Combining word embeddings as a tool for subject identification

Andreas Hamm
SC-IVS
Project: MeToDiO

Motivation

Typical tasks in automated document handling

- Given a subject, find documents related to that subject
- Given a document, find persons to whom the document might be relevant
- Given a set of documents, find which documents are related to each other

Test data

- Elib (DLR publication database) abstracts

Weighted word clouds

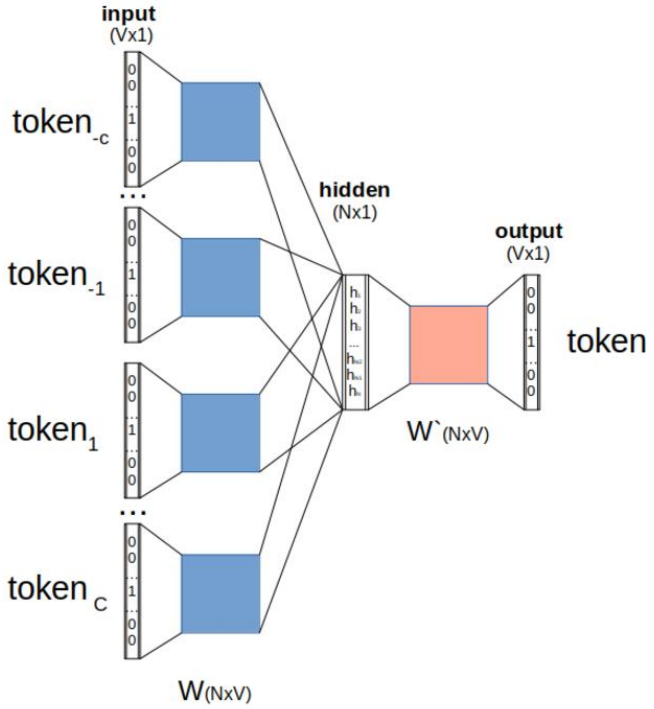
- Words and weights $\{((w_i, g_i) \in \mathcal{W} \times \mathbb{R} \mid i = 1, \ldots, k\}$
- Can characterize a subject
- Can capture the gist of a document
- Can outline common topics

Semantic word embeddings

- Embedding $\mathrm{emb}\colon \mathcal{W} \to \mathbb{R}^V$
- Turns word cloud into a geometric object
- Assess similarity and relatedness by geometric measures
- Presupposes semantic adequacy of embedding

Word2Vec (Mikolov et al., 2013)

- CBOW architecture
- Learns from the task of predicting target words from their context
- Pretrained on huge text collections from Google news
- 3 million word vocabulary
- Dimension V=300
- No vectors for out-of-vocabulary words



panel

The solar [yellow] *produces electricity.*

- Examples: Nearest words

Condensation	Venus	Aradish
Vapor	ERB	[OOV]
Dampness	Catherine	
Vapors	Jupiter	
Water vapor	Milton	
Mildew	Vinci	
Moisture	DAA	
Ice crystals	HOV	

- Examples: Arithmetic
  - **zoology – animal + flower ≈ botany**
  - **clustering – unsupervised + supervised ≈ clusters**

Word2Vec – trained on domain-specific corpus

- Trained on 16,896 entries in Elib
  - Last 6 years
  - Titles + English abstracts
- 18,837 word vocabulary

Observations

Pre-trained	Domain-trained
Convincing similarities and analogies for frequent words with unique meaning	Convincing similarities and analogies for domain-specific words
Specific words can be OOV	Because of small dictionary: many OOV words
Problems with homophones	

- Examples: Nearest words

Condensation	Venus	Aradish
Sublimation	Mercury	Arabidopsis
Flowing	Phobos	Thaliana
Droplets	Enceladus	Unicellular
Evaporation	Exoplanets	Kidney
Sand	Ryugu	Viruses
TGO	Jupiter	Halobacterium
Helium	Planet	Neon

- Examples: Arithmetic
  - **zoology – animal + flower ≈ [OOV]**
  - **clustering – unsupervised + supervised ≈ classifier**

fastText (Bojanowski et al., 2017)

