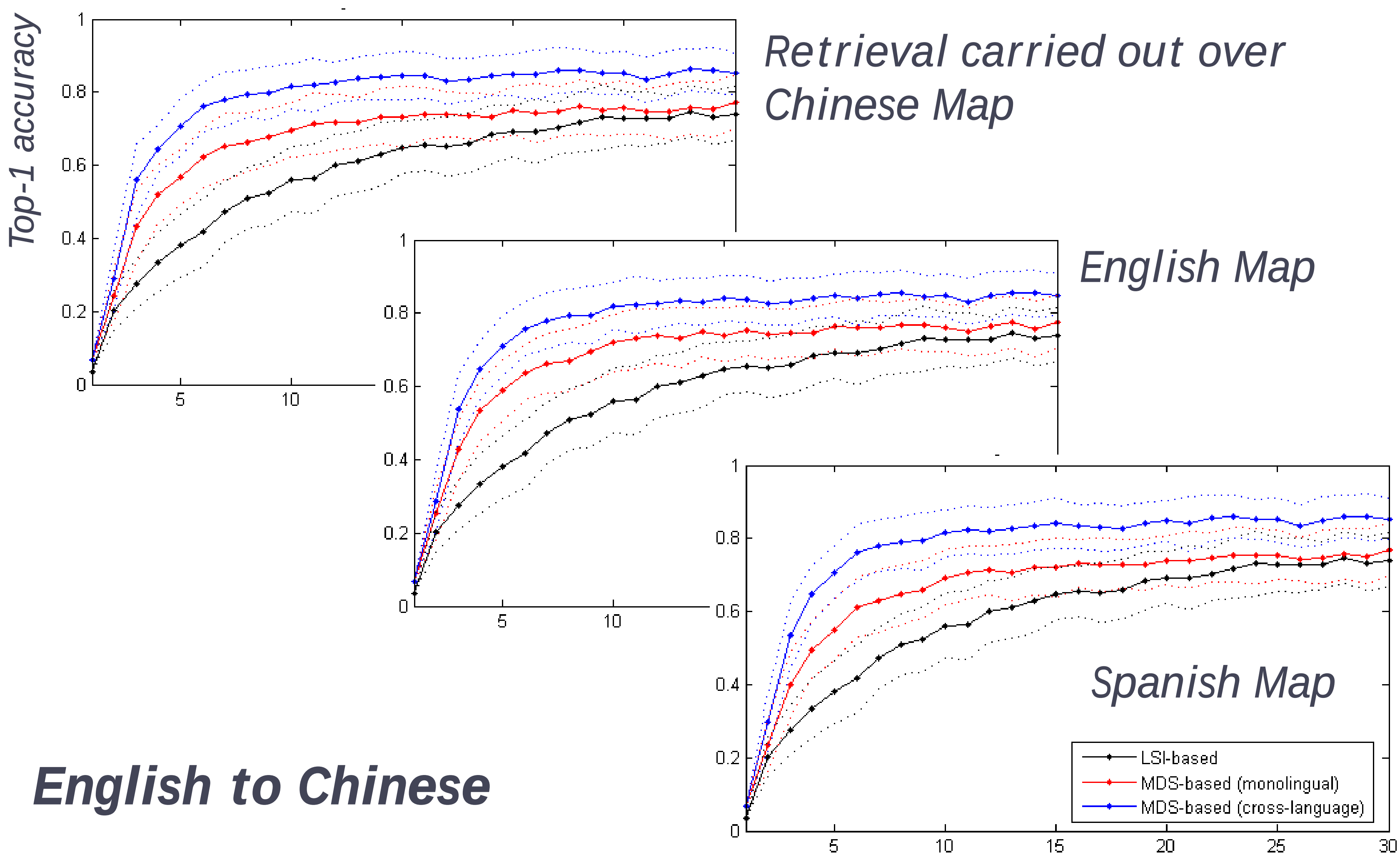
query in language B

Comparative Evaluations

We performed a comparative evaluation of the three methods described over the trilingual dataset:

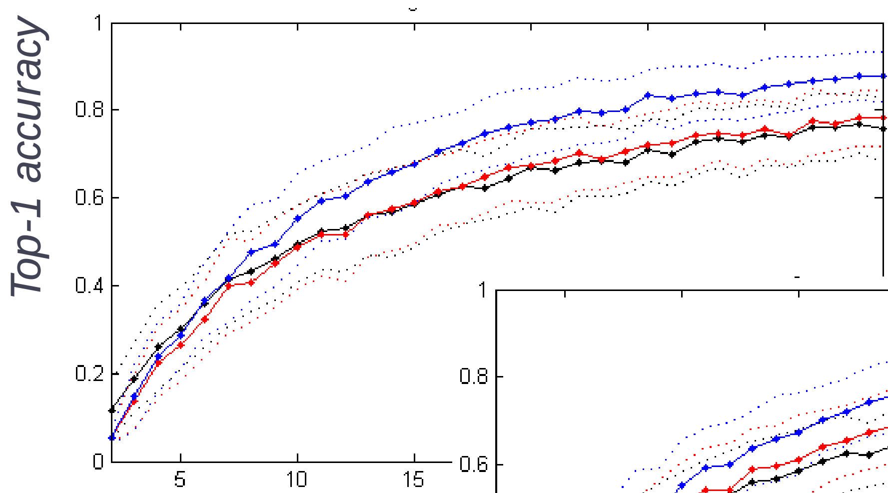
- Task 1: Retrieve a book using the same book in a different language as query:
  - Subtask 1.A: Dimensionality of the retrieval space is varied
  - Subtask 1.B: Anchor document set size is varied
- Task 2: Retrieve a chapter using the same chapter in a different language as a query

Task 1.A: Dimensionality of Space

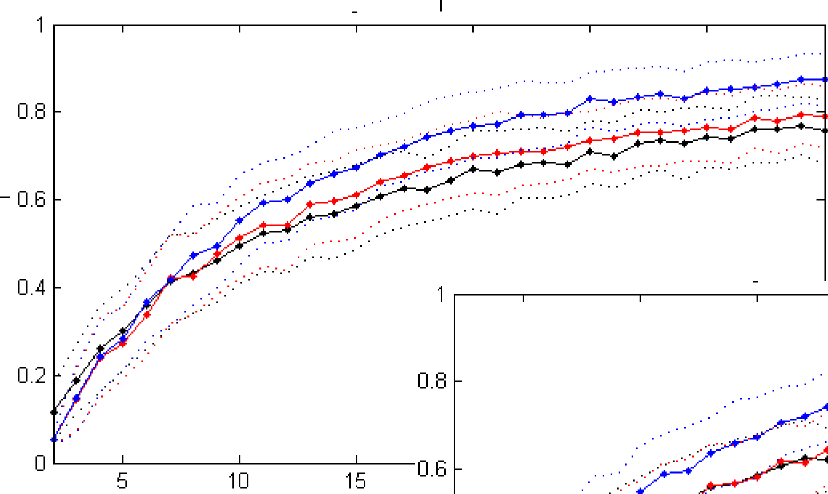


English to Chinese

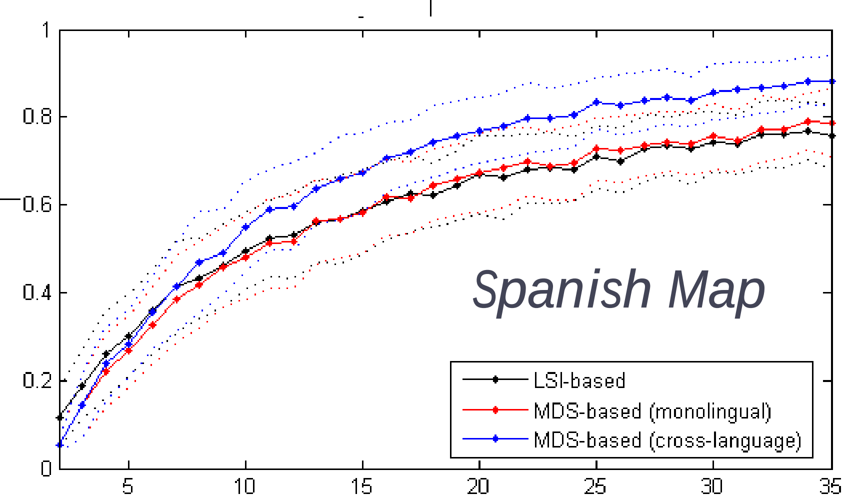
Task 1.B: Anchor Document Set



Retrieval carried out over Chinese Map



English Map

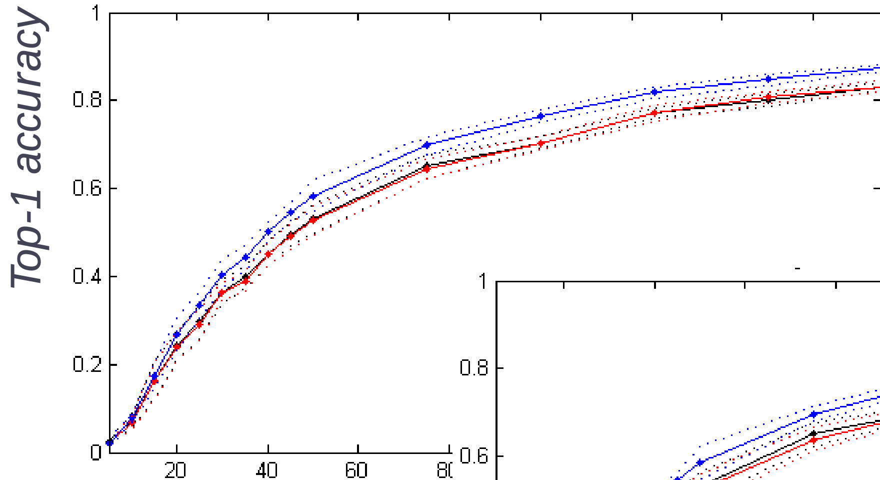


Spanish Map

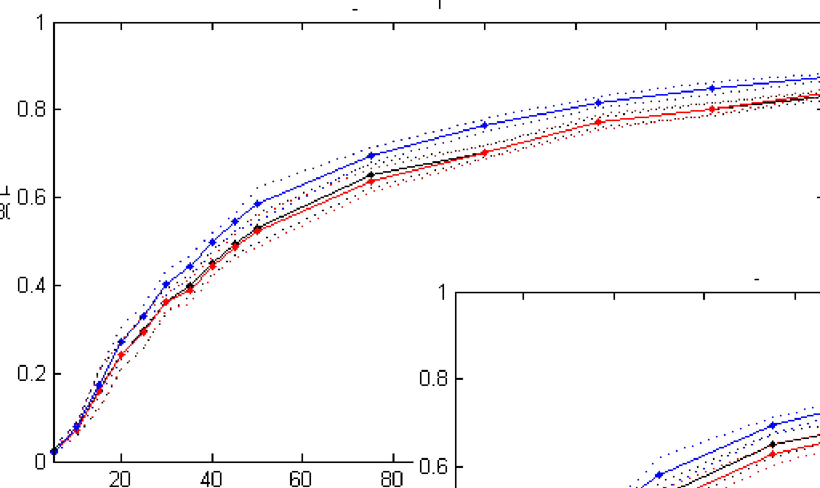
English to Chinese

(dimensionality of retrieval space is equal to anchor set size)

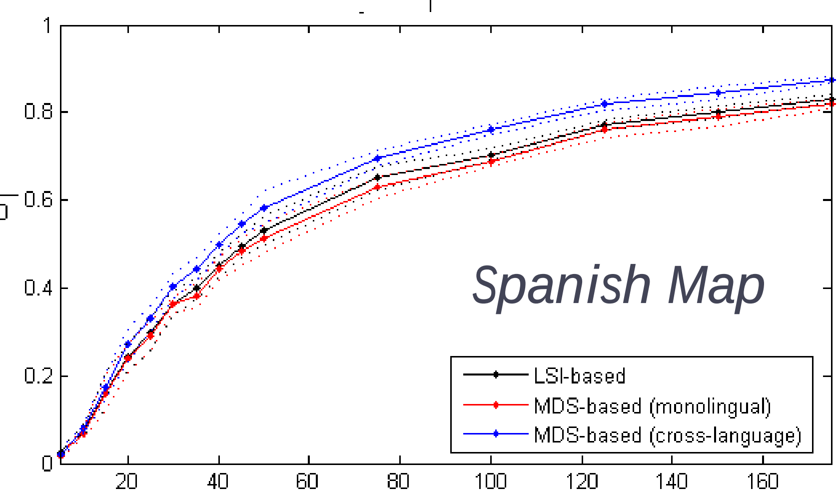
Task 2: Chapter Retrieval



Retrieval carried out over Chinese Map



English Map



English to Chinese

(dimensionality of retrieval space is equal to anchor set size)

Some Conclusions*

- Semantic maps, and more specifically MDS projections, can be exploited for CLIR tasks
- The cross-language projection matrix variant performs better than the monolingual projection matrix variant
- MDS maps perform better than LSI for the considered CLIR tasks

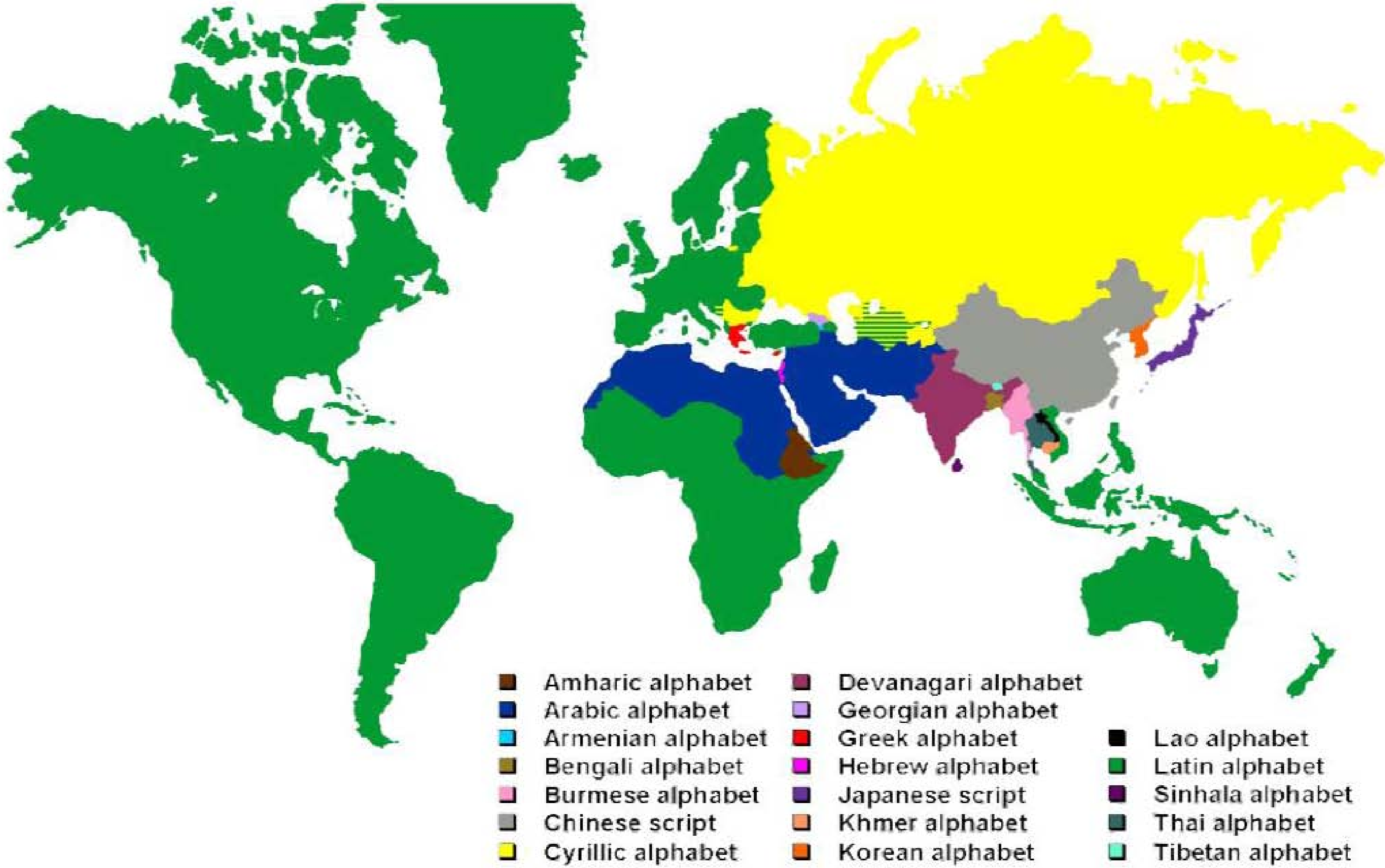
** Banchs R.E. and Costa-jussà M.R. (2013), Cross-Language Document Retrieval by using Nonlinear Semantic Mapping, International Journal of Applied Artificial Intelligence, 27(9), pp. 781-802*

