DISTRIBUTIONAL MODELS OF LEXICAL SEMANTICS

Web Mining & Information Retrieval a.a. 2013/2014
Prof. Roberto Basili

A distributional perspective of lexical semantics

- Distributional Hypothesis (Harris, 1964): The meaning of a word can be described by the set of its textual context :

Words with similar meanings will occur with similar neighbors if enough text material is available [Schutze and Pedersen(1995)]

- IDEA: acquire an artificial representation of a target word w, considering the **distribution** of all other words that co-occur with w,

    - two words sharing the same co-occurrences will be represented in a similar manner.

    - words are mapped into vectors expressing their corresponding contexts in the corpus

    - The similarity among words is estimated measuring the distance in the space of their vector representations.

- GOAL: design word vectors able to represent in a meaningful fashion the semantics of words

What kind of relation are we interested in? (1)

- **Topical relations**: Two words involved in a topical relation refers to a common topic (eg. Economy or Sport)

- **Syntagmatic relations** concern *positioning*, and relate entities that co-occur in the text;
  - it is a relation in *praesentia*.
  - This relation is a linear one, and applies to linguistic entities that *occur in sequential combinations*.
  - One example is represented by words that occur in a sequence, as in a normal sentence like "*the wolf is hungry*."
  - A syntagm is such an ordered combination of linguistic entities. For example, written words are syntagms of letters, sentences are syntagms of words, and paragraphs are syntagms of sentences.

What kind of relation are we interested in? (2)

- **Paradigmatic relations** concern *substitution*, and relate entities that do not co-occur in the text;
  - it is a relation in *absentia*.
  - Paradigmatic relations hold between linguistic entities that occur in the same context but not at the same time, like the words "hungry" and "thirsty" in the sentence "the wolf is [hungry|thirsty]".
  - Paradigmatic relations are substitutional relations, which means that linguistic entities have a paradigmatic relation when the choice of one excludes the choice of another.
  - A paradigm is thus a set of such substitutable entities.

What's the role of different word spaces?

- **Topic space** [Salton et al.(1975)] captures topical relations:
  - A document-based space, i.e. the context is an entire document
  - Words appearing in the same documents have a similar representation
  - individual score is computed according the TF-IDF schema

- **Co-occurrence word-based space** [Sahlgren(2006)] captures paradigmatic relations:
  - Contexts are words, as lemmas, appearing in a n-length window
  - Individual scores are computed according to the Point-wise Mutual Information (PMI) over the co-occurrence frequency
  - The window width is a parameter allowing the space to capture different aspects

- **Co-occurrence syntax-based space** [Pado and Lapata(2007)] captures paradigmatic relation (constrained by syntax)
  - Contexts words are enriched through information about syntactic relations

Co-occurrence word space
An Example

VerbNet (VN) (Kipper-Schuler 2006) is the largest on-line verb lexicon currently available for English. It is a hierarchical domain-independent, broad-coverage verb lexicon with mappings to other lexical resources such as WordNet (Miller, 1990; Fellbaum, 1998), Xtag (XTAG Research Group, 2001), and FrameNet (Baker et al., 1998). VerbNet is organized into verb classes extending Levin (1993) classes through refinement and addition of subclasses to achieve syntactic and semantic coherence among members of a class. Each verb class in VN is completely described by thematic roles, selectional restrictions on the arguments, and frames consisting of a syntactic description and semantic predicates with a temporal function, in a manner similar to the event decomposition of Moens and Steedman (1988).

Example – POS tagging

VerbNet::NNP (::(VN::NNP)::) (::(Kipper-Schuler::JJR 2006::CD)::) is::VBZ the::DT largest::JJS on-line::JJ verb::NN lexicon::NN currently::RB available::JJ for::IN English::NNP .::.

It::PRP is::VBZ a::DT hierarchical::JJ domain-independent::JJ ,::, broad-coverage::JJ verb::NN lexicon::NN with::IN mappings::NNS to::TO other::JJ lexical::JJ resources::NNS such::JJ as::IN WordNet::NNP (::(Miller::NNP ,::, 1990::CD ;::: Fellbaum::NNP ,::, 1998::CD)::) ,::, Xtag::NNP (::(XTAG::NNP Research::NNP Group::NNP ,::, 2001::CD)::) ,::, and::CC FrameNet::NNP (::(Baker::NNP et::CC al::NNP .::.

VerbNet::NN is::VBZ organized::VBN into::IN verb::NN classes::NNS extending::VBG Levin::NNP (::(1993::CD)::) classes::NNS through::IN refinement::NN and::CC addition::NN of::IN subclasses::NNS to::TO achieve::VB syntactic::JJ and::CC semantic::JJ coherence::NN among::IN members::NNS of::IN a::DT class::NN .::.