- Similar CBOW architecture
- But break down target words into subwords (character n-grams)

  word: <panel>

  3-grams: <pa, pan, ane, nel, el>
- Learn vector representations of subwords
- Form vectors for words as average of subword vectors

- Pretrained on huge text collections from Wikipedia and CommonCrawl
- 2 million word vocabulary
- Also vectors for out-of-vocabulary words

- Examples: Nearest words

Condensation	Venus	Aradish
Condesation	Fly-trap	QQFZAAEACwAAAAAG
Condensations	Jupiter	KellerTitusTOAM06TO
Evaporation	Neptune	DEky4M0BSpUOTPnSp
Dampness	Venusian	KitsAeroBedAgfaphoto
Vapor	Uranus	DEky4M0BSpUOTPnSp
Vapour	Mars	MobileCOOLPADCUBO
Condensing	Aphrodite	QQJCgAEACwJAAAAEA

- Examples: Arithmetic
  - **zoology – animal + flower ≈ botany**
  - **clustering – unsupervised + supervised ≈ clustered**

fastText – trained on domain-specific corpus

- Trained on 16,896 entries in Elib
  - Last 6 years
  - Titles + English abstracts
- 18,837 word vocabulary

Observations

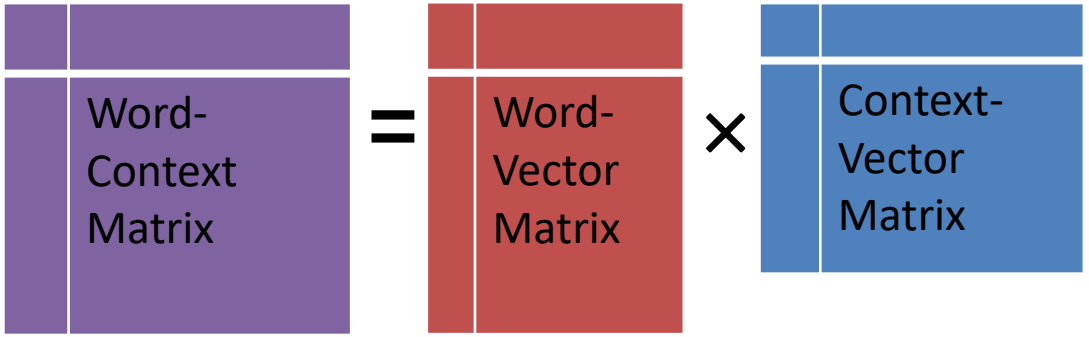
Pre-trained	Domain-trained
Can handle OOV words	Can handle OOV words
Specific words can be near to non-sensical words	Morphology has stronger influence than semantics
Mostly convincing similarities and analogies for frequent words with unique meaning	Mostly convincing similarities and analogies for domain-specific words

- Examples: Nearest words

Condensation	Venus	Aradish
Conjugation	Mercury	Radish
Conflation	Janus	Radiotherapy
Continuation	Genus	Radicals
Permeation	Phobos	Arcadia
Vaporization	Ryugu	Radiances
Supersaturation	Mars	Radios
Conversation	Moonmilk	Radii

- Examples: Arithmetic
  - **zoology – animal + flower ≈ phenomenology**
  - **clustering – unsupervised + supervised ≈ blistering**

GloVe (Pennington et al., 2014)



- Decomposition of word-context co-occurrence matrix
  - Approximation using a weighted least square regression
- Pretrained on huge text collections from Wikipedia + Gigaword + CommonCrawl
- 400,000 words in vocabulary

- Quality comparable to Word2Vec

- Examples: Nearest words

Condensation	Venus	Aradish
Vapor	Serena	[OOV]
Evaporation	Capriati	
Vapour	Hingis	
Droplets	Sharapova	
Moisture	Seles	
Oxidation	Henin	
Condense	Clijsters	

- Examples: Arithmetic
  - **zoology – animal + flower ≈ botany**
  - **clustering – unsupervised + supervised ≈ k-means**

ConceptNet Numberbatch (Speer et al., 2017)