Section 3

Vector Spaces in Cross-language NLP

- Semantic Map Similarities Across Languages
- Cross-language Information Retrieval in Vector Spaces
- **Cross-script Information Retrieval and Transliteration**
- Cross-language Sentence Matching and its Applications
- Semantic Context Modelling for Machine Translation
- Bilingual Dictionary and Translation-table Generation
- Evaluating Machine Translation in Vector Space

Main Scripts used Around the World



Transliteration and Romanization

- The process of phonetically representing the words of one language in a non-native script
- Due to socio-cultural and technical reasons, most languages using non Latin native scripts commonly implement Latin script writing rules: “Romanization”

你好 ➡ nǐ hǎo

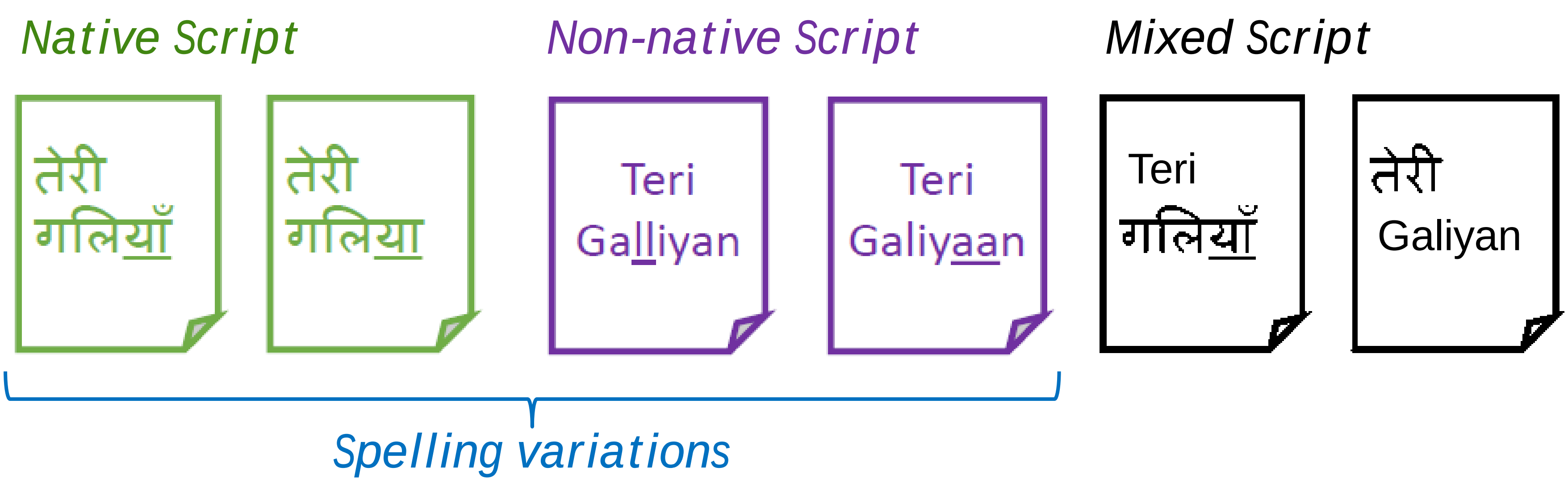
The Multi-Script IR (MSIR) Problem*

- There are many languages that use non Latin scripts (Japanese, Chinese, Arabic, Hindi, etc.)
- There is a lot of text for these languages in the Web that is represented into the Latin script
- For some of these languages, no standard rules exist for transliteration

** Gupta P., Bali K., Banchs R.E. Choudhury M. and Rosso P. (2014), Query Expansion for Multi-script Information Retrieval, in Proceedings of the 37st Annual International ACM SIGIR 2014*

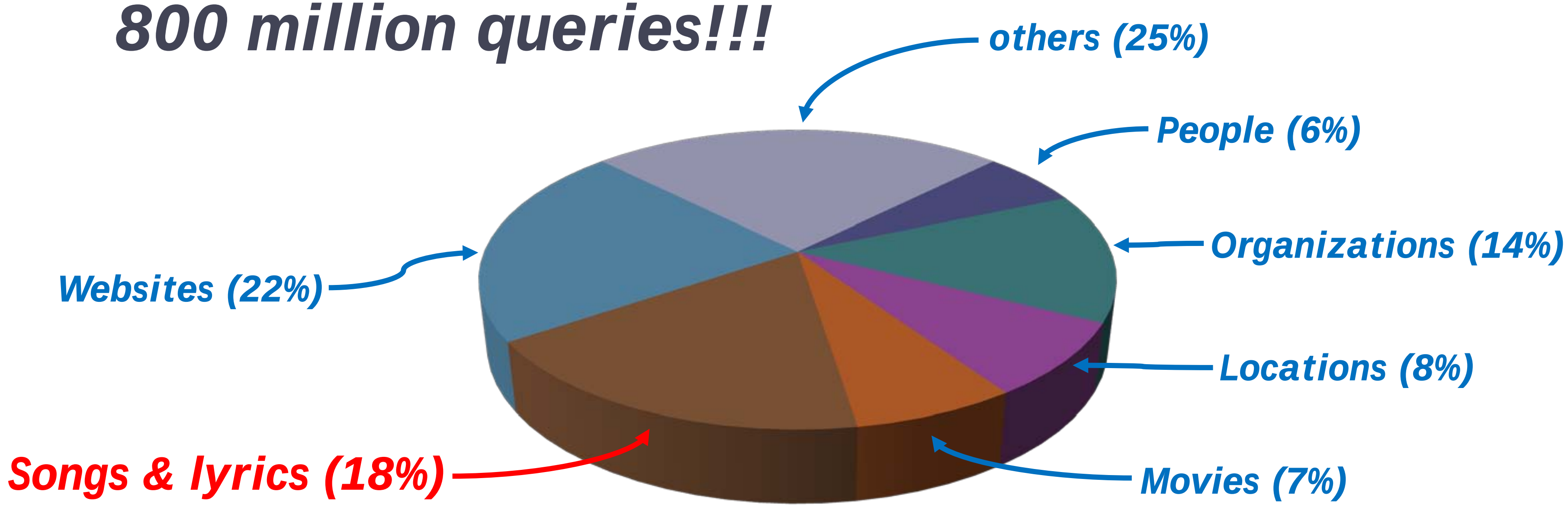
The Main Challenge of MSIR

- Mixed script queries and documents
- Extensive spelling variations



Significance of MSIR

- Only 6% of the queries issued in India to Bing contain Hindi words in Latin script
- From a total number of 13.78 billion queries!!! ***800 million queries!!!***

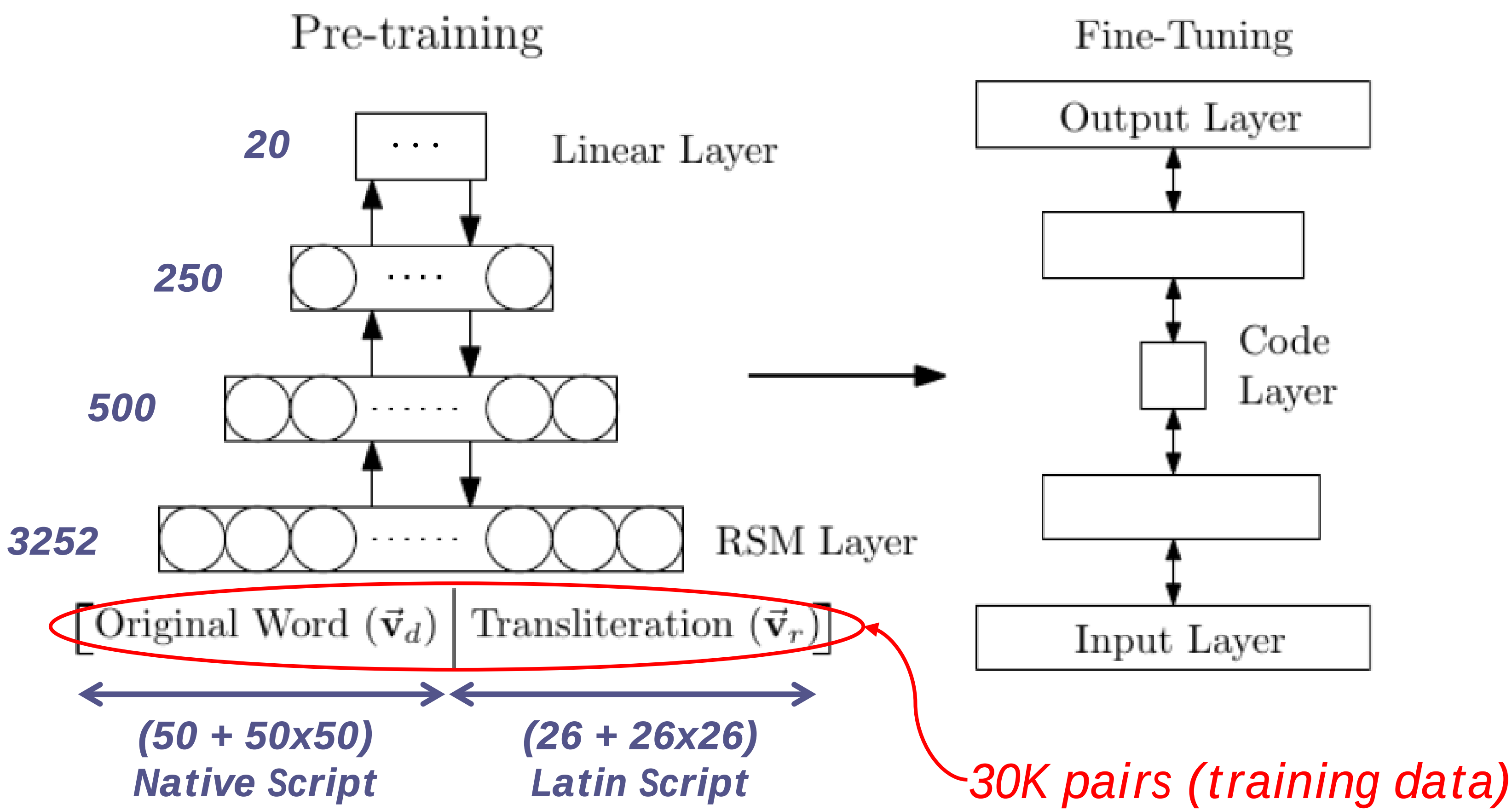


Proposed Method for MSIR*

- Use characters and bigram of characters as terms (features) and words as documents (observations)
- Build a cross-script semantic space by means of a deep autoencoder
- Use the cross-script semantic space for finding “equivalent words” within and across scripts
- Use “equivalent words” for query expansion

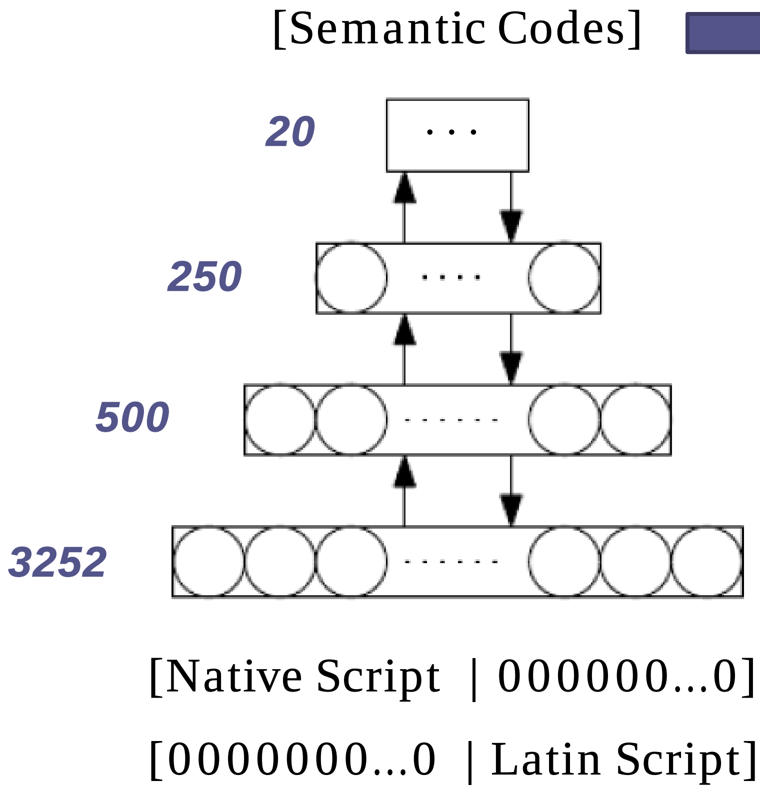
** Gupta P., Bali K., Banchs R.E. Choudhury M. and Rosso P. (2014), Query Expansion for Multi-script Information Retrieval, in Proceedings of the 37st Annual International ACM SIGIR 2014*

Training the Deep Autoencoder

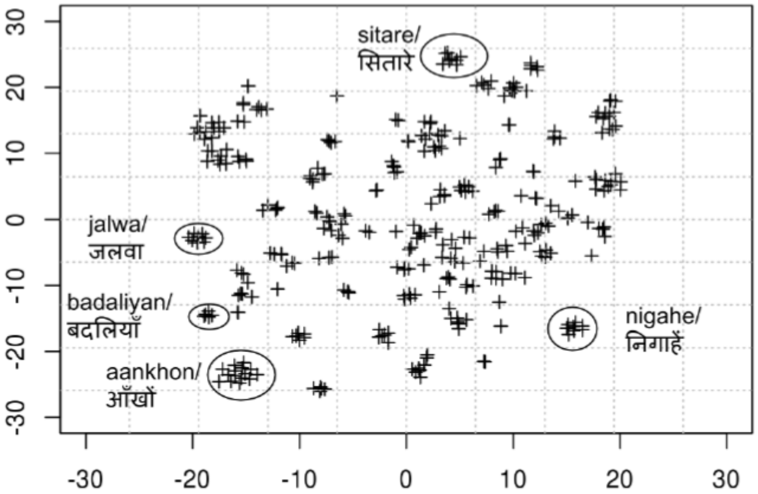


Images taken from Gupta P., Bali K., Banchs R.E. Choudhury M. and Rosso P. (2014), Query Expansion for Multi-script Information Retrieval, in Proc. of the 37st Annual International ACM SIGIR 2014

Building the Semantic Space



2D Visualization of the constructed cross-script semantic space



All available words used

Images taken from Gupta P., Bali K., Banchs R.E. Choudhury M. and Rosso P. (2014), Query Expansion for Multi-script Information Retrieval, in Proc. of the 37st Annual International ACM SIGIR 2014

