Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

Alexander Panchenko

INDUCTION AND EMBEDDING OF LINGUISTIC STRUCTURES FROM TEXT

Word sense induction

A lexical sample WSI task

- **Target word**, e.g. “bank”.

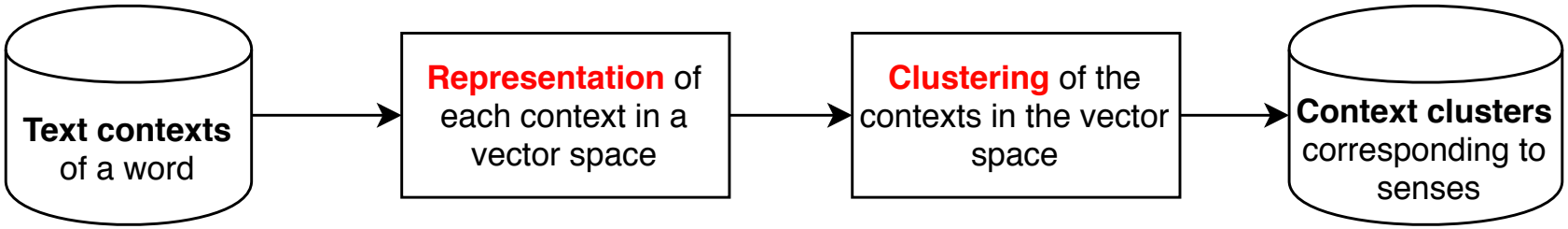
A lexical sample WSI task

- **Target word**, e.g. “bank”.
- **Contexts** where the word occurs, e.g.:
  - “river **bank** is a slope beside a body of water”
  - “**bank** is a financial institution that accepts deposits”
  - “Oh, the **bank** was robbed. They took about a million dollars.”
  - “**bank** of Elbe is a good and popular hangout spot complete with good food and fun”

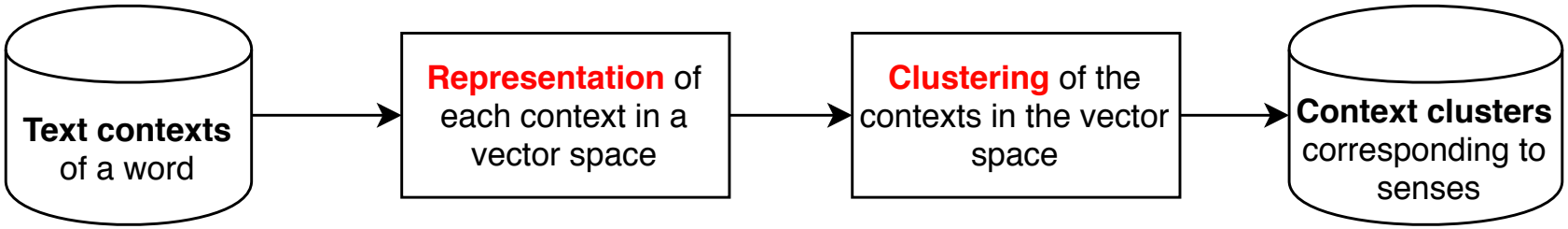
A lexical sample WSI task

- **Target word**, e.g. “bank”.
- **Contexts** where the word occurs, e.g.:
  - “river **bank** is a slope beside a body of water”
  - “**bank** is a financial institution that accepts deposits”
  - “Oh, the **bank** was robbed. They took about a million dollars.”
  - “**bank** of Elbe is a good and popular hangout spot complete with good food and fun”
- You need to **group the contexts by senses**:
  - “river **bank** is a slope beside a body of water”
  - “**bank** of Elbe is a good and popular hangout spot complete with good food and fun”
  - “**bank** is a financial institution that accepts deposits”
  - “Oh, the **bank** was robbed. They took about a million dollars.”