- Uses Word2Vec and Glove vectors as starting point
- Applies retrofitting with the knowledge graph *ConceptNet*
  - Transform original embedding vectors in a way that edge vectors in the knowledge graph are close to each other in embeddings space

$$\Psi(Q) = \sum_{i=1}^{n} \left[\alpha_i \|q_i - \hat{q}_i\|^2 + \sum_{(i,j)\in E} \beta_{ij} \|q_i - q_j\|^2 \right]$$

- 1,900,000 words in vocabulary

Observations

Less convincing in simple similarities and analogies
Many OOV words
Exposes insightful conceptual connections

- Examples: Nearest words

Condensation	Venus	Aradish
Steaming up	Evening star	[OOV]
Recondensation	Cytherean	
Condensing	Venerian	
Steam up	Dii majores	
Steam condenser	Venus‘s flower basket	
Fog up	Roman god	
Condensate	Inner planet	

- Examples: Arithmetic
  - **zoology – animal + flower ≈ albiflourus**
  - **clustering – unsupervised + supervised ≈ clusterization**

Test of embeddings for subject identification

Can embedding reveal the subject structure in a collection of DLR-typical texts?

- Data set: Elib entries with English abstracts from past 365 days – 1749 documents
- For each document D: Calculate a posIdfRank* word ranking

$$\{((w_i, g_i) \in \mathcal{W} \times \mathbb{R} \mid i = 1, \ldots, k\}$$

- Calculate weighted average document vector

$$d = \sum_{i:w_i \in D \cap \mathrm{Vocab}} g_i \mathrm{emb}(w_i) \,/ \sum_{i:w_i \in D \cap \mathrm{Vocab}} g_i$$

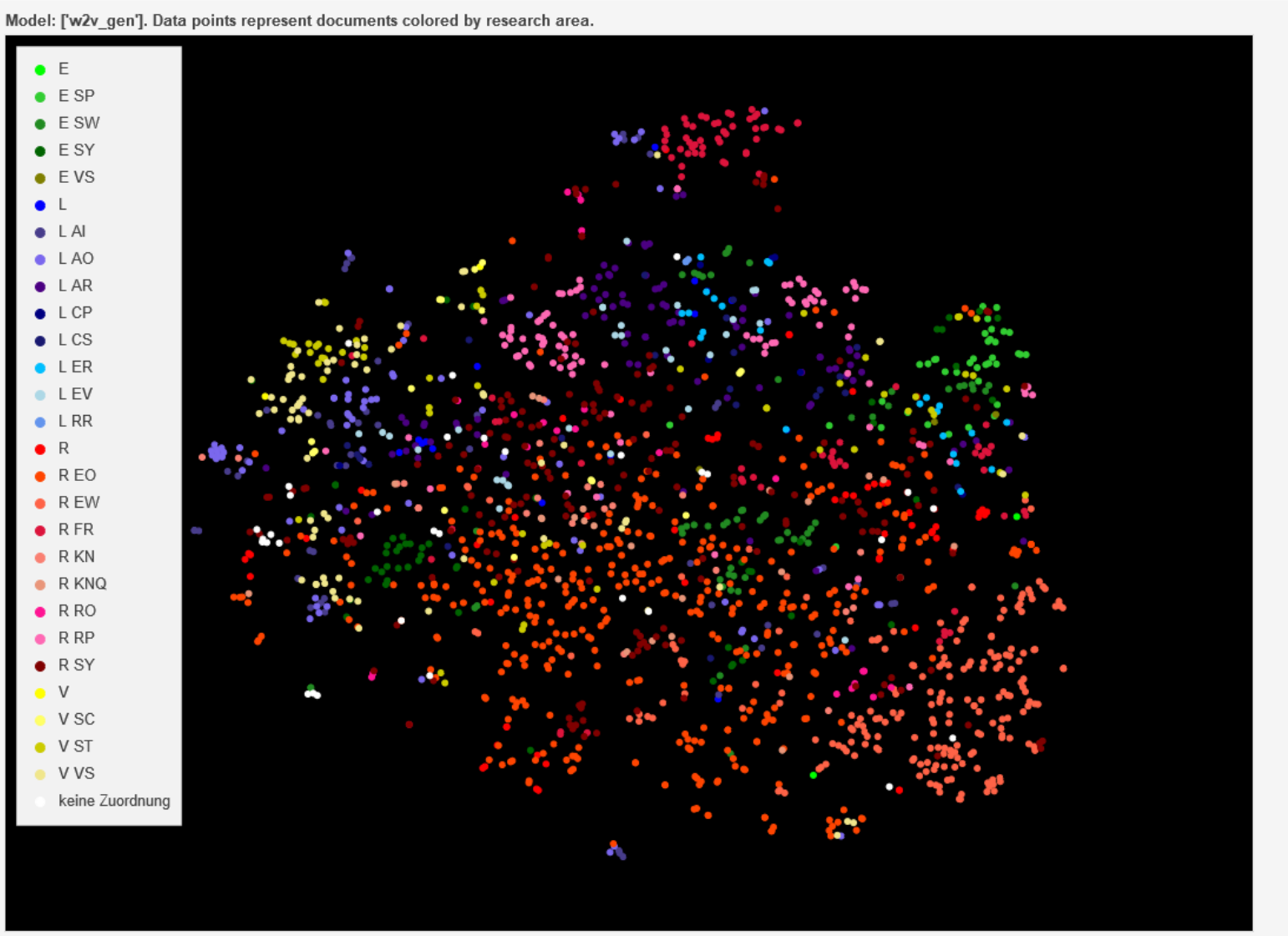
- Try to discover the subject area classification of Elib in the position of the document vectors