Cross-script query expansion

Original Query	**ik din ayega**
Query Variants of “ik”	“ik”, “ek”, “एक”
Variants of “din”	“din”, “diin”, “दिन”
Variants of “ayega”	“ayega”,“aeyega”,“ayegaa”, “आयेगा”
Formulated Query (bigram)	ikdin, ikdiin, ··· diin$ayegaa, “एक$दिन”, “दिन$आयेगा”

Baseline Systems

The proposed method is compared to:

- Naïve system: no query expansion used
- LSI: uses cross-language LSI to find the word equivalents
- CCA: uses Canonical Correlation Analysis* to find the word equivalents

** Kumar S. and Udupa R. (2011), Learning hash functions for cross-view similarity search, in Proceedings of IJCAI, pp.1360-1365*

Comparative Evaluation Results

Method	Mean Average Precision	Similarity Threshold
Naïve	29.10%	NA
LSI	35.22%	0.920
CCA	38.91%	0.997
Autoencoder	50.39%	0.960

Number of "Word Equivalents"

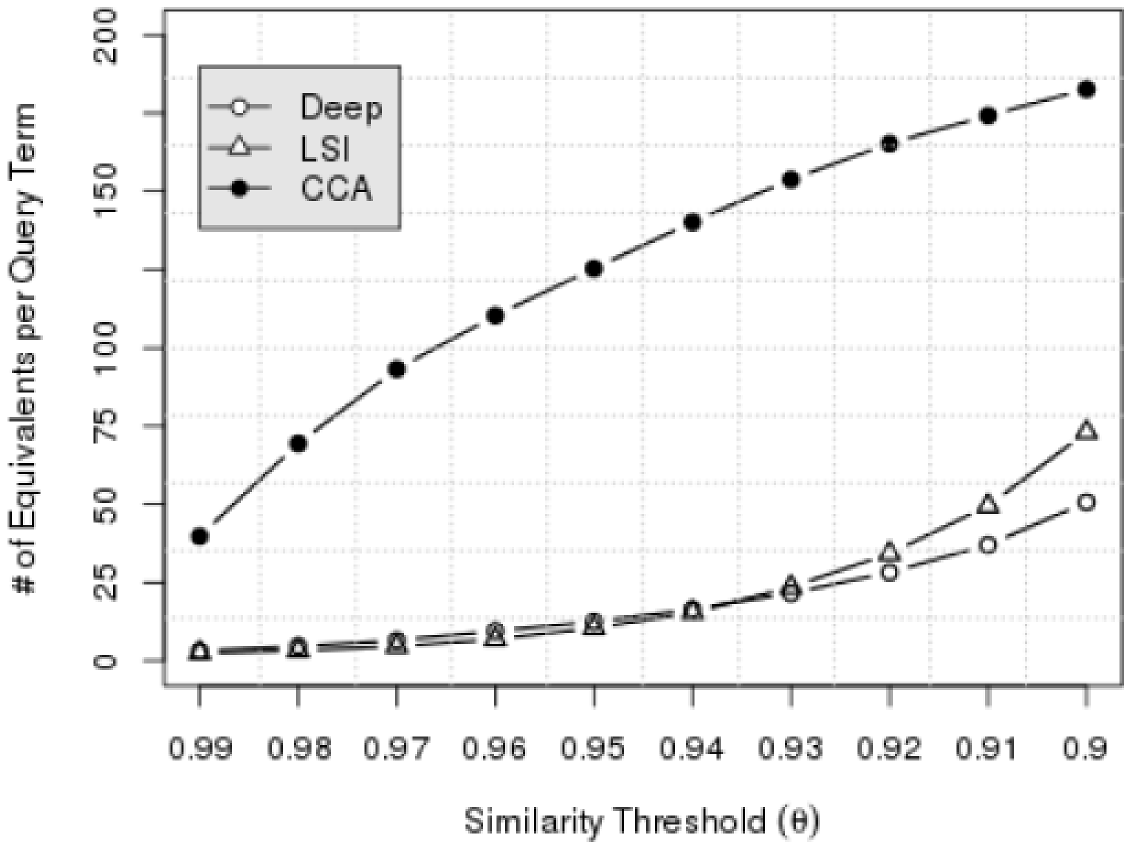


Image taken from Gupta P., Bali K., Banchs R.E. Choudhury M. and Rosso P. (2014), Query Expansion for Multi-script Information Retrieval, in Proc. of the 37st Annual International ACM SIGIR 2014

Section 3

Vector Spaces in Cross-language NLP

- Semantic Map Similarities Across Languages
- Cross-language Information Retrieval in Vector Spaces
- Cross-script Information Retrieval and Transliteration
- **Cross-language Sentence Matching and its Applications**
- Semantic Context Modelling for Machine Translation
- Bilingual Dictionary and Translation-table Generation
- Evaluating Machine Translation in Vector Space

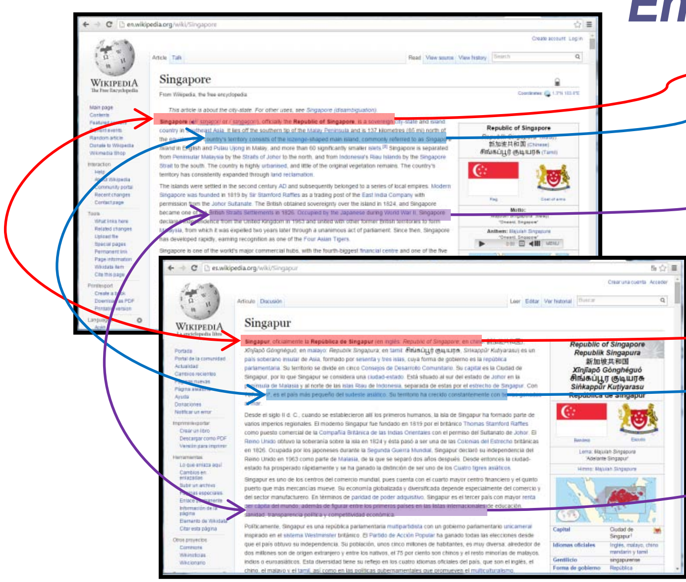
Cross-language Sentence Matching

- Focuses on the specific problem of text matching at the sentence level
- A segment of text in a given language is used as a query for retrieving a similar segment of text in a different language
- This task is useful to some specific applications:
  - Parallel corpora compilation
  - Cross-language plagiarism detection

Parallel Corpora Compilation*

- Deals with the problem of extracting parallel sentence from comparable corpora

English



1. Singapore, officially the Republic of Singapore
2. is a sovereign city-state and island country in Southeast Asia
3. and from Indonesia's Riau Islands by the Singapore Strait to the south
4. ...

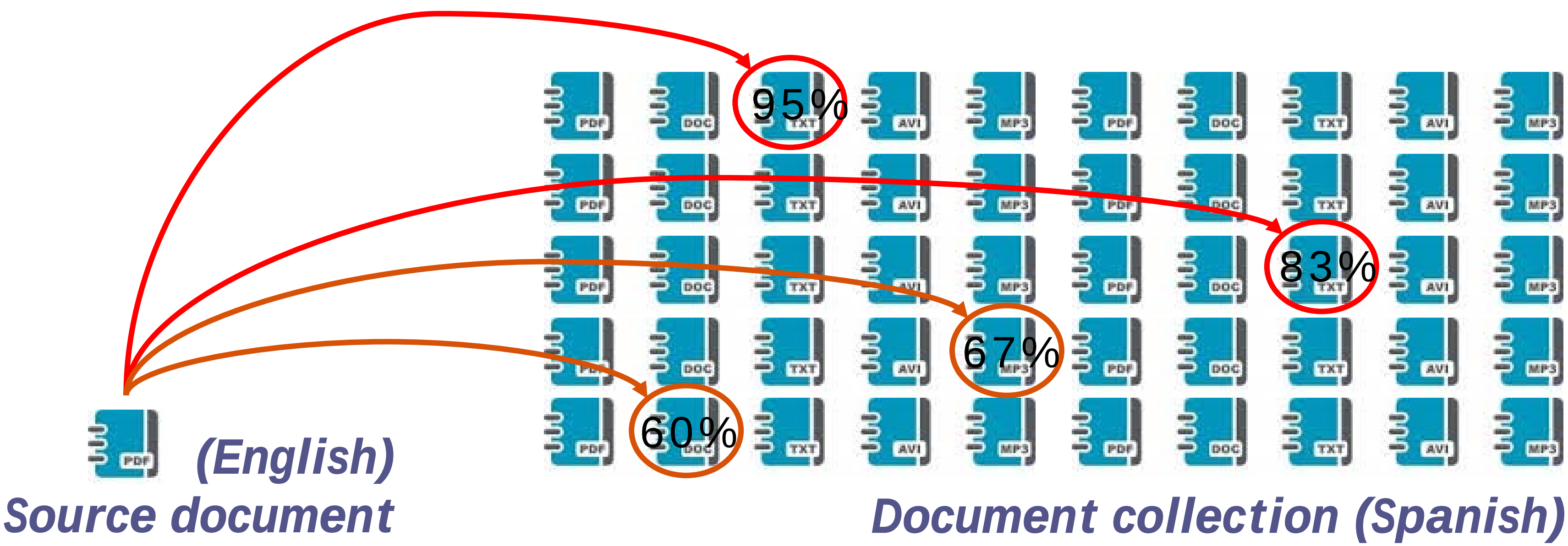
Spanish

1. Singapur, oficialmente la República de Singapur
2. Es un país soberano insular de Asia
3. y al norte de las islas Riau de Indonesia, separada de estas por el estrecho de Singapur
4. ...

** Utiyama M. and Tanimura M. (2007), Automatic construction technology for parallel corpora, Journal of the National Institute of Information and Communications Technology, 54(3), pp.25-31*

CL Plagiarism Detection*

- Deals with the problem of identifying copied documents or fragments across languages



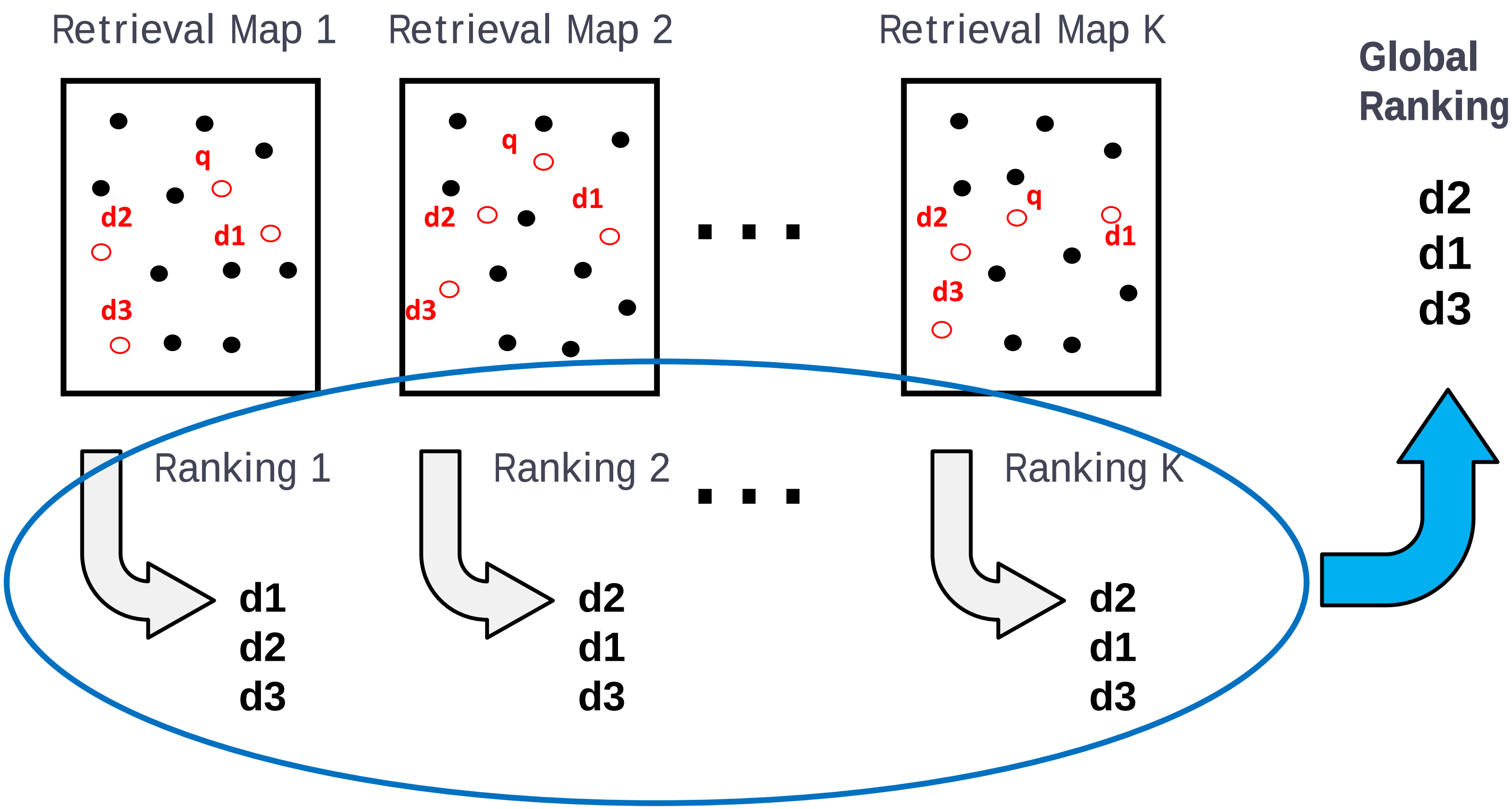
** Potthast M., Stein B., Eiselt A., Barrón A. and Rosso P. (2009), Overview of the 1st international competition on plagiarism detection, Workshop on Uncovering Plagiarism, Authorship, and Social Software Misuse*

Proposed Method

- The previously described MDS-based Semantic Map approach to CLIR is used
  - Cross-language projection matrix variant*
  - Additionally, a majority voting strategy over different semantic retrieval maps is implemented and tested

** Banchs R.E. and Costa-jussà M.R. (2010), A non-linear semantic mapping technique for cross-language sentence matching, in Proceedings of the 7th international conference on Advances in natural language processing (IceTAL'10), pp. 57-66.*

Majority Voting Strategy



Penta-lingual Data Collection

Extracted from the Spanish Constitution

	English	Spanish	Català	Euskera	Galego
Number of sentences	611	611	611	611	611
Number of words	15285	14807	15423	10483	13760
Vocabulary size	2080	2516	2523	3633	2667
Average sentence length	25.01	24.23	25.24	17.16	22.52

Language	Sample sentence
English	This right may not be restricted for political or ideological reasons
Spanish	Este derecho no podrá ser limitado por motivos políticos o ideológicos
Català	Aquest dret no podrà ser limitat por motius polítics o ideològics
Euskera	Eskubide hau arrazoi politiko edo idiologikoek ezin dute mugatu
Galego	Este dereito non poderá ser limitado por motivos políticos ou ideolóxicos

Task Description

- To retrieve a sentence from the English version of the Spanish Constitution using the same sentence in any of the other four languages as a query
- Performance quality is evaluated by means of top-1 and top-5 accuracies measured over a 200-sentence test set
- One retrieval map is constructed for each language available in the collection (400 anchor documents)
- Retrieval Map dimensionality for all languages: 350

