From Latent Semantic Spaces to Word spaces:
Distributional Models of Lexical Semantics

Web Mining & Information Retrieval a.a. 2022/2023
Roberto Basili

A distributional perspective on 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::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::N

Left context – windows 2

N classes::NNS extending::VBG Levin::NNP

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::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 su … c::JJ and::CC semantic::JJ coherence::NN among::IN members::NNS of::IN a::DT class::NN .::.

Right 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

- 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 target word *lexicon::N* (in row) occurs with (columns)
  - *verb::N* Left (feature 8) 2
  - *with::IN* Right (feature 25) 1
  - *available::J* Right (feature 56) 1
  - *online::J* Left (feature 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**:

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

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

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

- Vectors are then normalized to be comparable

The resulting matrix W

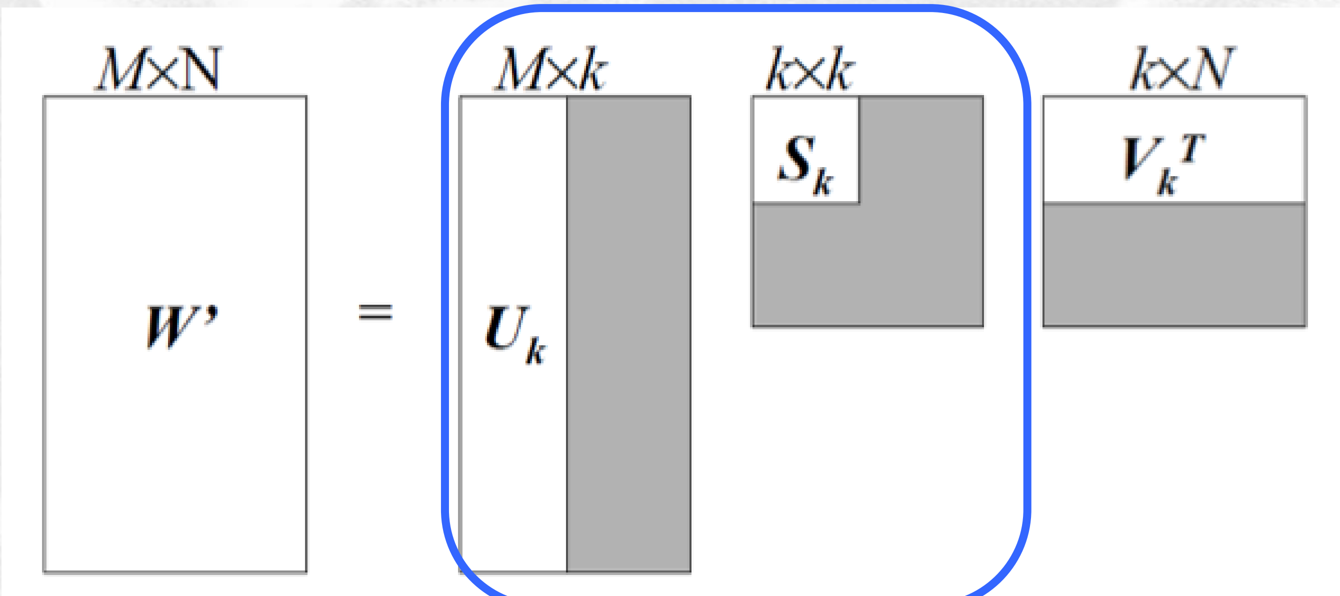
- Matrix with t=2 and C=100

	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
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, SVD *(Golub & Kahan 1965)* is applied to source co-occurrence matrix:

$W\sqrt{S_k}$ $\quad W = USV^T \approx W' = U_k S_k V_k^T \xrightarrow{\phi} U\sqrt{S_k}$



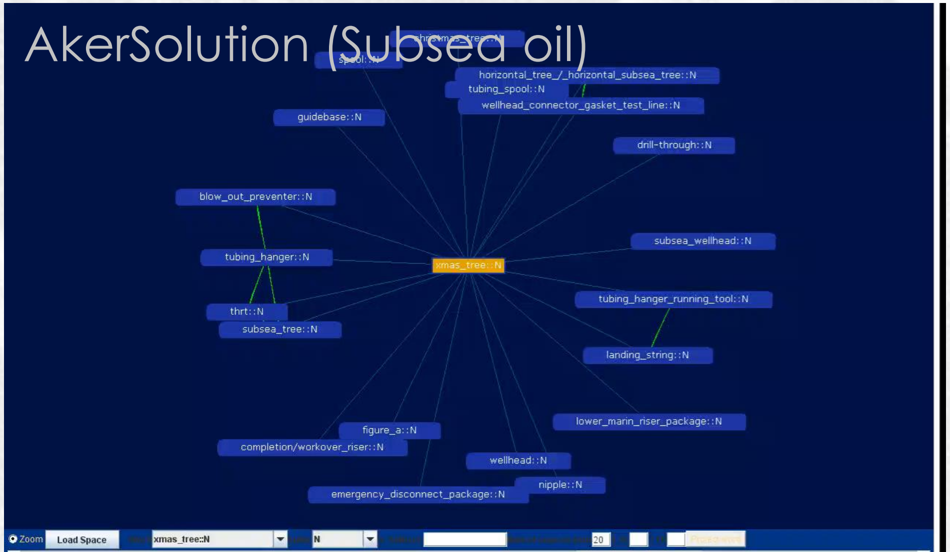
Latent Semantic Analysis (2)