E	Energie
E SP	Energiespeicher
E SW	Solar- und Windenergie
E SY	Energiesystemtechnologie und -analyse
E VS	Verbrennungssysteme
L	Luftfahrt
L AI	Luftverkehr und Auswirkungen
L AO	Air Traffic Management and Operation
L AR	Aircraft Research
L CP	Umweltschonender Antrieb
L CS	Komponenten und Systeme
L ER	Engine Research
L EV	Effizientes Luftfahrzeug
L RR	Rotorcraft Research
R	Raumfahrt
R EO	Erdbeobachtung
R EW	Erforschung des Weltraums
R FR	Forschung unter Weltraumbedingungen
R KN	Kommunikation und Navigation
R KNQ	Kommunikation, Navigation und Quantentechnologie
R RO	Robotik
R RP	Raumtransport
R SY	Technik für Raumfahrtsysteme
V	Verkehr
V SC	Schienenverkehr
V ST	Straßenverkehr
V VS	Verkehrssysteme
keine Zuordnung	

* Hamm, A. and Odrowski, S. *Term-Community-Based Topic Detection with Variable Resolution.* Information, 12 (6). MDPI 2021. doi: 10.3390/info12060221.

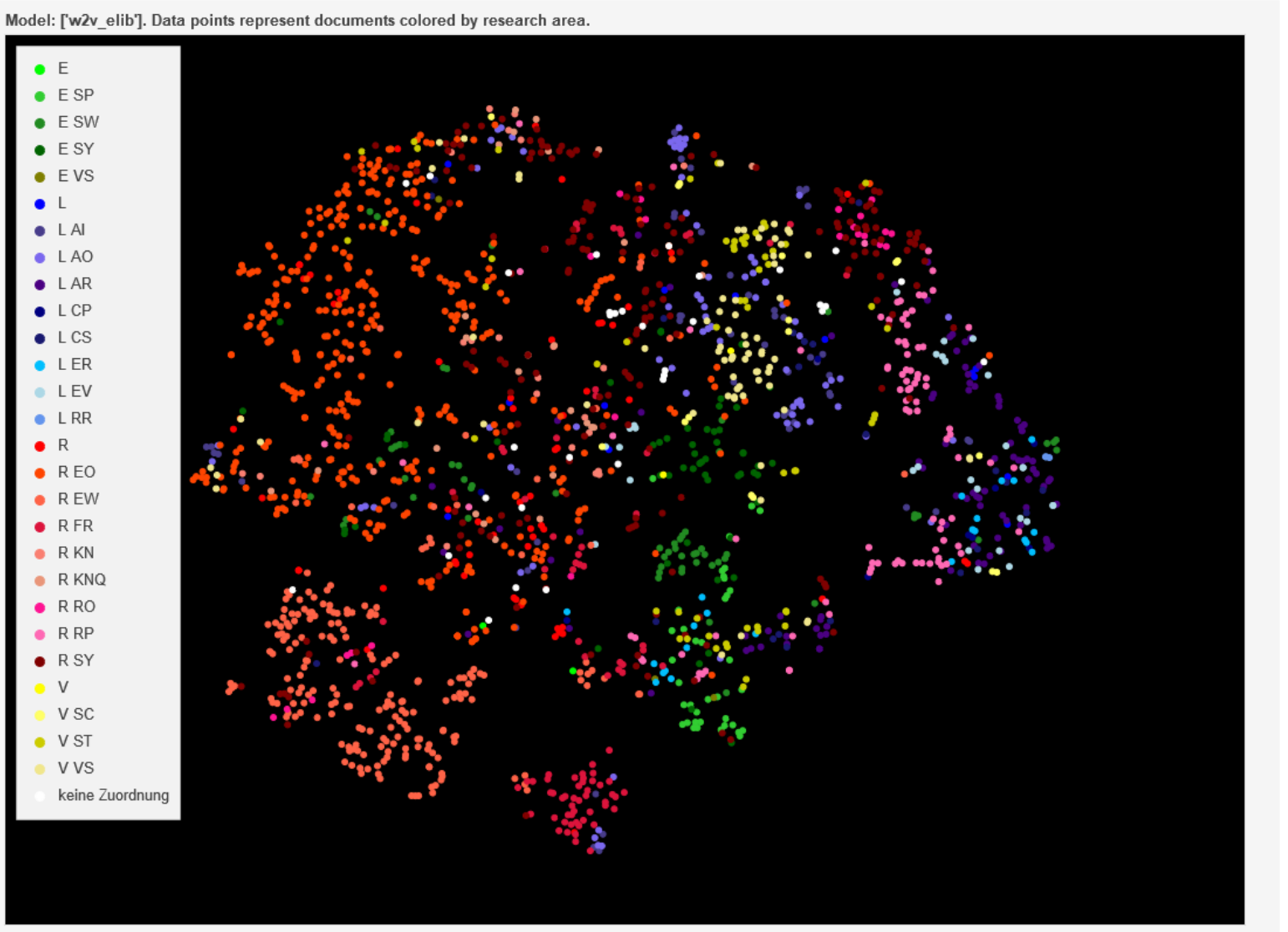
Distribution of research areas in embedding space Pretrained Word2Vec

- Projection of Elib document vectors into 2-dim plane using tSNE
- Colour according to one of 28 subject areas
- Colours show clustering tendency
- Colour clusters mostly not clearly separated