Sense induction using clustering

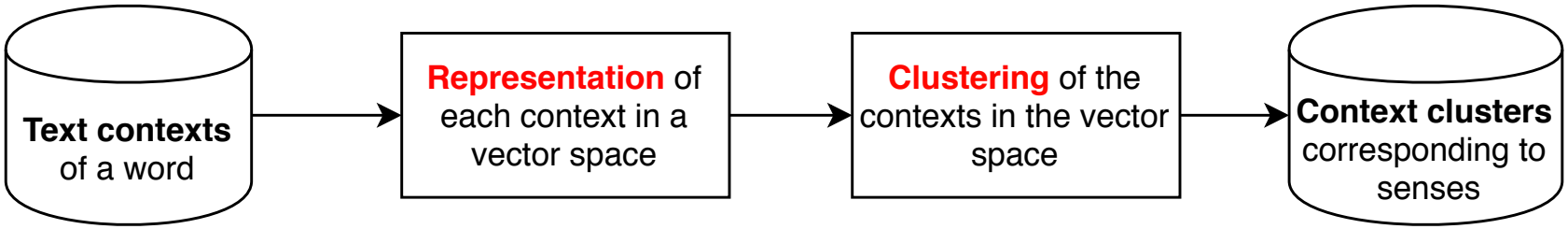


Sense induction using clustering



- **Representation**
  - Sparse vector model (TF-IDF, etc.)
  - Weighted (TF-IDF, χ^2, etc.) sum of word embeddings
  - Sentence embeddings (InterSent, Skip-Thougts, doc2vec, etc.)

Sense induction using clustering



- **Representation**
  - Sparse vector model (TF-IDF, etc.)
  - Weighted (TF-IDF, χ^2, etc.) sum of word embeddings
  - Sentence embeddings (InterSent, Skip-Thougts, doc2vec, etc.)
- **Clustering**
  - Affinity Propagation
  - Agglomerative Clustering
  - k-means

Sense induction using neighbors

1. **Get the neighbors** of a target word, e.g. **"bank"**:
   1. lender
   2. river
   3. citybank
   4. slope
   5. ...
2. Get **similar to "bank"** and **dissimilar to "lender"**:
   1. river
   2. slope
   3. land
   4. ...
3. **Compute distances** to **"lender"** and **"river"**.

Graph-vector sense induction

1. **For** i-th neighbor of the target word w among k neigbours:
   1. Get a pair of opposite words for the i neighbor: (w_j, w_k)
   2. Add them as as nodes: $V = V \cup \{w_j, w_k\}$
   3. Remember the pair as an anti-edge: $A = A \cup (w_j, w_k)$

Graph-vector sense induction

1. **For** i-th neighbor of the target word w among k neigbours:
   1. Get a pair of opposite words for the i neighbor: (w_j, w_k)
   2. Add them as as nodes: $V = V \cup \{w_j, w_k\}$
   3. Remember the pair as an anti-edge: $A = A \cup (w_j, w_k)$
2. **Build an ego network** $G = (V, E)$ of the word w:
   1. E are computed based on word similarities;
   2. E are pruned based on the anti-edge constraints: $E = E \setminus A$.

Graph-vector sense induction

1. **For** i-th neighbor of the target word w among k neigbours:
   1. Get a pair of opposite words for the i neighbor: (w_j, w_k)
   2. Add them as as nodes: $V = V \cup \{w_j, w_k\}$
   3. Remember the pair as an anti-edge: $A = A \cup (w_j, w_k)$
2. **Build an ego network** $G = (V, E)$ of the word w:
   1. E are computed based on word similarities;
   2. E are pruned based on the anti-edge constraints: $E = E \setminus A$.
3. **Cluster** the ego network of the word w.

Graph-vector sense induction

1. **For** i-th neighbor of the target word w among k neigbours:
   1. Get a pair of opposite words for the i neighbor: (w_j, w_k)
   2. Add them as as nodes: $V = V \cup \{w_j, w_k\}$
   3. Remember the pair as an anti-edge: $A = A \cup (w_j, w_k)$
2. **Build an ego network** $G = (V, E)$ of the word w:
   1. E are computed based on word similarities;
   2. E are pruned based on the anti-edge constraints: $E = E \setminus A$.
3. **Cluster** the ego network of the word w.
4. **Find cluster labels** by finding the central nodes in a cluster.

Graph-vector sense induction

- **Get the neighbors** of a target word, e.g. **“java”**:
  1. Python
  2. Borneo
  3. C++
  4. Sumatra
  5. Arabica
  6. Robusta
  7. Ruby
  8. JavaScript
  9. Bali
  10. ...