Each::DT verb::NN class::NN in::IN VN::NNP is::VBZ completely::RB described::VBN by::IN thematic::JJ roles::NNS ,::, selectional::JJ restrictions::NNS on::IN the::DT arguments::NNS ,::, and::CC frames::NNS consisting::VBG of::IN a::DT syntactic::JJ description::NN and::CC semantic::JJ predicates::NNS with::IN a::DT temporal::JJ function::NN ,::, in::IN a::DT manner::NN similar::JJ to::TO the::DT event::NN decomposition::NN of::IN Moens::NNP and::CC Steedman::NNP (::(1988::CD)::) .::.

Example: lexicon::NN

VerbNet::NNP (::(VN::NNP)::) (::(Kipper-Schuler::JJR 2006::CD)::) is::VBZ the::DT largest::JJS on-line::JJ verb::NN lexicon::NN currently::RB available::JJ for::IN English::NNP .::.

It::PRP is::VBZ a::DT hierarchical::JJ domain-independent::JJ ,::, broad-coverage::JJ verb::NN lexicon::NN with::IN mappings::NNS to::TO other::JJ lexical::JJ resources::NNS such::JJ as::IN WordNet::NNP (::(Miller::NNP ,::, 1990::CD ;::: Fellbaum::NNP ,::, 1998::CD)::) ,::, Xtag::NNP (::(XTAG::NNP Research::NNP Group::NNP ,::, 2001::CD)::) ,::, and::CC FrameNet::NNP (::(Baker::NNP et::CC al::NNP .::.

VerbNet::NN is::VBZ organized::VBN into::IN verb::NN classes::NNS extending::VBG Levin::NNP (::(1993::CD)::) classes::NNS through::IN refinement::NN and::CC addition::NN of::IN subclasses::NNS to::TO achieve::VB syntactic::JJ and::CC semantic::JJ coherence::NN among::IN members::NNS of::IN a::DT class::NN .::.

Each::DT verb::NN class::NN in::IN VN::NNP is::VBZ completely::RB described::VBN by::IN thematic::JJ roles::NNS ,::, selectional::JJ restrictions::NNS on::IN the::DT arguments::NNS ,::, and::CC frames::NNS consisting::VBG of::IN a::DT syntactic::JJ description::NN and::CC semantic::JJ predicates::NNS with::IN a::DT temporal::JJ function::NN ,::, in::IN a::DT manner::NN similar::JJ to::TO the::DT event::NN decomposition::NN of::IN Moens::NNP and::CC Steedman::NNP (::(1988::CD)::) .::.

Example

VerbNet::NNP (::(VN::NNP)::) (::(Kipper-Schuler::JJR 2006::CD)::) is::VBZ the::DT largest::JJS on-line::JJ verb::NN lexicon::NN currently::RB available::JJ for::IN English::NNP .::.

It::PR~ ~::VBZ a::DT hierarchical::JJ domain-independent::JJ ,::, broad-coverage::JJ verb::NN lexicon::NN with::IN ~ppings::NNS to::TO other::JJ lexical::JJ resources::NNS such::JJ as::IN WordNet::NNP (::(Miller::NN~ ~, 1990::CD ;::: Fellbaum::NNP ,::, 1998::CD)::) ,::, X~~~NNP (::(XTAG::NNP Research::NN~ ~Group::NNP ,::, 2001::CD)::) ,::, and::CC Fr~~~Net::NNP (::(Baker::NNP et::CC al::NNP .::.

VerbNet::NN is::Vb~ organized::VBN into::IN v~~~NN classes::NNS extending::VBG Levin::NNP (::(1993::CD)::)   ~C addition::NN of::IN subclasses::NNS to::TO achiev~   ~ce::NN among::IN members::NNS of::IN a::DT class::NN .::.

Left context – windows 2

Each::DT verb::NN class::NN in::IN VN::NNP is::VBZ completely::RB described::VBN by::IN thematic::JJ roles::NNS ,::, selectional::JJ restrictions::NNS on::IN the::DT arguments::NNS ,::, and::CC frames::NNS consisting::VBG of::IN a::DT syntactic::JJ description::NN and::CC semantic::JJ predicates::NNS with::IN a::DT temporal::JJ function::NN ,::, in::IN a::DT manner::NN similar::JJ to::TO the::DT event::NN decomposition::NN of::IN Moens::NNP and::CC Steedman::NNP (::(1988::CD)::) .::.