Evaluation Results

	Spanish		**Català**		**Euskera**		**Galego**	
Retrieval Map	***top-1***	***top-5***	***top-1***	***top-5***	***top-1***	***top-5***	***top-1***	***top-5***
English	97.0	**100**	96.0	99.0	69.5	91.0	**95.0**	98.5
Spanish	95.5	99.0	94.5	99.5	**77.0**	**93.0**	94.0	**99.5**
Català	95.0	**100**	94.5	99.5	74.5	90.5	93.0	99.0
Euskera	96.5	99.0	95.0	99.5	70.0	86.5	**95.0**	98.5
Galego	96.5	**100**	94.5	**100**	73.0	91.5	93.0	98.0
Majority voting	**97.5**	**100**	**96.5**	99.5	76.0	92.5	94.5	**99.5**

Comparative Evaluation

- The proposed method (majority voting result) is compared to other two methods:
  - Cross-language LSI* (previously described)
  - Query translation** (a cascade combination of machine translation and monolingual information retrieval)

** Dumais S.T., Letsche T.A., Littman M.L. and Landauer T.K. (1997), Automatic Cross-Language Retrieval Using Latent Semantic Indexing, in AAAI-97 Spring Symposium Series: Cross-Language Text and Speech Retrieval, pp. 18-24*

*** Chen J. and Bao Y. (2009), Cross-language search: The case of Google language tools, First Monday, 14(3-2)*

Comparative Evaluation Results

	Spanish		Català		Euskera		Galego	
CLIR Method	***top-1***	***top-5***	***top-1***	***top-5***	***top-1***	***top-5***	***top-1***	***top-5***
LSI based	96.0	99.0	95.5	98.5	75.5	90.5	93.5	97.5
Query transl.	96.0	99.0	95.5	99.5	*	*	93.5	98.0
Semantic maps	**97.5**	**100**	**96.5**	**99.5**	**76.0**	**92.5**	**94.5**	**99.5**

* *Euskera-to-English* translations were not available

Section 3

Vector Spaces in Cross-language NLP

- Semantic Map Similarities Across Languages
- Cross-language Information Retrieval in Vector Spaces
- Cross-script Information Retrieval and Transliteration
- Cross-language Sentence Matching and its Applications
- **Semantic Context Modelling for Machine Translation**
- Bilingual Dictionary and Translation-table Generation
- Evaluating Machine Translation in Vector Space

Statistical Machine Translation

Developing context-awareness in SMT systems

- Original noisy channel formulation:

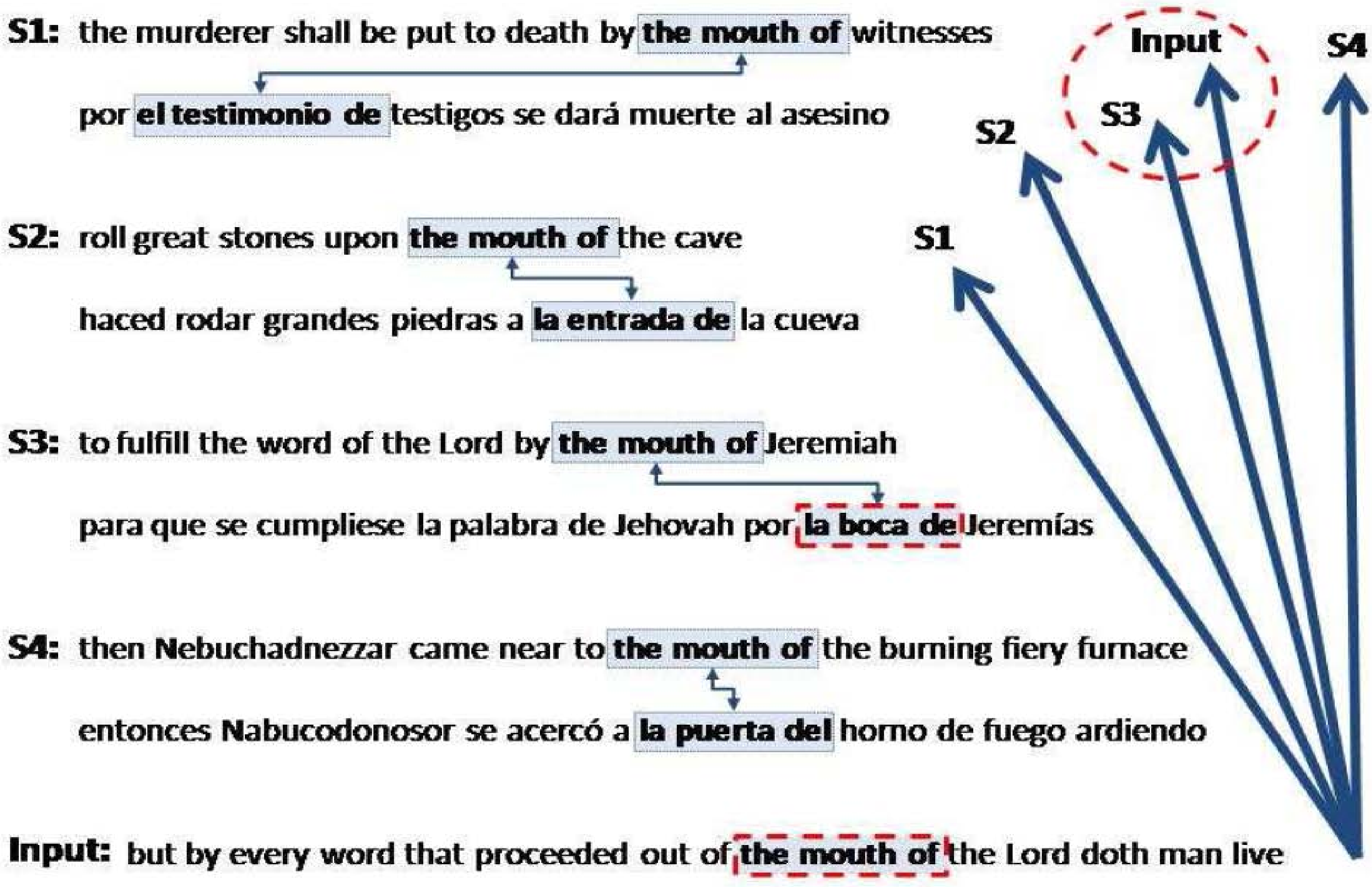
$$\hat{T} = \underset{T}{\operatorname{argmax}}\, P(T|S) = \underset{T}{\operatorname{argmax}}\, P(S|T)\, P(T)$$

- Proposed model reformulation*: **Context Awareness Model**

$$\hat{T} = \underset{T}{\operatorname{argmax}}\, P(T|S,C) = \underset{T}{\operatorname{argmax}}\, P(C|S,T)\, P(S|T)\, P(T)$$

** Banchs R.E. (2014), A Principled Approach to Context-Aware Machine Translation, in Proceedings of the EACL 2014 Third Workshop on Hybrid Approaches to Translation*

Unit Selection Depends on Context



An Actual Example...

"WINE" sense of "VINO"

SC1: *No habéis comido pan ni tomado* ***vino*** *ni licor...*

Ye have not eaten bread, neither have ye drunk ***wine*** *or strong drink...*

SC2: *...dieron muchas primicias de grano,* ***vino*** *nuevo, aceite, miel y de todos ...*

... brought in abundance the first fruits of corn, ***wine****, oil, honey, and of all ...*

"CAME" sense of "VINO"

SC3: *Al tercer día* ***vino*** *Jeroboam con todo el pueblo a Roboam ...*

So Jeroboam and all the people ***came*** *to Rehoboam the third day ...*

SC4: *Ella* ***vino*** *y ha estado desde la mañana hasta ahora ...*

She ***came*** *, and hath continued even from the morning until now ...*

IN1: *... una tierra como la vuestra, tierra de grano y de* ***vino****, tierra de pan y de viñas ...*

(wine)

IN2: *Cuando amanecía, la mujer* ***vino*** *y cayó delante de la puerta de la casa de aquel ...*

(came)

Translation probabilities

- Translation probabilities:

Phrase	$\phi(f\|e)$	$lex(f\|e)$	$\phi(e\|f)$	$lex(e\|f)$
{*vino*|||*wine*}	0.665198	0.721612	0.273551	0.329431
{*vino*|||*came*}	0.253568	0.131398	0.418478	0.446488

- Proposed context-awareness model:

	SC1	SC2	SC3	SC4
sense	{*vino*|||*wine*}		{*vino*|||*came*}	
IN1	**0.0636**	**0.2666**	0.0351	0.0310
IN2	0.0023	0.0513	**0.0888**	**0.0774**

Comparative evaluation*

	Development	Test
Baseline System	39.92	38.92
Vector Space Model	40.61	39.43
Statistical Class Model	40.62	39.72
Latent Dirichlet Allocation	40.63	39.82
Latent Semantic Indexing	**40.80**	**39.86**

***** *Banchs R.E. and Costa-jussà M.R. (2011), A Semantic Feature for Statistical Machine Translation, in Fifth Workshop on Syntax, Semantics and Structure in Statistical Translation, ACL 2011, pp. 126–134*

Neural Network Models for MT*

- The Neural Network framework can be used to incorporate source context information in both:
  - the target language model:
    Neural Network Joint Model (NNJM)
  - the translation model:
    Neural Network Lexical Translation Model (NNLTM)

** Devlin J., Zbib R., Huang Z., Lamar T., Schwartz R. and Makhoul J. (2014), Fast and Robust Neural Network Joint Models for Statistical Machine Translation, in Proceedings of the 52 Annual Meeting of the Association for Computational Linguistics, pp. 1370-1380*

Joint Model (NNJM)

- Estimates the probability of a target word given its previous word history and a source context window

$$P(T|S) \approx \prod_{i=1}^{|T|} P(\underbrace{t_i}_{\text{Target word}} \mid \underbrace{t_{i-1}, t_{i-2} \ldots t_{i-n}}_{\text{Target history}}, \underbrace{s_{j+m}, s_{j+m-1} \ldots s_j \ldots s_{j-m+1}, s_{j-m}}_{\text{Source context window}})$$

with $j = f_a(i)$

Lexical Translation Model (NNLTM)

- Estimates the probability of a target word given a source context window

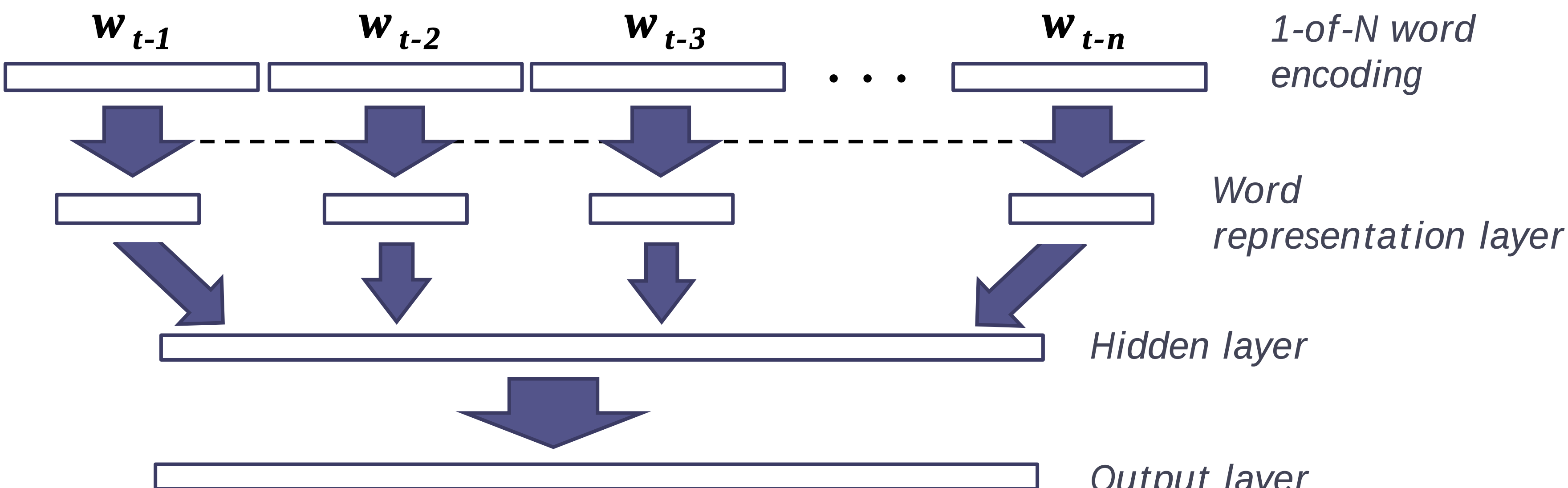
$$P(T|S) \approx \prod_{j=1}^{|S|} P(t_i \mid \overbrace{s_{j+m}, s_{j+m-1} \dots s_j \dots s_{j-m+1}, s_{j-m}}^{\textit{Source context window}})$$

Target word

with $i = f_a(j)$

Neural Network Architecture

- Feed-forward Neural Network Language Model*

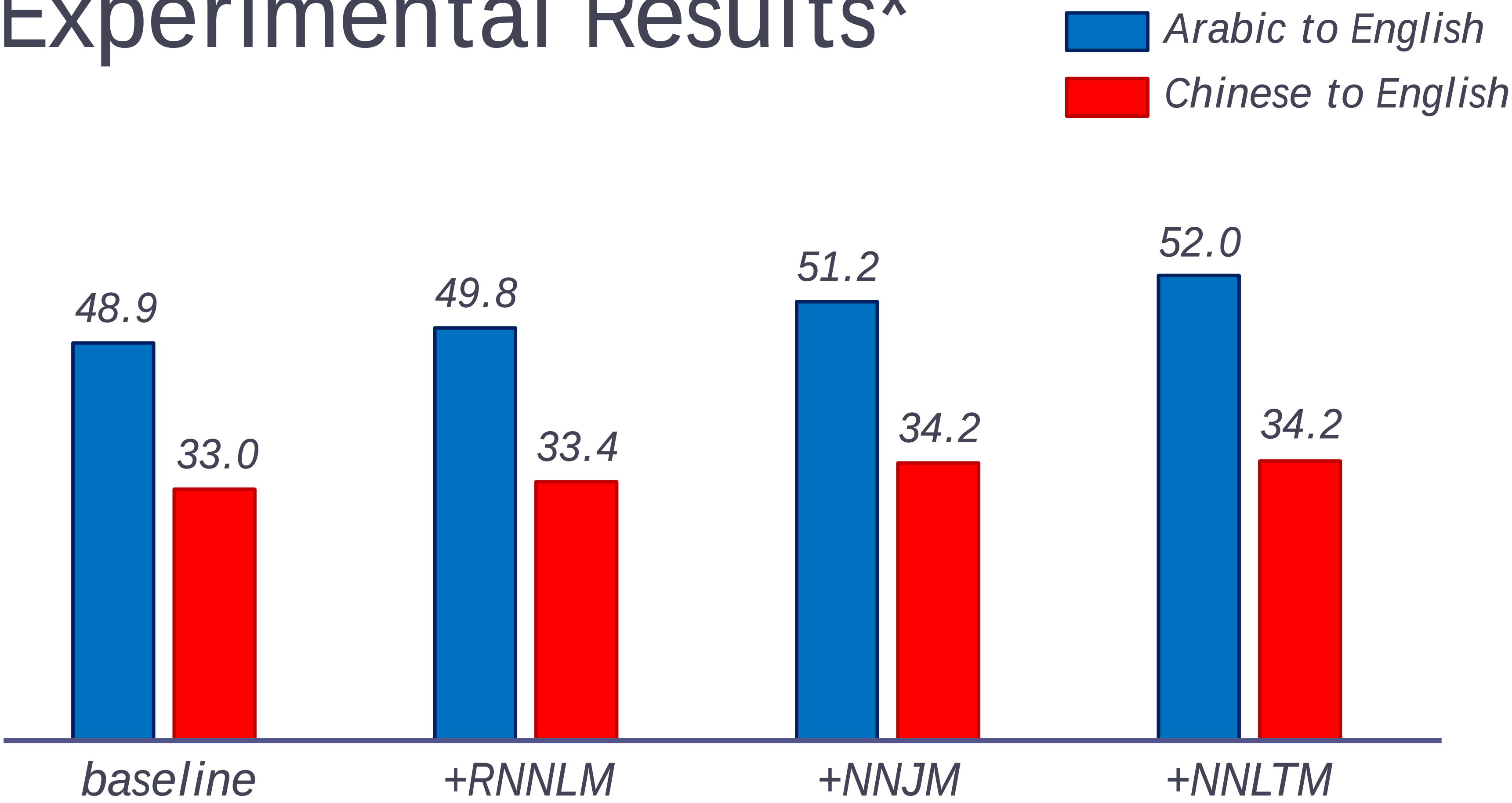


$$\boldsymbol{y} = \boldsymbol{V} f(\boldsymbol{b} + \boldsymbol{W} [\boldsymbol{C}\, \boldsymbol{w}_{t-1}, \boldsymbol{C}\, \boldsymbol{w}_{t-2} \ldots \boldsymbol{C}\, \boldsymbol{w}_{t-n}])$$

$$y_i = p(w_t = i \mid context)$$

* *Bengio J., Ducharme R., Vincent P. and Jauvin C. (2003), A neural probabilistic language model, Journal of Machine Learning Research, 3, pp.1137-1155*