- 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

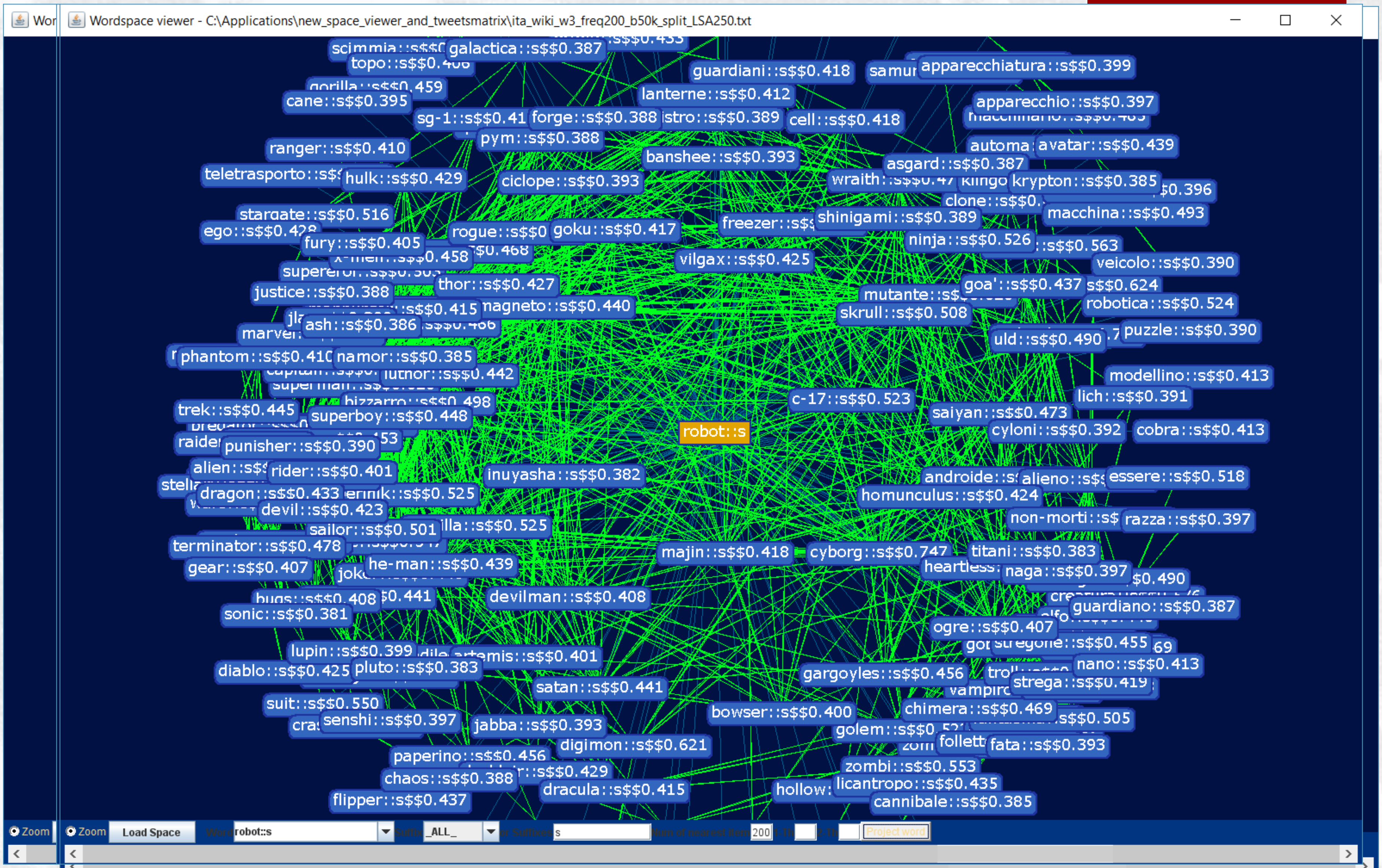
Results

- A new truncated matrix $U_k S_k^{1/2}$ by which representing information about *lexical entries* (i.e. the rows of W) such as:
  - *lexicon*::N
  - *verb*::N
  - …
- These vectors are representative of
  - Paradigmatic (*company* vs. *enterprise*, *rat* vs. *mouse*)
  - Topical (*company* vs. *market*, *triangle* vs. *geometry*, …)
  - Associative (*company* vs. *investments*, *triangle* vs. *perimeter*, …)
- … relations according to varying sizes of the context window [Schutze and Pedersen(1995)] [Sahlgren(2006)] [P. D. Turney and P. Pantel (2010)] [Croce et al., 2019]

Latent Semantic Spaces:
Encoding & Domain Corpora



Lexical Acquisition on the Web



Word spaces: clustering and classification

- This geometrical representation is suitable as a basic representation for several learning algorithms
  - Unsupervised learning
    - clustering of verbs that show similar behaviour (e.g. a process model)
  - Supervised Learning
    - Classification of words among semantic classes (e.g. Frame rec.)
    - Selection of Contexts that better represent classes
    - Initialization for Neural Networks: embedding lexical input features
  - Overall Semi-supervised learning
    - Language-specific representations
    - Pre-Training for complex multitask (neural) models, e.g. LSTM or CNNS and encoders input

Recap

- Documents are traditionally represented through a bag-of-word model where individual words play the role of independent axes of the space where documents are lying
- Documents are thus column vector of weights in a M dimensional space, whereas M is the dimension of the vocabulary
- Terms (i.e. words) are (row) vectors in N dimensional spaces, whereas N (>> M) is the number of different documents

Recap (2)

- Two terms are similar is their *n*-dimensional vectors have an high value of the cosine similarity ... but
- ... this DO mean that they share documents, i.e. they must occurr in a large number of documents
- As a result word senses (e.g. multiple meanings of the same term) do not influence document modeling as well as term similarity estimation
- This is not capturing the different role word meanings play in a document
- IDEA: find a space whereas word senses are bettere expressed. We call this spaces *latent semantic spaces*
- HOW:
  - 1. Describe words through their local co-occurrence with other words in sentences of a large corpus. The first words are called targets, while the second words are the contextual words (or features)