Example

VerbNet::NNP (::(VN::NNP)::) (::(Kipper-Schuler::JJR 2006::CD)::) is::VBZ the::DT largest::JJS on-line::JJ verb::NN lexicon::NN currently::RB available::JJ for::IN English::NNP .::.
It::PRP is::VBZ a::DT hierarchical::JJ domain-independent::JJ ,::, broad-coverage::JJ verb::NN lexicon::NN with::IN mappings::NNS to::TO other::JJ lexical::JJ resources::NNS such::JJ as::IN WordNet::NNP (::(Miller::NNP ::, 1990::CD ;::: Fellbaum::NNP ,::, 1998::CD)::) ,::, Xtag::NNP (::(XTAG::NNP Research::NNP Group::NNP ,::, 2001::CD)::) ,::, and::CC FrameNet::NNP (::(Baker::NNP et::CC al::NNP .::.
VerbNet::NN is::VBZ organized::VBN into::IN verb::NN classes::NNS extending::VBG Levin::NNP (::(1993::CD)::) classes::NNS through::IN refinement::NN and::CC addition::NN of::IN subclasses::NNS to::TO achieve::VB syntactic::JJ and::CC semantic::JJ coherence::NN among::IN members::NNS of::IN a::DT class::NN .::.
Each::DT verb::NN class::NN in::IN VN::NNP is::VBZ completely::RB described::VBN by::IN thematic::JJ roles::NNS ,::, selectional::JJ restrictions::NNS on::IN the::DT arguments::NNS ,::, and::CC frames::NNS consisting::VBG of::IN a::DT syntactic::JJ description::NN and::CC semantic::JJ predicates::NNS with::IN a::DT temporal::JJ function::NN ,::, in::IN a::DT manner::NN similar::JJ to::TO the::DT event::NN decomposition::NN of::IN Moens::NNP and::CC Steedman::NNP (::(1988::CD)::) .::.

Right context – windows 2

Example

- The word space is expressed by a co-occurrence matrix M
  - Rows: The target words occurring more than a *t(hreshold)* are selected (e.g 200)
  - Columns : The C most frequent word-context are selected (e.g. 20,000)
  - Each matrix item is the co-occurrence frequency between the target word and contextual word

- Example: the word *lexicon::N* occurs with
  - *verb::N* Left (feat 8)   2
  - *with::IN* Right (feat 25)   1
  - *available::J* Right (feat 56)   1
  - *online::J* Left (feat 78)   1
  - ...

- It will be represented by the frequency vector
  - 8:2 25:1 56:1 78:1 98:1 110:1 137:1

Pointwise Mutual Information (PMI)

- Context with high frequency (e.g. stopwords) have higher score

- PMI is a commonly used metric in Information Theory [Fano, 1961] for measuring this strength of association between two events *x* and *y*.

$$I(x,y) = \log_2 \frac{P(x,y)}{P(x)P(y)}$$

P(x)= probability of *x*

P(y)= probability of *y*

P(x,y)= joint probability of *x* e *y*

- Two words *x* e *y* that often co-occur (respect to their occurrence) show a high degree of association

- Words with high frequency are penalized

Pointwise Mutual Information (PMI)

□ The previous definition is adapted [Church and Hanks, 1989] to our word-occurrence problem:

  ▫ P(x) = probability of the word *x* inside a corpus

  ▫ P(y) = probability of the word *y* inside a corpus

  ▫ P(x,y) = probability that *x* co-occur with *y*

□ This probability is estimated through the **Maximum Likelihood Estimation:**

$$I(x, y) \approx \log_2 \frac{\dfrac{c_{xy}}{N}}{\dfrac{c_x}{N} \times \dfrac{c_y}{N}}$$

c_x = number of occurence of *x*

c_{xy} = number of co-occurence of *x* and *y*

N = total number of token

PMI

- The PMI between lexicon::N and verb::N

    - c_x: lexicon::N occurrs 2 times

    - c_y: verb::N occurrs 4 times

    - c_{xy}: 2 co-occurences (left side)

    - N: 142 tokens

    - PMI=5,14

$$I(x, y) \approx \log_2 \frac{\dfrac{c_{xy}}{N}}{\dfrac{c_x}{N} \times \dfrac{c_y}{N}}$$