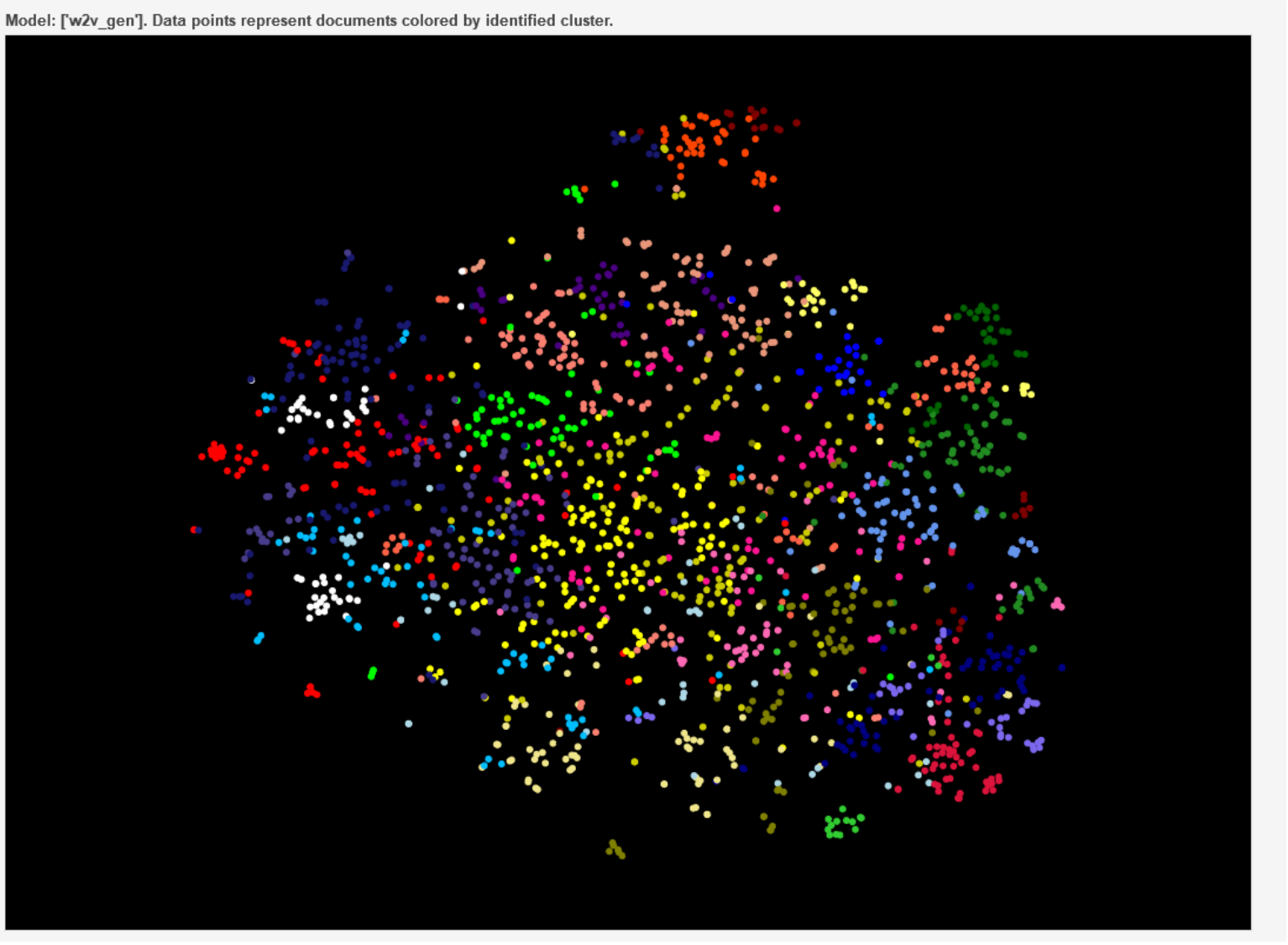
Distribution of research areas in embedding space
Elib-trained Word2Vec

- Projection of Elib document vectors into 2-dim plane using tSNE
- Colour according to one of 28 subject areas
- Colours show clustering tendency
- Colour clusters are more clearly separated