  - The resulting target word-by-context word matrix W has targets in rows and contexts in columns

Recap (3)

- HOW (continued)
  - 3. Apply to the obtained M×N matrix W, the Singular Value Decomposition as a search for the latent structure of the space underlying the dcument collection
    - It extracts eigenvalues (i.e. eigenspaces of the term co-occurrence statistics) that are dimensions of maximal covariance of W
    - Truncated SVD transformations approximate W with a W'. They allow to maintain limited the number of dimensions (usually *k*) employed to represents target term vectors
  - 4. Compile individual *k*-dimensional semantic representations of the target terms into a general and reusable dictionary, called embedding lexicon
  - Apply learning tasks to the obtained lexicon:
    - Term Clustering: looking for wor classes as clusters of tearget term vectors
    - Term Classification: use word vectors to obtain a representaion of training documents (e.g. via weighted linear combinations) and train your classifier onto the labeled document vectors

Recap (4)

- Given the **unsupervised** nature of the SVD the target term vectors can be used as basic representations, called embeddings, for a variety of text processing tasks,
  - Semisupervised Document classification,
  - Question classification,
  - Sentiment Analysis
- Term vector are extacted without relying on any labeled data
  - They generalize word meanings and are better representations than the original, but uninterpreted, words

References

- Susan T. Dumais, Michael Berry, Using Linear Algebra for Intelligent Information Retrieval, SIAM Review, 1995, 37, 573–595
- Sahlgren, M. (2006): The Word-Space Model: *Using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces*. Ph.D. dissertation, Department of Linguistics, Stockholm University.
- Hinrich Schutze and Jan O. Pedersen. 1995. Information retrieval based on word senses. In Symposium on Document Analysis and Information Retrieval. [pdf]
- Hinrich Schutze, Automatic word sense discrimination, Computational Linguistics, 24(1), 1998.
- P. D. Turney and P. Pantel (2010) "From Frequency to Meaning: Vector Space Models of Semantics", JAIR, Volume 37, pages 141-188 [pdf].
- D. Croce, S. Filice, R. Basili, Making sense of kernel spaces in neural learning, Computer Speech & Language, Volume 58, 2019, Pages 51-75, ISSN 0885-2308,https://doi.org/10.1016/j.csl.2019.03.006.