- Vectors are then normalized to be comparable

The resulting matrix

- Matrix with t=2 – C=100
- It can be sparse and affected by the "*curse of dimensionality*"

|	and::CC R	and::CC L	a::DT R	a::DT L	verb::N R	verb::N L	be::V R	be::V L	class::N R	of::IN R	class::N L	of::IN L	lexicon::N R	verbnet::N L	v
and::CC:	0	0	0	0	0	0	0	0	0	0,142	0	0,142	0	0	
a::DT:	0	0	0	0	0	0	0	0,155	0,155	0	0	0,210	0	0	
verb::N:	0	0	0	0	0	0	0	0	0,244	0	0	0	0,302	0	
be::V:	0	0	0,174	0	0	0	0	0	0	0	0	0	0	0,255	
of::IN:	0,147	0,147	0,219	0	0	0	0	0	0,180	0	0	0	0	0	
class::N:	0	0	0,000	0,184	0	0,271	0	0	0	0	0	0,205	0	0	
the::DT:	0	0	0	0	0	0	0	0,214	0	0	0	0	0	0	
to::TO:	0	0	0	0	0	0	0	0	0	0	0	0,200	0	0	
in::IN:	0	0	0,295	0	0	0,320	0,320	0	0	0	0,320	0	0	0	
xtag::N:	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
lexicon::N:	0	0	0	0	0	0,331	0	0	0	0	0	0	0	0	
syntactic::J:	0,344	0	0	0,289	0	0	0	0	0	0	0	0,313	0	0	
with::IN:	0	0	0,259	0	0	0,280	0	0	0	0	0	0	0	0	
semantic::J:	0	0,304	0	0	0	0	0	0	0	0	0	0	0	0	

Latent Semantic Analysis

☐ In LSA approach, SVD is applied to source co-occurrence matrix:

$$M = USV^T \approx U_k S_k V_k^T$$

Target terms space

$$U_k S_k^{\frac{1}{2}}$$
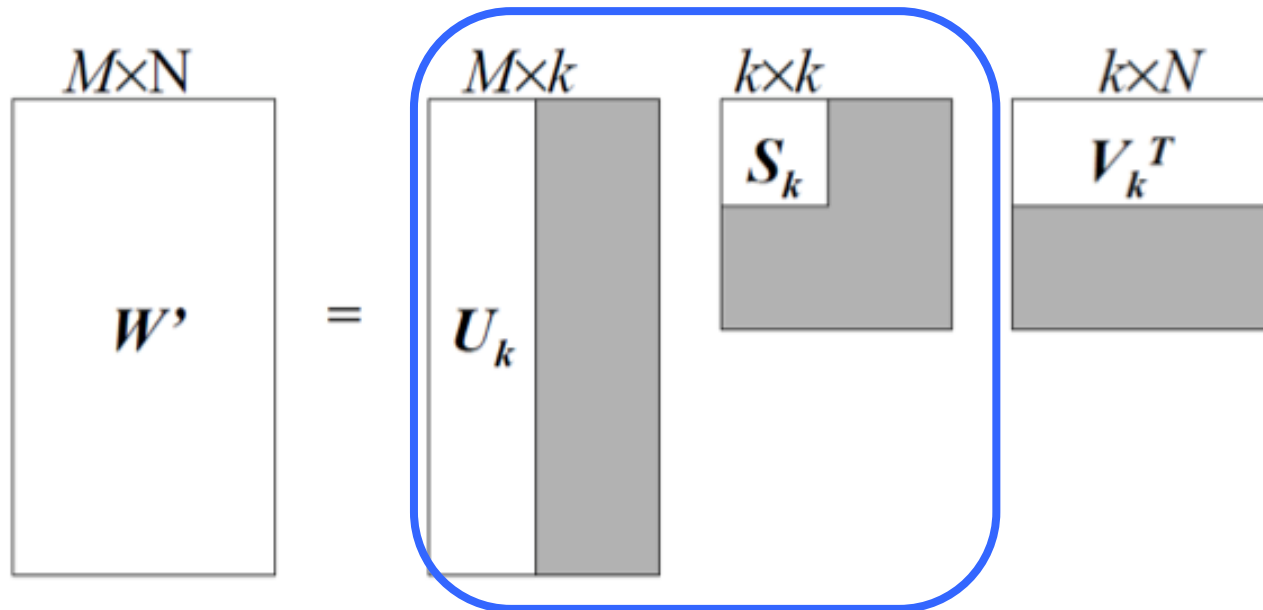
Latent Semantic Analysis 2

- Minimize the global reconstruction error

- Reduce noise over the data distribution

- SVD let the principal components of the distribution emerge (max covariance)

- Principal components are linear combinations of the original dimensions, i.e. pseudo concepts, as captured in the entire space

- Capture second order relations among targets words

Word spaces: clustering and classification

- This geometrical representation is suitable for several learning algorithm
  - Unsupervised learning
    - clustering of verbs that show similar behaviour
  - Supervised Learning
    - Classification of verbs among the verb classes
    - Selection of Contexts that better represent classes
  - Semi-supervised learning