Experimental Results*



** Devlin J., Zbib R., Huang Z., Lamar T., Schwartz R. and Makhoul J. (2014), Fast and Robust Neural Network Joint Models for Statistical Machine Translation, in Proceedings of the 52 Annual Meeting of the Association for Computational Linguistics, pp. 1370-1380*

Section 3

Vector Spaces in Cross-language NLP

- Semantic Map Similarities Across Languages
- Cross-language Information Retrieval in Vector Spaces
- Cross-script Information Retrieval and Transliteration
- Cross-language Sentence Matching and its Applications
- Semantic Context Modelling for Machine Translation
- **Bilingual Dictionary and Translation-table Generation**
- Evaluating Machine Translation in Vector Space

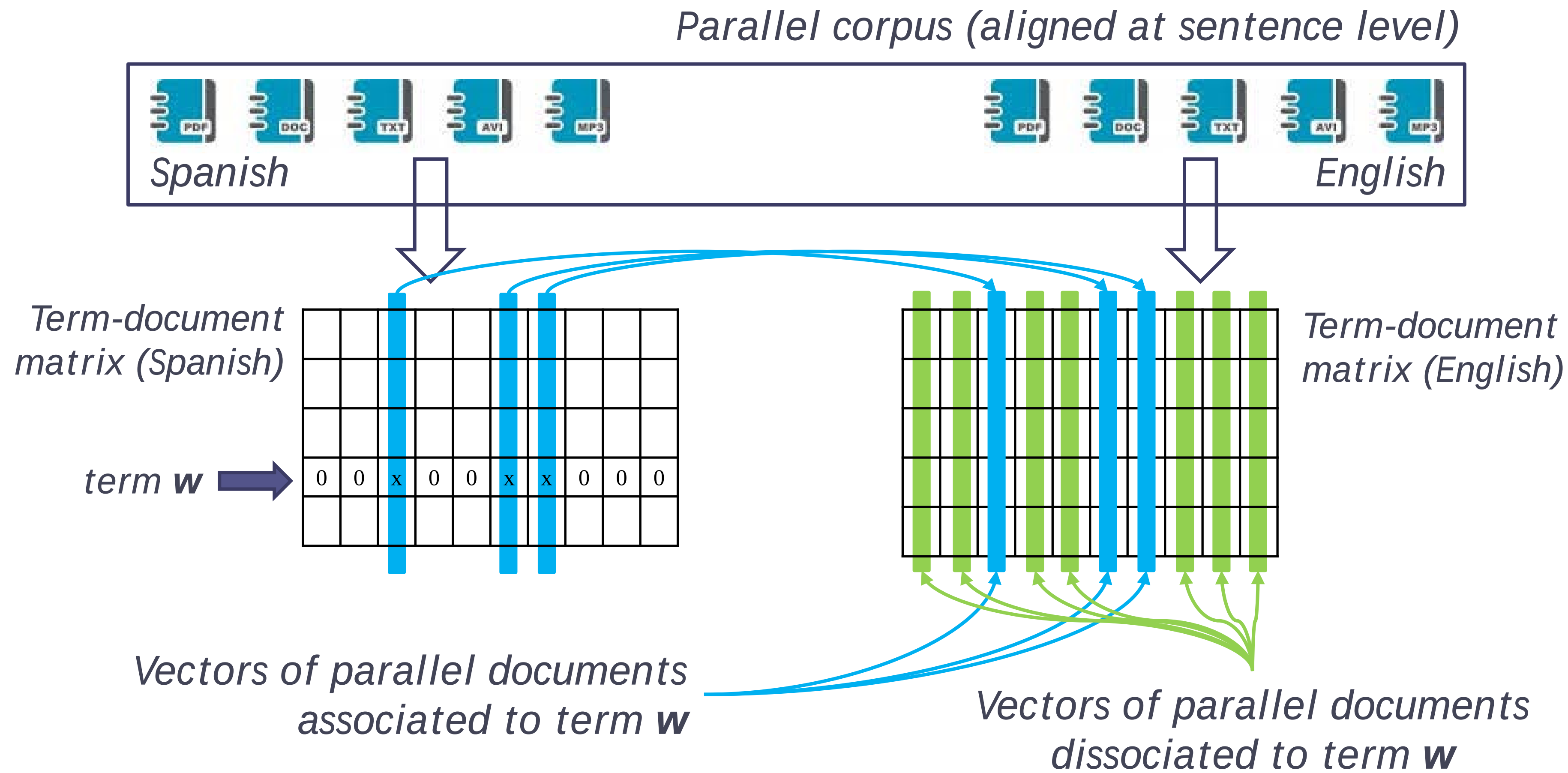
Word Translations in Vector Space

- Semantic similarities across languages can be exploited to “discover” word translation pairs from parallel data collections by:
  - either operating in the term-document matrix space*
  - or learning transformations across reduced spaces**

** Banchs R.E. (2013), Text Mining with MATLAB, Springer , chap. 11, pp. 277-311*

*** Mikolov T., Le Q.V. and Sutskever I. (2013), Exploiting Similarities among Languages for Machine Translation, arXiv:1309.4168v1*

Operating in Term-document Space*



* Banchs R.E. (2013), Text Mining with MATLAB, Springer , chap. 11, pp. 277-311

Obtaining the Translation Terms*

- Compute $\boldsymbol{V}^{+}$, the average vector of parallel documents associated to term $\boldsymbol{w}$
- Compute $\boldsymbol{V}^{-}$, the average vector of parallel documents dissociated to term $\boldsymbol{w}$
- Obtain the most relevant terms (with largest weights) for the difference vector $\boldsymbol{V}^{+} - \boldsymbol{V}^{-}$

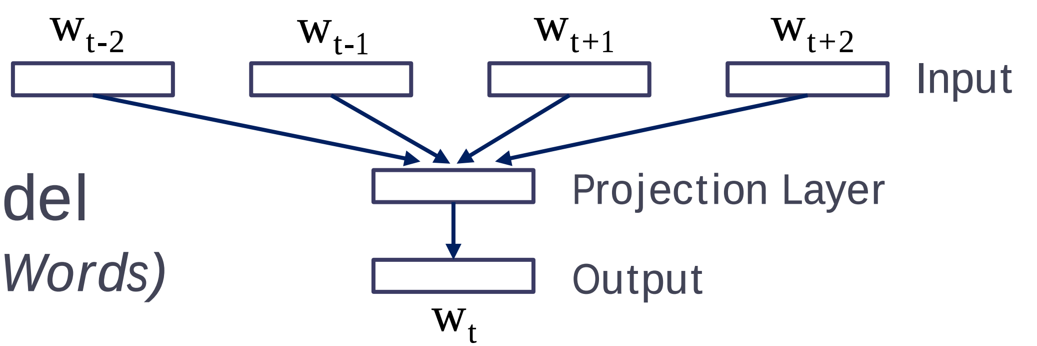
** Banchs R.E. (2013), Text Mining with MATLAB, Springer , chap. 11, pp. 277-311*

Some Sample Translations

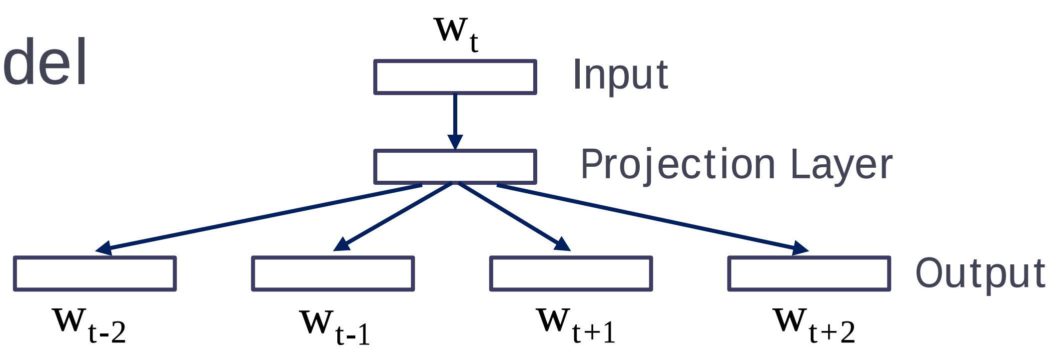
- English translations to Spanish terms:
  - casa: house, home
  - ladrón: thief, sure, fool
  - caballo: horse, horseback
- Spanish translations to English terms:
  - city: ciudad, fortaleza
  - fields: campo, vida
  - heart: corazón, ánimo, alma

Learning Projections*

- Construct projection spaces by means of
  - either CBOW model
    (Continuous Bag-Of-Words)



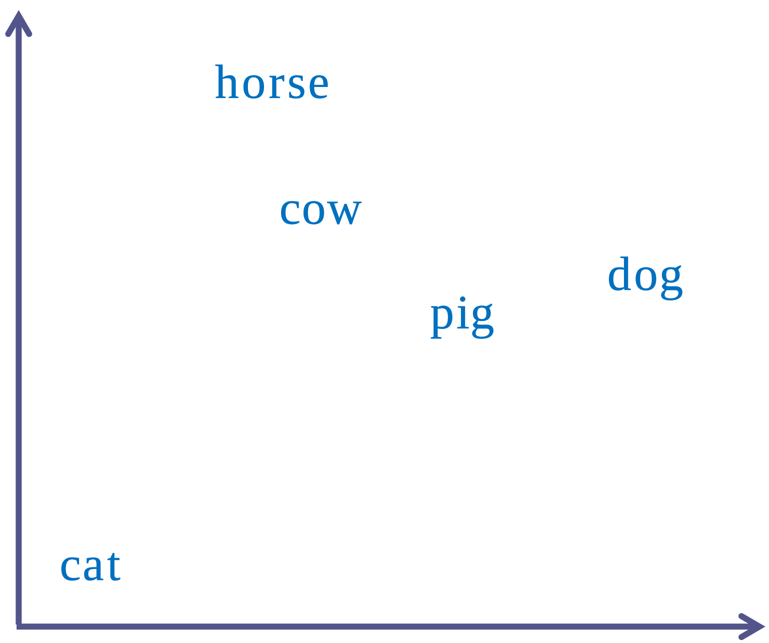
  - or Skip-gram model



* *Mikolov T., Le Q.V. and Sutskever I. (2013), Exploiting Similarities among Languages for Machine Translation, arXiv:1309.4168v1*

Some Sample Projections

English Semantic Map for Animals



Spanish Semantic Map for Animals

Images taken from Mikolov T., Le Q.V. and Sutskever I. (2013), Exploiting Similarities among Languages for Machine Translation, arXiv:1309.4168v1

Obtaining the Translation Terms

- Use some bilingual word pairs $\{s_i, t_i\}$ to train a "translation matrix" $\boldsymbol{W}$ such that:

$$t_i \approx W s_i$$

- Use $\boldsymbol{W}$ for projecting a new term s_j into the target space
- Collect the terms in target space that are closest to the obtained projection

Some Sample Translations*

- English translations to Spanish terms:
  - emociones: emotions, emotion, feeling
  - imperio: dictatorship, imperialism, tyranny
  - preparada: prepared, ready, prepare
  - millas: kilometers, kilometres, miles
  - hablamos: talking, talked, talk

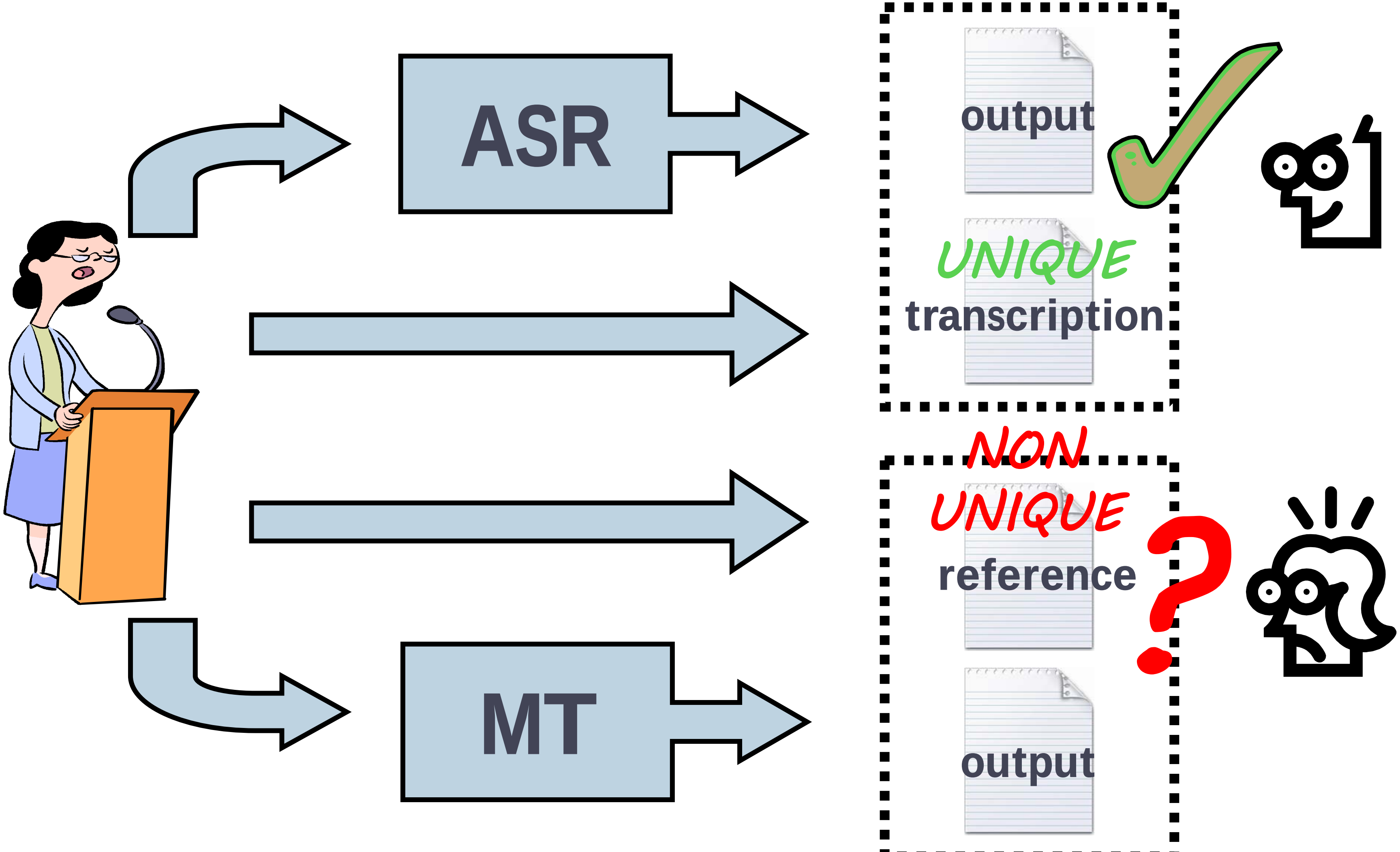
** Mikolov T., Le Q.V. and Sutskever I. (2013), Exploiting Similarities among Languages for Machine Translation, arXiv:1309.4168v1*