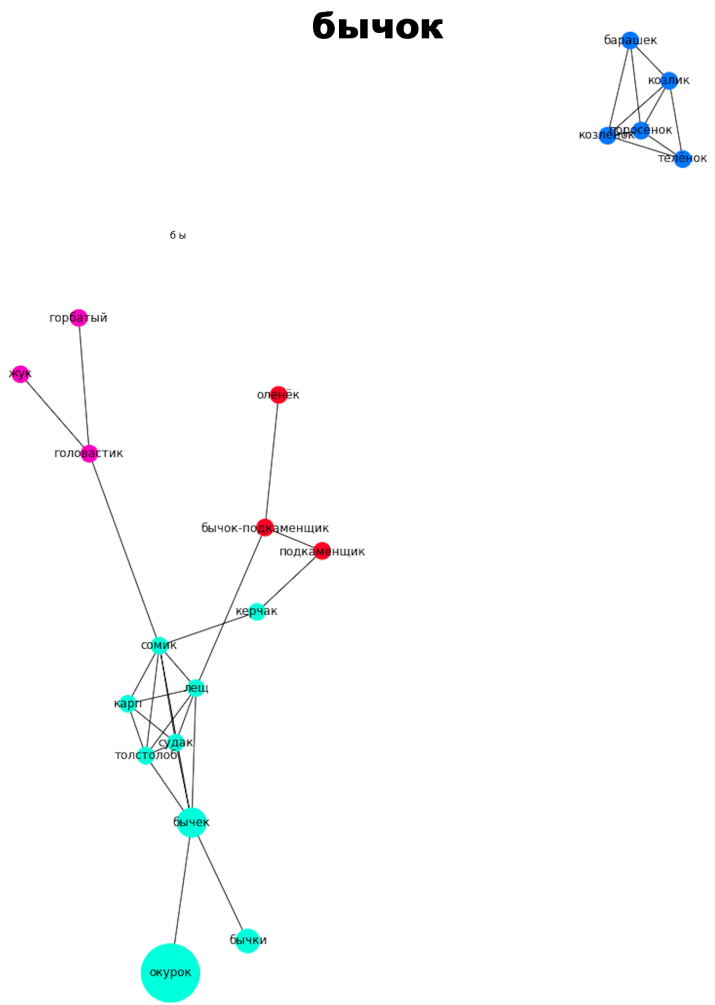
Graph-vector sense induction

- **Get the neighbors** of a target word, e.g. **"java"**:
  1. Python ≠ Borneo
  2. Borneo ≠ Scala
  3. C++ ≠ Borneo
  4. Sumatra ≠ highway
  5. Arabica ≠ Python
  6. Robusta ≠ Python
  7. Ruby ≠ Arabica
  8. Bali ≠ North