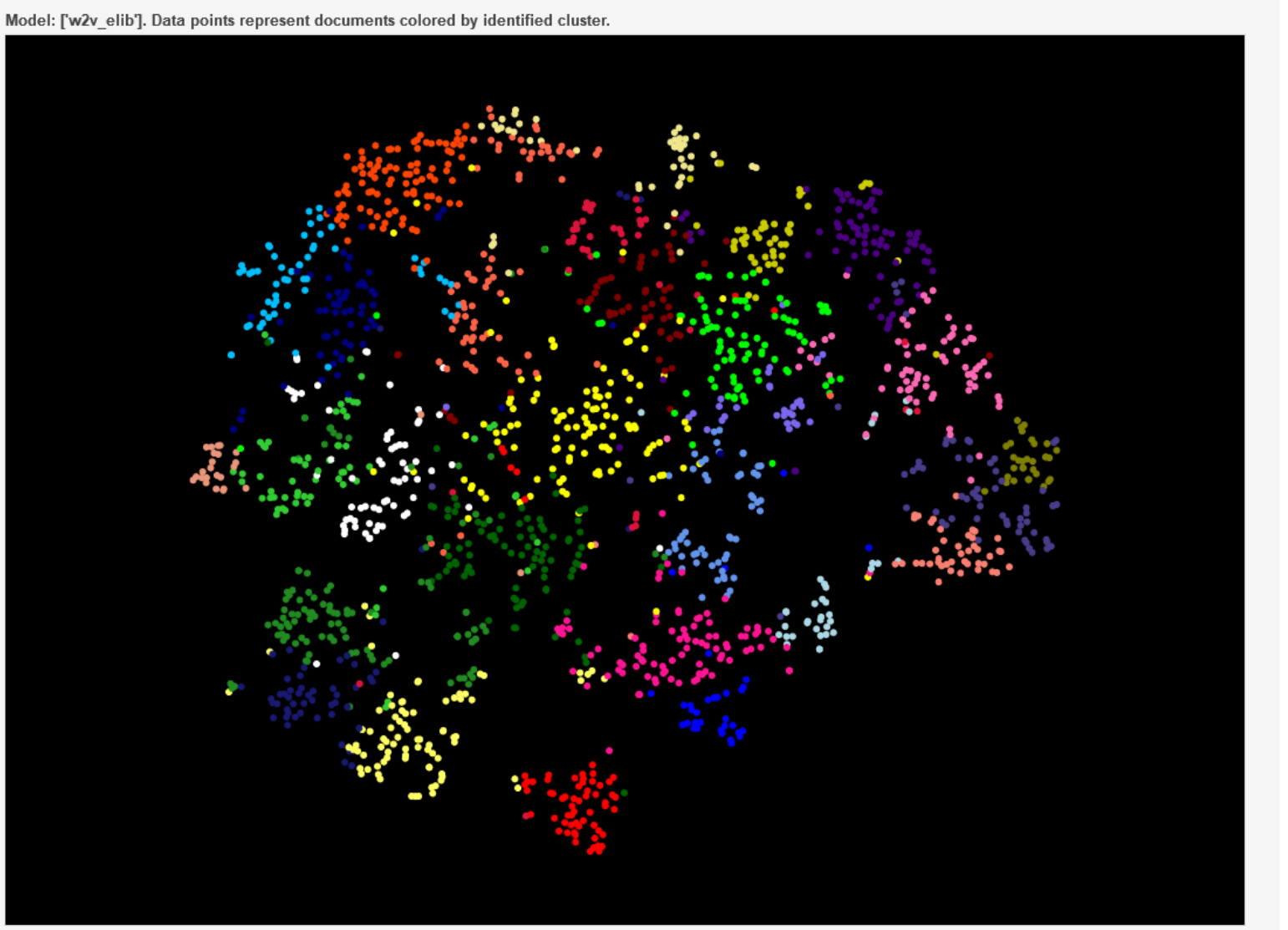
Clustering in embedding space
Pretrained Word2Vec

- Projection of Elib document vectors into 2-dim plane using tSNE
- Colour according to results of k-means clustering (random colours for 28 clusters)
- Colour clusters mostly not clearly separated



Clustering in embedding space
Elib-trained Word2Vec

- Projection of Elib document vectors into 2-dim plane using tSNE
- Colour according to results of k-means clustering (random colours for 28 clusters)
- Colour clusters mostly not clearly separated



Measuring the quality of clustering in embedding space

- Intrinsic quality measure
  - Silhouette of a data point: $s = {}^{b-a}/_{\max(a,b)}$
    - a: mean distance between that point and all other points of its cluster
    - b: mean distance between the point and all points of the nearest cluster
  - **Silhouette coefficient** is the mean of all silhouettes
- Extrinsic quality measure by comparison with subject area classification
  - **Homogeneity** – Ideal: All clusters contain only documents of one class

    $h = 1 - {}^{H(C|K)}/_{H(C)}$, where $H(C|K)$ is the conditional entropy of the class distribution given the clustering
  - **Completeness** – Ideal: All documents of one class belong to the same cluster