Section 3

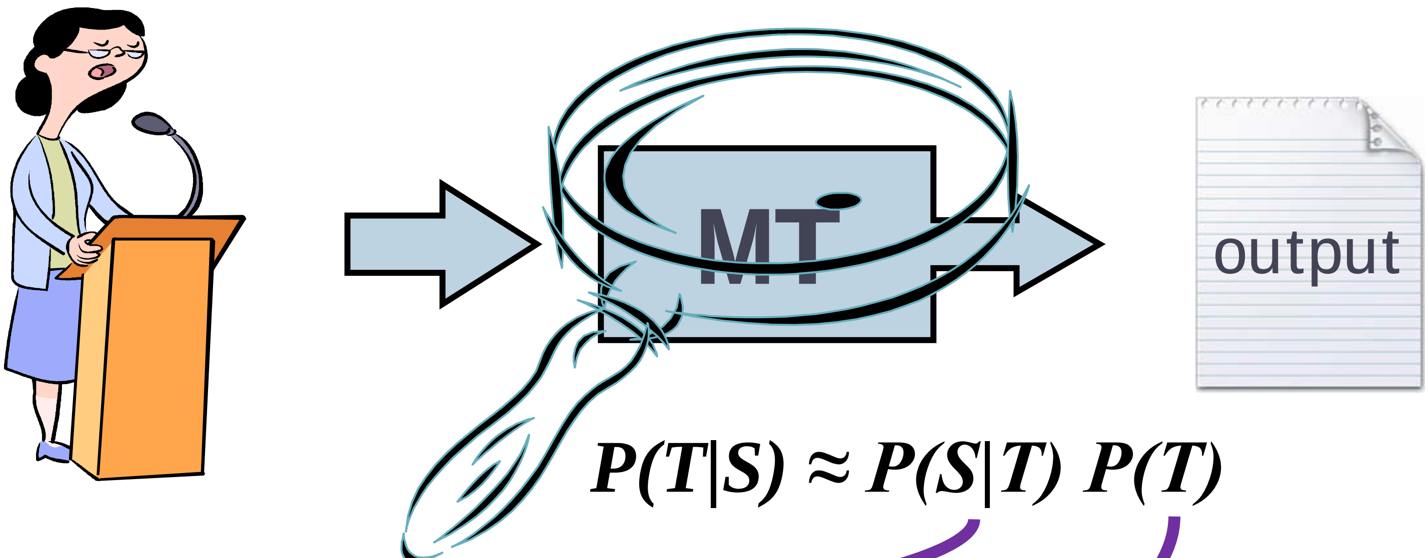
Vector Spaces in Cross-language NLP

- Semantic Map Similarities Across Languages
- Cross-language Information Retrieval in Vector Spaces
- Cross-script Information Retrieval and Transliteration
- Cross-language Sentence Matching and its Applications
- Semantic Context Modelling for Machine Translation
- Bilingual Dictionary and Translation-table Generation
- **Evaluating Machine Translation in Vector Space**

Automatic Evaluation of MT



Human Evaluation of MT*



$$P(T|S) \approx P(S|T)\,P(T)$$

ADEQUACY

How much of the source information is preserved?

FLUENCY

How good is the generated target language quality?

** White J.S., O'Cornell T. and Nava F.O. (1994), The ARPA MT evaluation methodologies: evolution, lessons and future approaches, in Proc. of the Assoc. for Machine Translation in the Americas, pp. 193-205*

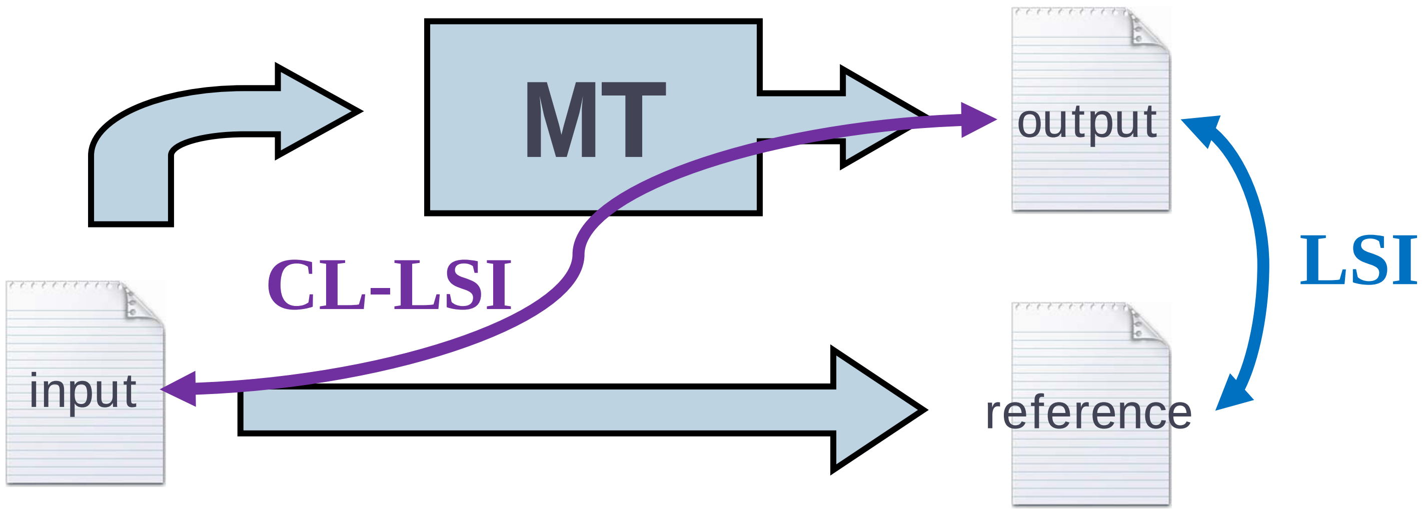
Proposed Evaluation Framework*

- Approximate adequacy and fluency by means of independent models:
  - Use a “semantic approach” for adequacy
  - Use a “syntactic approach” for fluency
- Combine both evaluation metrics into a single evaluation score

***** *Banchs R.E. and Li H. (2011), AM-FM: A Semantic Framework for Translation Quality Assessment, in Proceedings of the 49th Annual Meeting of the ACL, shortpapers, pp. 153–158*

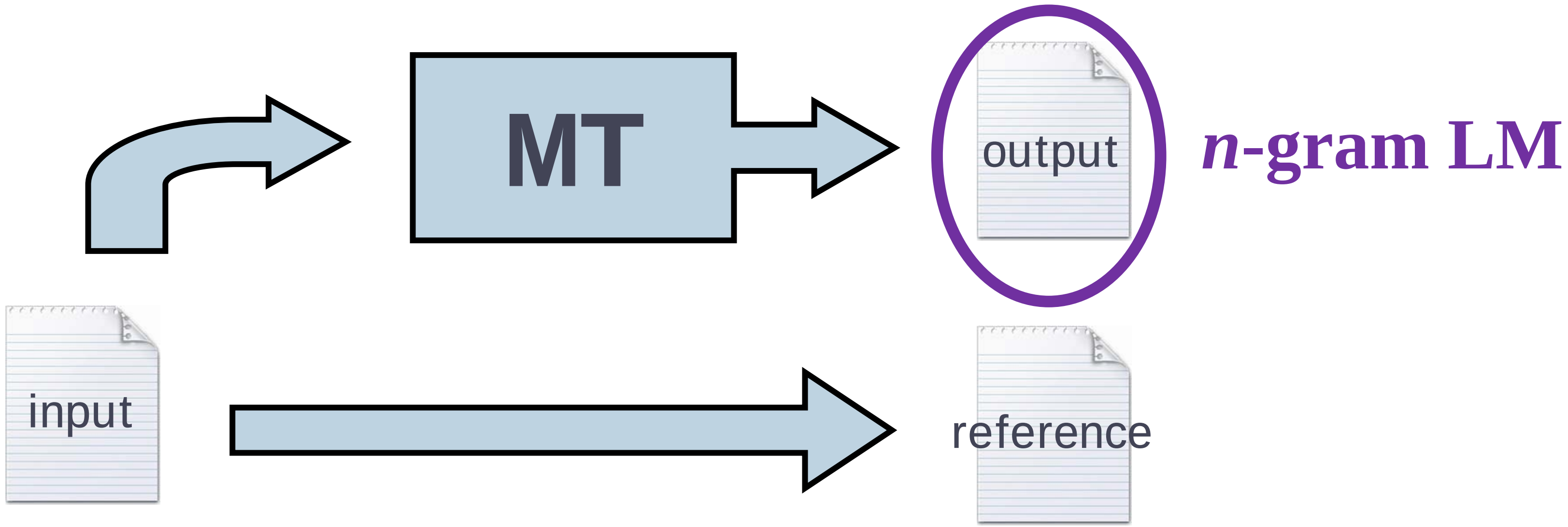
AM: Adequacy-oriented Metric

- Compare sentences in a semantic space
  - Monolingual AM (***mAM***): compare output vs. reference
  - Cross-language AM (***xAM***): compare output vs. input



FM: Fluency-oriented Metric

- Measures the quality of the target language with a language model
- Uses a compensation factor to avoid effects derived from differences in sentence lengths



AM-FM Combined Score

Both components can be combined into a single metric according to different criteria

- Weighted Harmonic Mean: $$H\text{-}AM\text{-}FM = \frac{AM \;\; FM}{\alpha\, AM + (1-\alpha)\, FM}$$
- Weighted Mean: $$M\text{-}AM\text{-}FM = (1-\alpha)\, AM + \alpha\, FM$$
- Weighted L2-norm: $$N\text{-}AM\text{-}FM = \sqrt{(1-\alpha)\, AM^2 + \alpha\, FM^2}$$

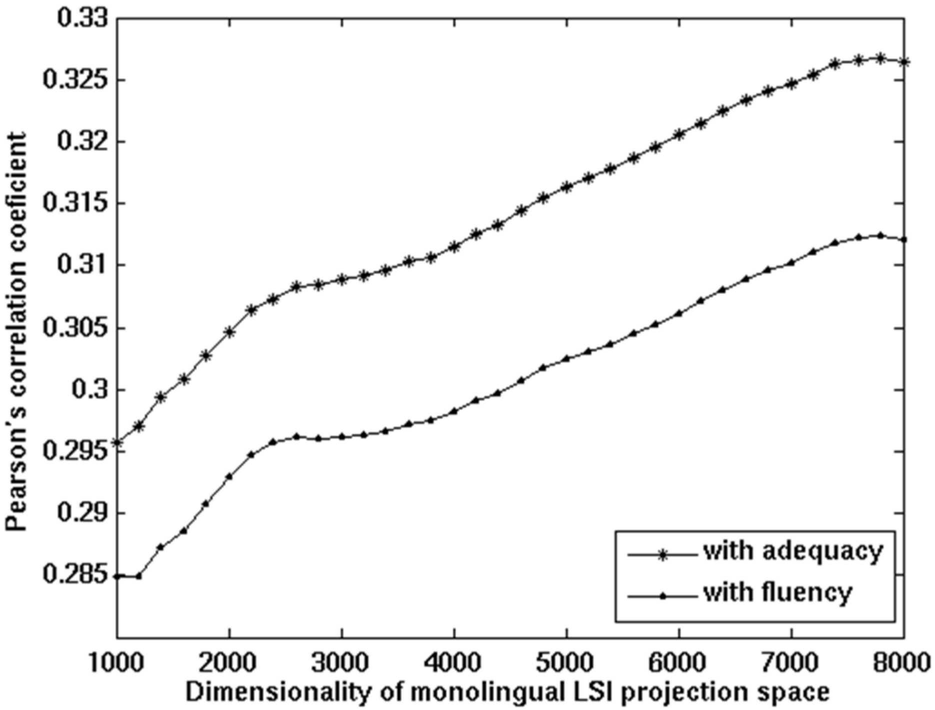
WMT-2007 Dataset*

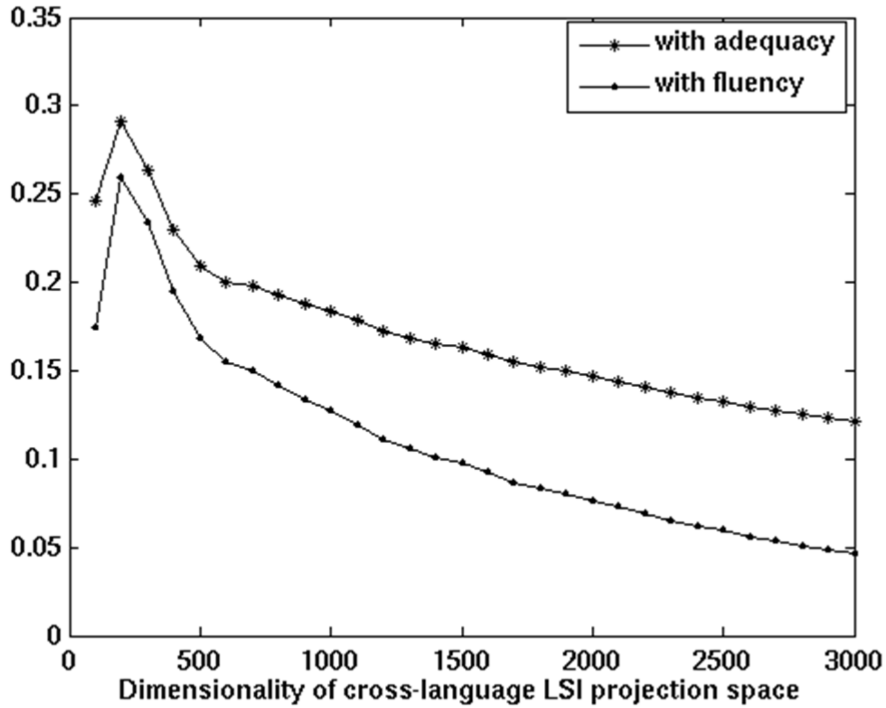
- Fourteen tasks:
  - five European languages (EN, ES, DE, FR, CZ) and
  - two different domains (News and EPPS).
- Systems outputs available for fourteen of the fifteen systems that participated in the evaluation.
- 86 system outputs for a total of 172,315 individual sentence translations, from which 10,754 were rated for both adequacy and fluency by human judges.