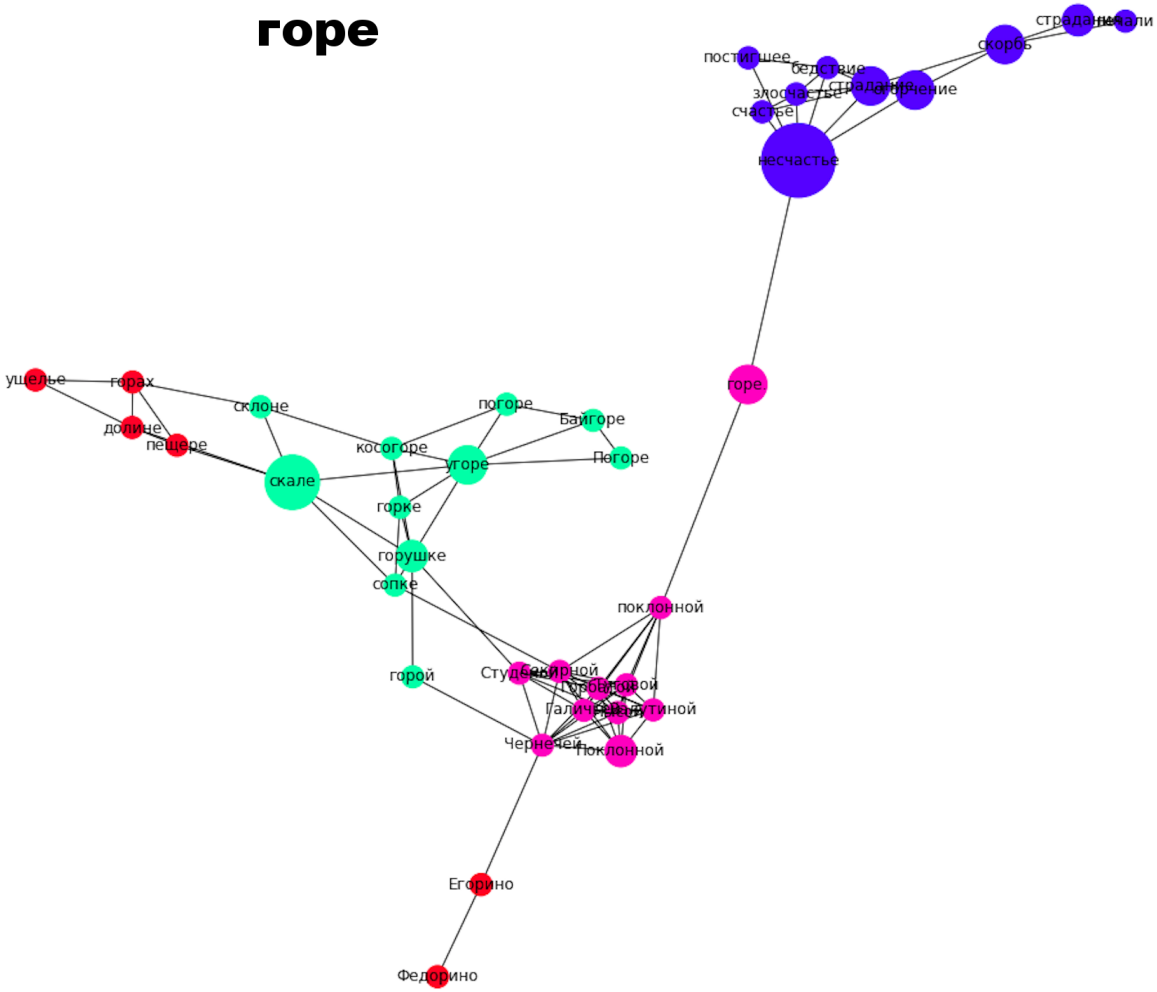
Graph-vector sense induction

- **Nodes**:
  1. Python
  2. Borneo
  3. C++
  4. Arabica
  5. Robusta
  6. Ruby