***** *Callison-Burch C., Fordyce C., Koehn P., Monz C. and Schroeder J. (2007), (Meta-) evaluation of machine translation, in Proceedings of Statistical Machine Translation Workshop, pp. 136-158*

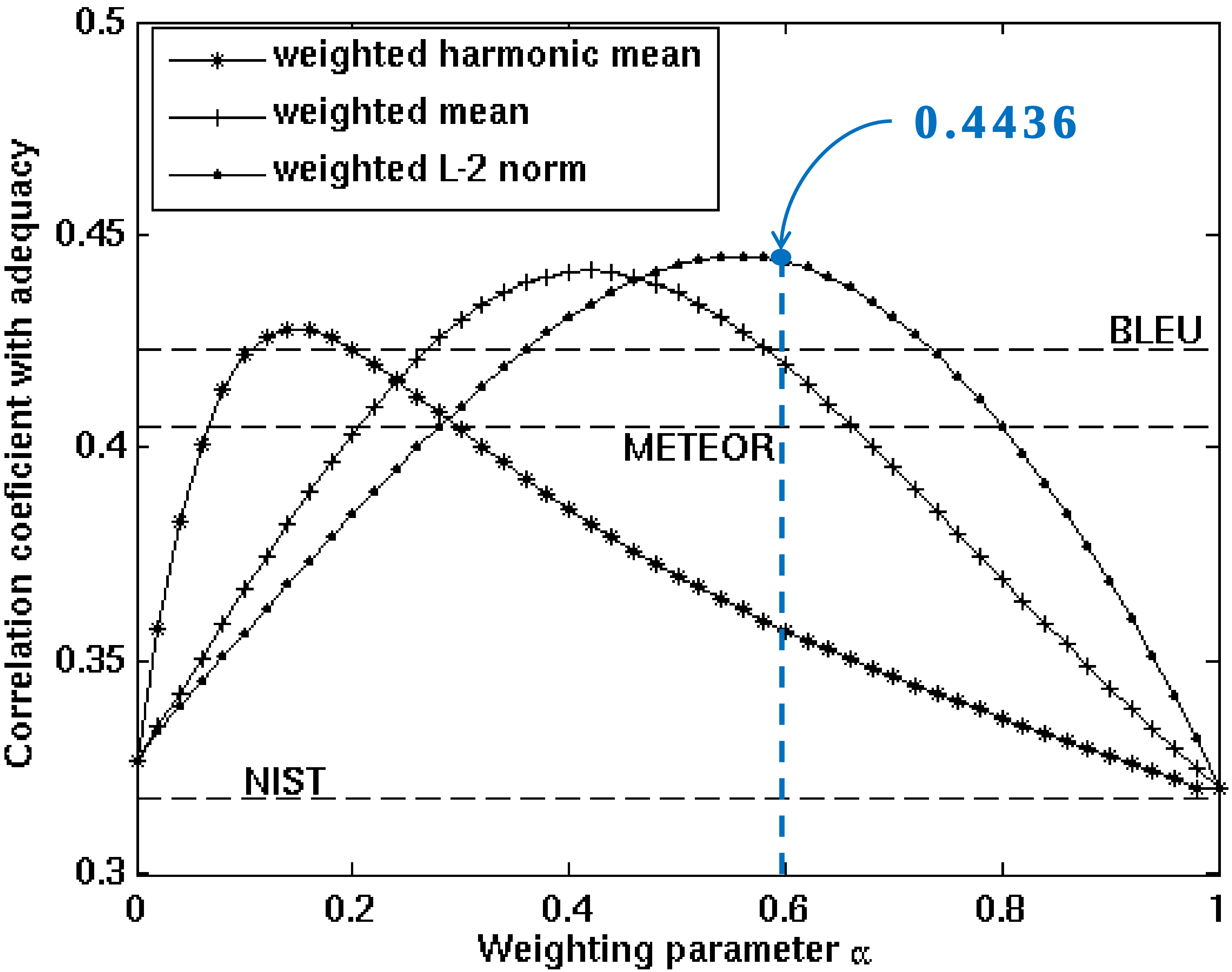
Dimensionality Selection

Pearson's correlation coefficients between the ***mAM*** (left) and ***xAM*** (right) components and human-generated scores

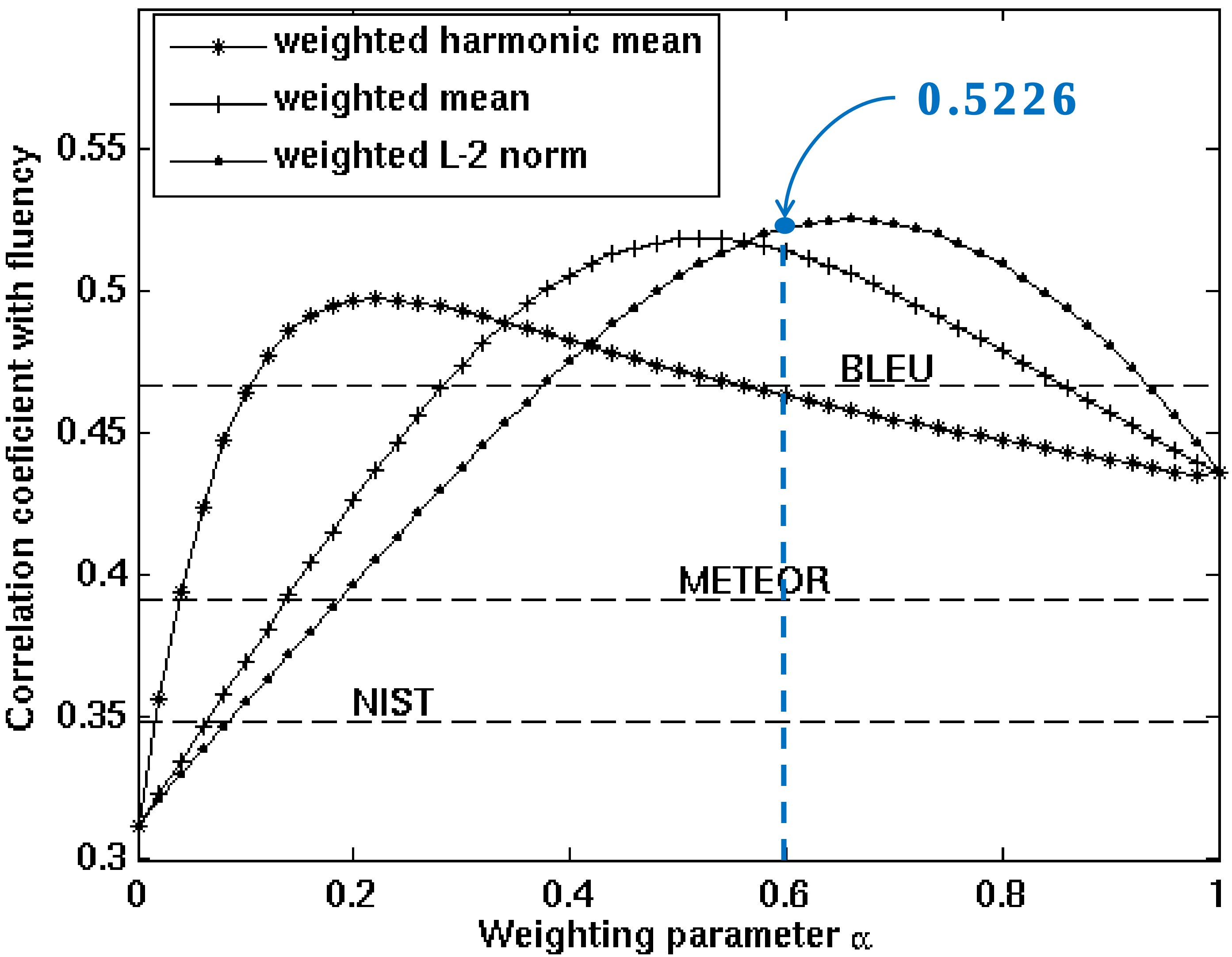




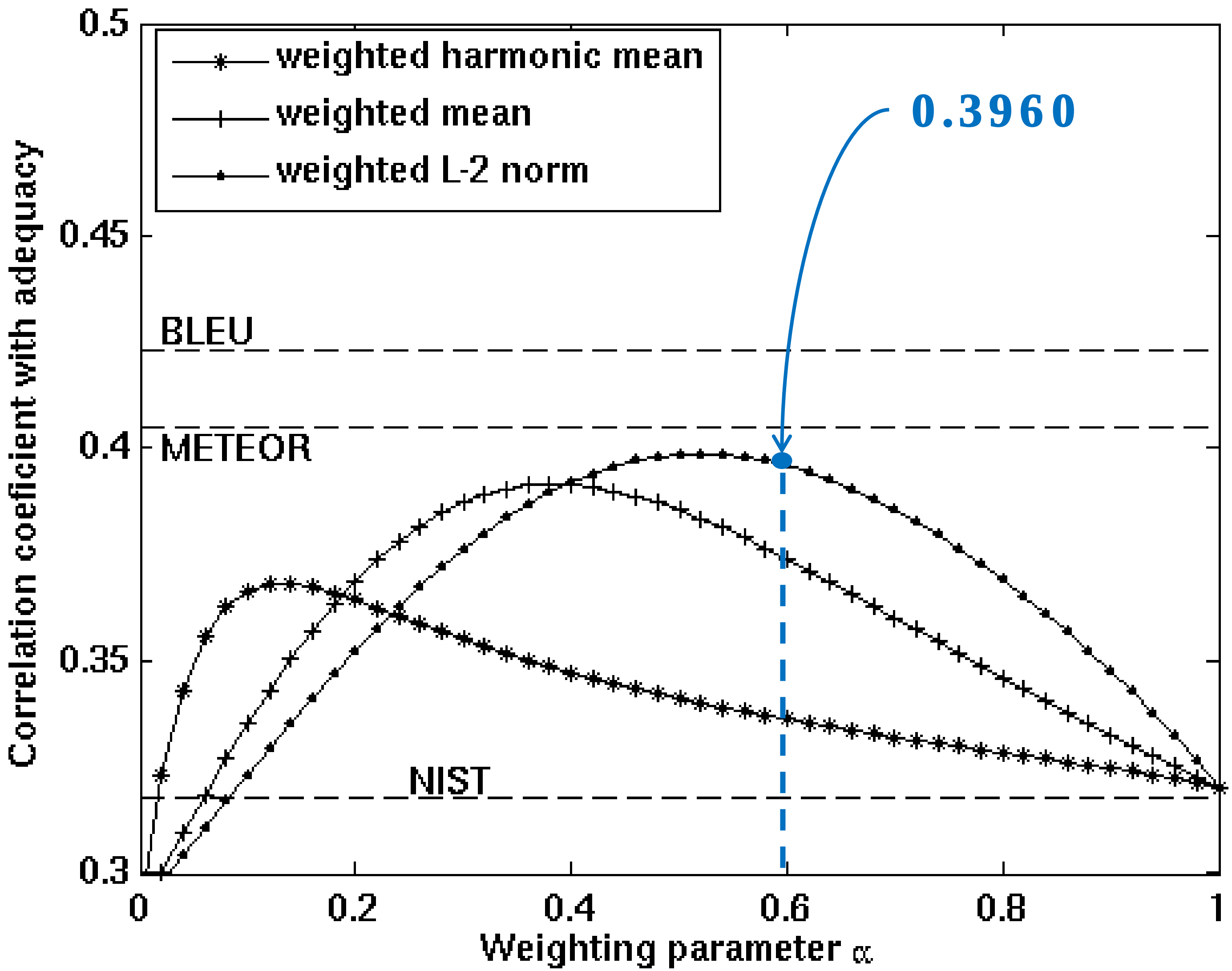
mAM-FM and Adequacy



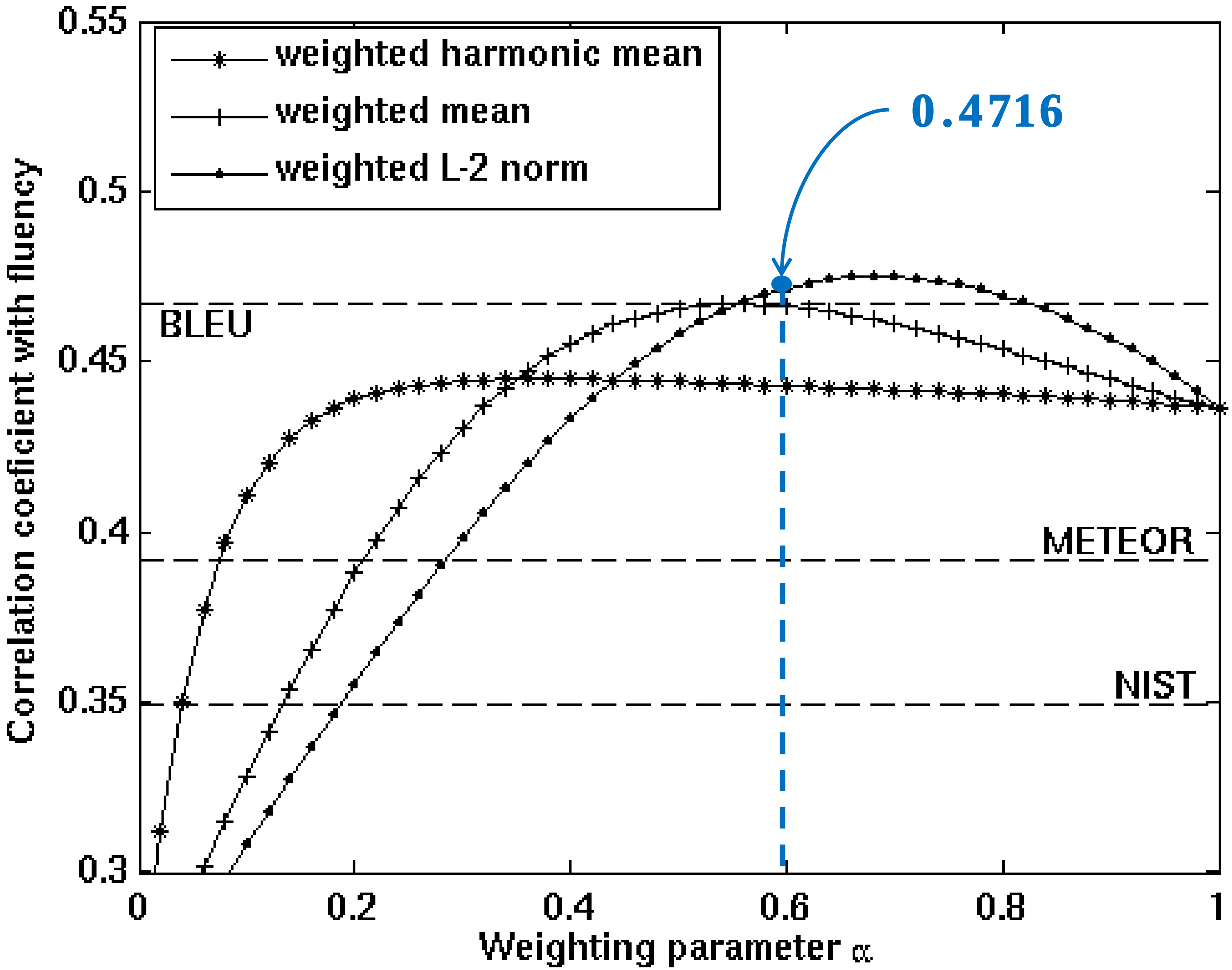
mAM-FM and Fluency



xAM-FM and Adequacy



xAM-FM and Fluency



Section 3

Main references for this section

- R. E. Banchs and A. Kaltenbrunner, 2008, “Exploring MDS projections for cross-language information retrieval”
- P. Gupta, K. Bali, R. E. Banchs, M. Choudhury and P. Rosso, 2014, “Query Expansion for Multi-script Information Retrieval”
- R. E. Banchs and M. R. Costa-jussà, 2010, “A non-linear semantic mapping technique for cross-language sentence matching”
- R. E. Banchs and M. R. Costa-jussà, 2011, “A Semantic Feature for Statistical Machine Translation”

Section 3

Main references for this section

- J. Devlin, R. Zbib, Z. Huang, T. Lamar, R. Schwartz and J. Makhoul,2014, "Fast and Robust Neural Network Joint Models for Statistical Machine Translation"
- T. Mikolov, Q. V. Le and I. Sutskever, 2013, "Exploiting Similarities among Languages for Machine Translation"
- R. E. Banchs and H. Li, 2011, "AM-FM: A Semantic Framework for Translation Quality Assessment"

Section 3

Additional references for this section

- Banchs R.E. and Costa-jussà M.R. (2013), Cross-Language Document Retrieval by using Nonlinear Semantic Mapping, International Journal of Applied Artificial Intelligence, 27(9), pp. 781-802
- Dumais S.T., Letsche T.A., Littman M.L. and Landauer T.K. (1997), Automatic Cross-Language Retrieval Using Latent Semantic Indexing, in AAAI-97 Spring Symposium Series: Cross-Language Text and Speech Retrieval, pp. 18-24
- Kumar S. and Udupa R. (2011), Learning hash functions for cross-view similarity search, in Proceedings of IJCAI, pp.1360-1365
- Utiyama M. and Tanimura M. (2007), Automatic construction technology for parallel corpora, Journal of the National Institute of Information and Communications Technology, 54(3), pp.25-31
- Potthast M., Stein B., Eiselt A., Barrón A. and Rosso P. (2009), Overview of the 1st international competition on plagiarism detection, Workshop on Uncovering Plagiarism, Authorship, and Social Software Misuse

Section 3

Additional references for this section

- Chen J. and Bao Y. (2009), Cross-language search: The case of Google language tools, First Monday, 14(3-2)
- Banchs R.E. (2014), A Principled Approach to Context-Aware Machine Translation, in Proceedings of the EACL 2014 Third Workshop on Hybrid Approaches to Translation
- Bengio J., Ducharme R., Vincent P. and Jauvin C. (2003), A neural probabilistic language model, Journal of Machine Learning Research, 3, pp.1137-1155
- Banchs R.E. (2013), Text Mining with MATLAB, Springer , chap. 11, pp. 277-311
- White J.S., O'Cornell T. and Nava F.O. (1994), The ARPA MT evaluation methodologies: evolution, lessons and future approaches, in Proc. of the Assoc. for Machine Translation in the Americas, pp. 193-205
- Callison-Burch C., Fordyce C., Koehn P., Monz C. and Schroeder J. (2007), (Meta-) evaluation of machine translation, in Proceedings of Statistical Machine Translation Workshop, pp. 136-158

Section 4

Future Research and Applications

- **Current limitations of vector space models**
- Encoding word position information into vectors
- From vectors and matrices to tensors
- Final remarks and conclusions

Conceptual vs. Functional

- Vector Space Models are very good to capture the conceptual aspect of meaning
  - {dog, cow, fish, bird} vs. {chair, table, sofa, bed}
- However, they still fail to properly model the functional aspect of meaning
  - "Give me **a** pencil" vs. "Give me **that** pencil"

Word Order Information Ignored

- Differently from Formal Semantics*, VSM lacks of a clean interconnection between the syntax and semantic phenomena
- In part, a consequence of the Bag-Of-Words nature of VSM

VSMs completely ignore word order information

** Montague R. (1970), Universal Grammar, Theoria, 36, pp. 373-398*

Non-unique Representations

- Consider the two following sentences*
  - *"That day the office manager, who was drinking, hit the problem sales worker with a bottle, but it was not serious"*
  - *"It was not the sales manager, who hit the bottle that day, but the office worker with a serious drinking problem"*
- Although they are completely different, they contain **exactly the same set of words**, so they will produce **exactly the same VSM representation!**

* *Landauer T.K. and Dumais S.T. (1997), A solution to Plato's problem: the latent semantic analysis theory of acquisition, induction and representation of knowledge, Psychological Review, 104(2), pp. 211-240*

Other Limitations

Additionally…

- VSMs are strongly data-dependent
- VSMs noisy in nature (spurious events)
- Uncertainty or confidence estimation becomes an important issue
- Multiplicity of parameters with not clear relation to the outcomes

Section 4

Future Research and Applications

- Current limitations of vector space models
- **Encoding word position information into vectors**
- From vectors and matrices to tensors
- Final remarks and conclusions

Semantics and Word Order

- It is estimated that the meaning of English comes from*
  - Word choice ➡ *80%*
  - Word order ➡ *20%*

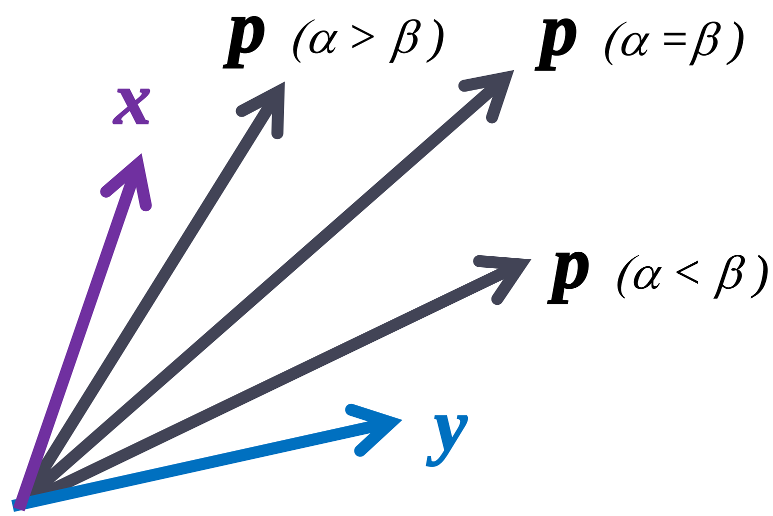
*** *Landauer T.K. (2002), On the computational basis of learning and cognition: Arguments from LSA, in Ross B.H. (ed.) The Psychology of Learning and Motivation: Advances in Research and Theory, 41, pp. 43-84***

Word Order in Additive Models

- Additive composition can be sensitive to word order by weighting the word contributions*

$$\boldsymbol{p} = \boldsymbol{x} + \boldsymbol{y}$$

$$\boldsymbol{p} = \alpha\,\boldsymbol{x} + \beta\,\boldsymbol{y}$$



* *Mitchell J. and Lapata M. (2008), Vector-based models of semantic composition, in Proceedings of ACL –HLT 2008, pp. 236-244*

Circular Convolution Model

- Word order encoded into a vector by collapsing outer-product matrix of word vectors*

$$p_i = \sum_j x_j \ y_{(i-j)\ mod\text{-}n}$$

$$p_i = (\ p_0\ ,\ p_1\ ,\ p_2\ )$$

$$\begin{array}{ccc} x_0\,y_0 & x_0\,y_1 & x_0\,y_2 \\ x_1\,y_0 & x_1\,y_1 & x_1\,y_2 \\ x_2\,y_0 & x_2\,y_1 & x_2\,y_2 \end{array}$$

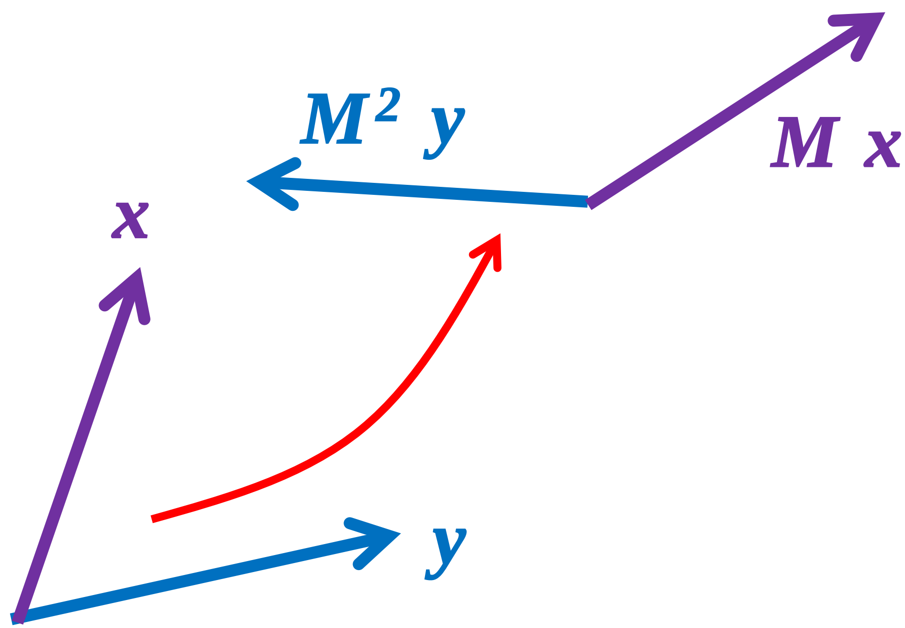
** Jones M.N. and Mewhort D.J.K (2007), Representing word meaning and order information in a composite holographic lexicon, Psychological Review, 114, pp. 1-37*

The Random Permutation Model

- Use permutation functions to randomly shuffle the vectors to be composed*

$$p = M\,x + M^2 y$$

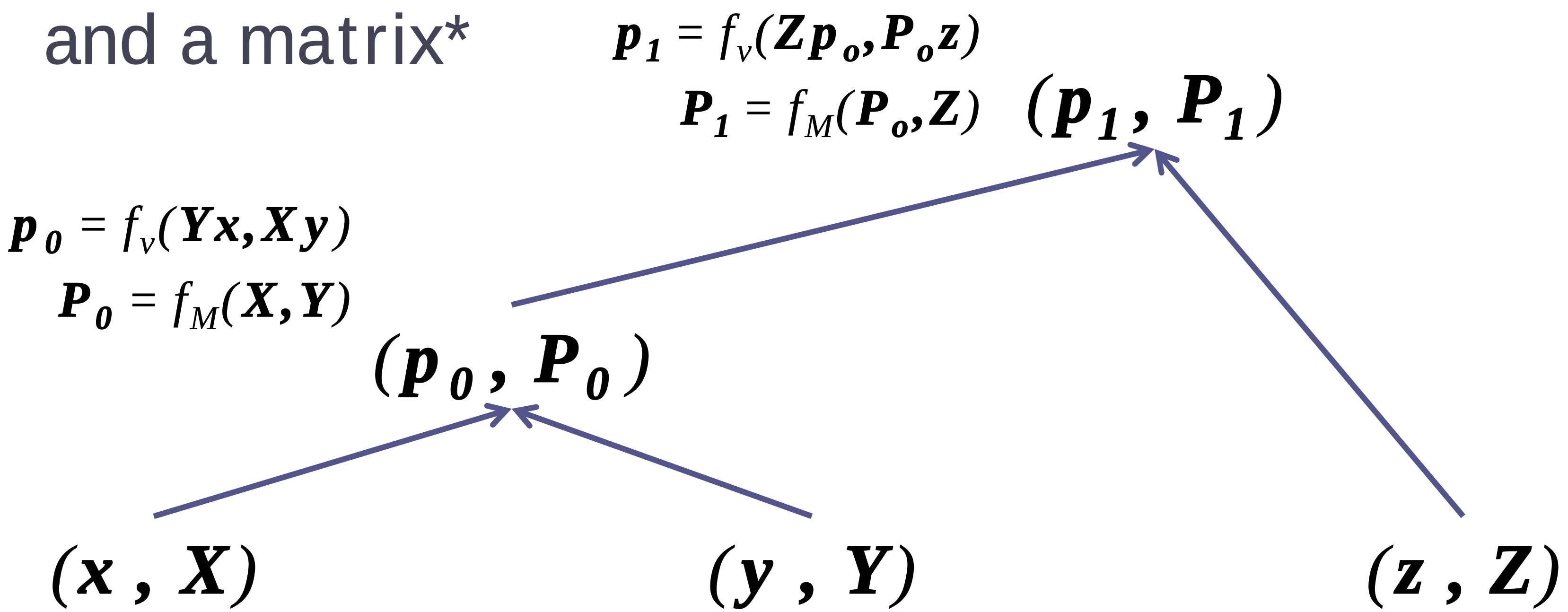
Random Permutation Operator



* *Sahlgren M., Holst A. and Kanerva P. (2008), Permutations as a means to encode order in word space, in Proceedings of the 30th Annual Conference of the Cognitive Science Society, pp. 1300-1305*

Recursive Matrix Vector Spaces

- Each word and phrase is represented by a vector and a matrix*



** Socher R., Huval B., Manning C.D., Ng A.Y. (2012), Semantic Compositionality through Recursive Matrix-Vector Spaces, in Proceedings of Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pp. 1201-1211*

Section 4

Future Research and Applications

- Current limitations of vector space models
- Encoding word position information into vectors
- **From vectors and matrices to tensors**
- Final remarks and conclusions

Union/Intersection Limited Binding

- Multiplicative operations limit vector interaction to those common non-zero components only

$$[\cancel{1}, 0, 3, 0, 1, 0] \times [0, \cancel{2}, 1, 0, 4, 0] = [0, 0, 3, 0, 4, 0]$$

- Additive operations limit vector interaction to both common and non-common non-zero components

$$[1, 0, 3, 0, 1, 0] + [0, 2, 1, 0, 4, 0] = [1, 2, 3, \overset{?}{0}, 4, \overset{?}{0}]$$

- Can we define operations to model richer interactions across vector components?

Vector Binding with Tensor Product*

- The tensor product of two vectors

$$\boldsymbol{a} \otimes \boldsymbol{b} = \{ a_i b_j \} \;\; for\; i = 1, 2 \ldots N_a \; and \;\; j = 1, 2 \ldots N_b$$

- All possible interactions across components are taken into account
- **But, the resulting vector representation is of higher dimensionality!**

*** Smolensky P. (1990), Tensor product variable binding and the representation of symbolic structures in connectionist systems, Artificial Intelligence, 46, pp.159-216**

Compressing Tensor Products

- Compress the result to produce a composed representations with the same dimensionality of the original vector space
- One representative example of this is the ***circular convolution model***
- ***Can tensor representations be exploited at high dimensional space?***

Section 4

Future Research and Applications

- Current limitations of vector space models
- Encoding word position information into vectors
- From vectors and matrices to tensors
- **Final remarks and conclusions**

VSMs in Monolingual Applications

Vector Space Models have been proven useful for many monolingual NLP applications, such as:

- Clustering
- Classification
- Information Retrieval
- Question Answering
- Essay grading
- Spelling Correction
- Role Labeling
- Sense Disambiguation
- Information Extraction
- *and so on...*

VSMs in Cross-language Applications

Vector Space Models are also starting to be proven useful for cross-language NLP applications:

- Cross-language information retrieval
- Cross-script information retrieval
- Parallel corpus extraction and generation
- Automated bilingual dictionary generation
- Machine Translation (decoding and evaluation)
- Cross-language plagiarism detection

Future Research

Seems to be moving in two main directions:

- Improving the representation capability of current VSM approaches by:
  - Using neural network architectures
  - Incorporating word order information
  - Leveraging on more complex operators
- Developing a more comprehensive framework by combining formal and distributional approaches

Section 4

Main references for this section

- T. K. Landauer S. T. and Dumais S.T. , 1997, “A solution to Plato’s problem: the latent semantic analysis theory of acquisition, induction and representation of knowledge”
- J. Mitchell and M. Lapata, 2008, “Vector-based models of semantic composition”
- M. N. Jones and D. J. K. Mewhort, 2007, “Representing word meaning and order information in a composite holographic lexicon”
- M. Sahlgren, A. Holst and P. Kanerva, 2008, “Permutations as a means to encode order in word space”

Section 4

Additional references for this section

- Montague R. (1970), Universal Grammar, Theoria, 36, pp. 373-398
- Landauer T.K. (2002), On the computational basis of learning and cognition: Arguments from LSA, in Ross B.H. (ed.) The Psychology of Learning and Motivation: Advances in Research and Theory, 41, pp. 43-84
- Socher R., Huval B., Manning C.D., Ng A.Y. (2012), Semantic Compositionality through Recursive Matrix-Vector Spaces, in Proceedings of Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pp. 1201-1211
- Smolensky P. (1990), Tensor product variable binding and the representation of symbolic structures in connectionist systems, Artificial Intelligence, 46, pp.159-216

Vector Spaces for Cross-Language NLP Applications

Rafael E. Banchs

Human Language Technology Department,
Institute for Infocomm Research, Singapore

INTERSPEECH 2015
September 6 - 10
Dresden, Germany