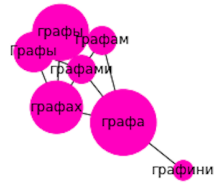
Sense induction example



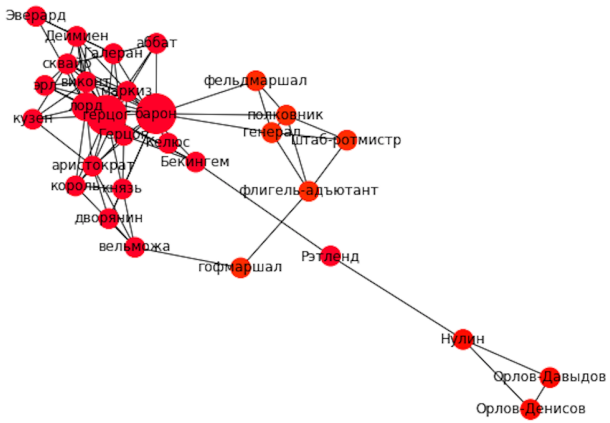
Sense induction example



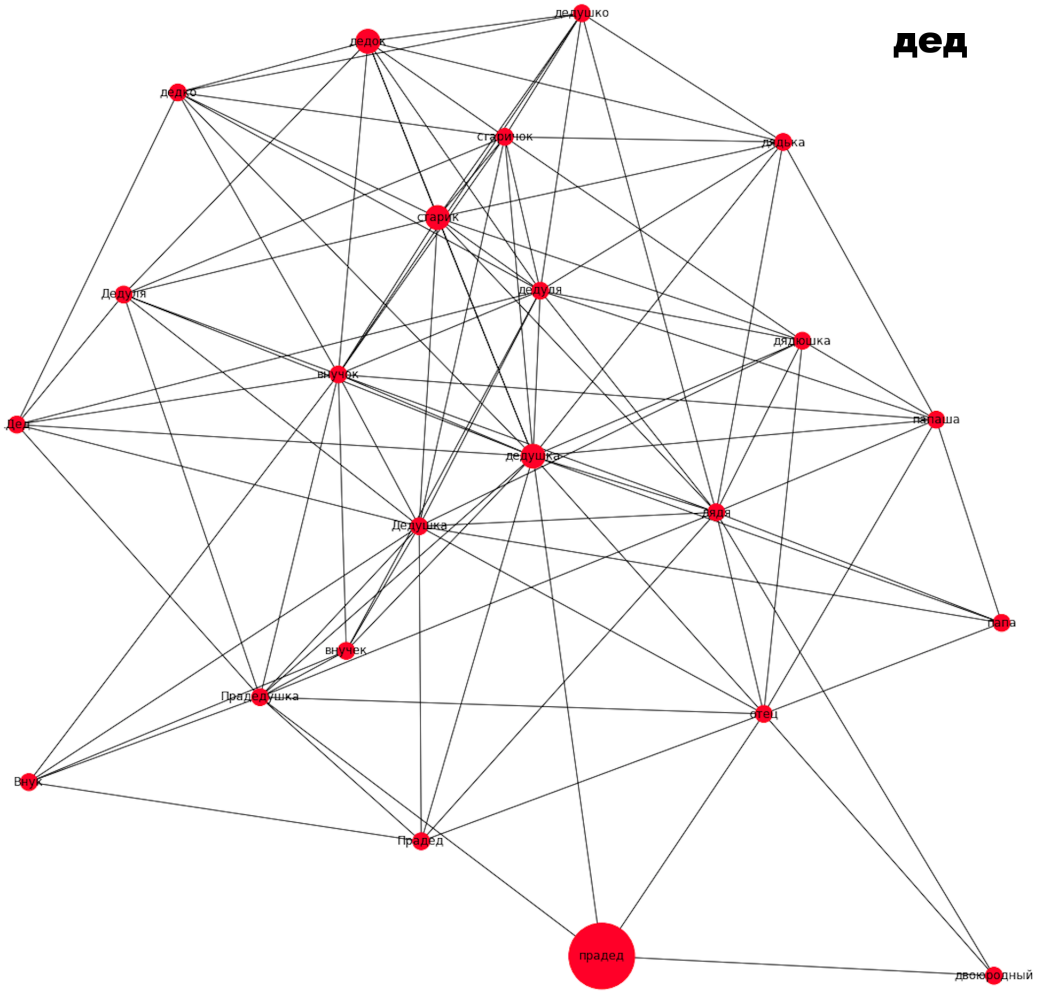
Sense induction example



граф



Sense induction example



Datasets

1. SemEval 2007
2. SemEval 2010
3. RUSSE 2018
4. **SemEval 2019 Task 2 Subtask 1**:
   - Clustering of verb occurrences
   - Assign occurrences of the target verbs to a number of clusters, in such a way that verbs belonging to the same cluster evoke the same frame type.
   - gold annotations for this subtask are based on FrameNet

Datasets

1. SemEval 2007
2. SemEval 2010
3. RUSSE 2018
4. **SemEval 2019 Task 2 Subtask 1**:
   - Clustering of verb occurrences
   - Assign occurrences of the target verbs to a number of clusters, in such a way that verbs belonging to the same cluster evoke the same frame type.
   - gold annotations for this subtask are based on FrameNet

- Trump **leads** the world, backward.
- Disrespecting international laws **leads** to many complications.
- Rosenzweig **heads** the climate impacts section at NASA's Goddard Institute.

Datasets

1. SemEval 2007
2. SemEval 2010
3. RUSSE 2018
4. **SemEval 2019 Task 2 Subtask 1**:
   - Clustering of verb occurrences
   - Assign occurrences of the target verbs to a number of clusters, in such a way that verbs belonging to the same cluster evoke the same frame type.
   - gold annotations for this subtask are based on FrameNet

- Trump **leads** the world, backward.
- Disrespecting international laws **leads** to many complications.
- Rosenzweig **heads** the climate impacts section at NASA's Goddard Institute.