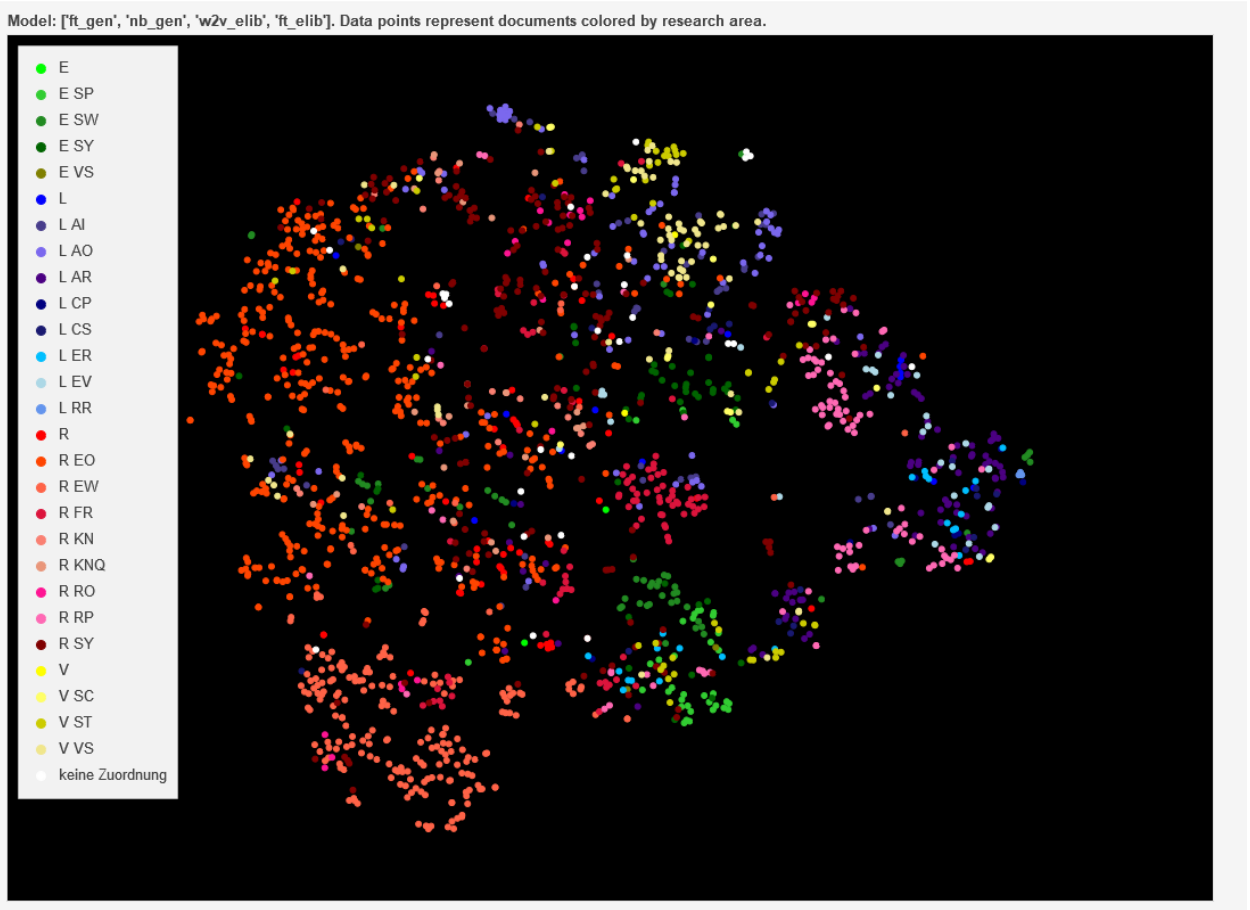
    $h = 1 - {}^{H(K|C)}/_{H(K)}$, where $H(K|C)$ is the conditional entropy of the cluster distribution given the class
  - **V-Measure**: Harmonic mean of homogeneity and completeness.

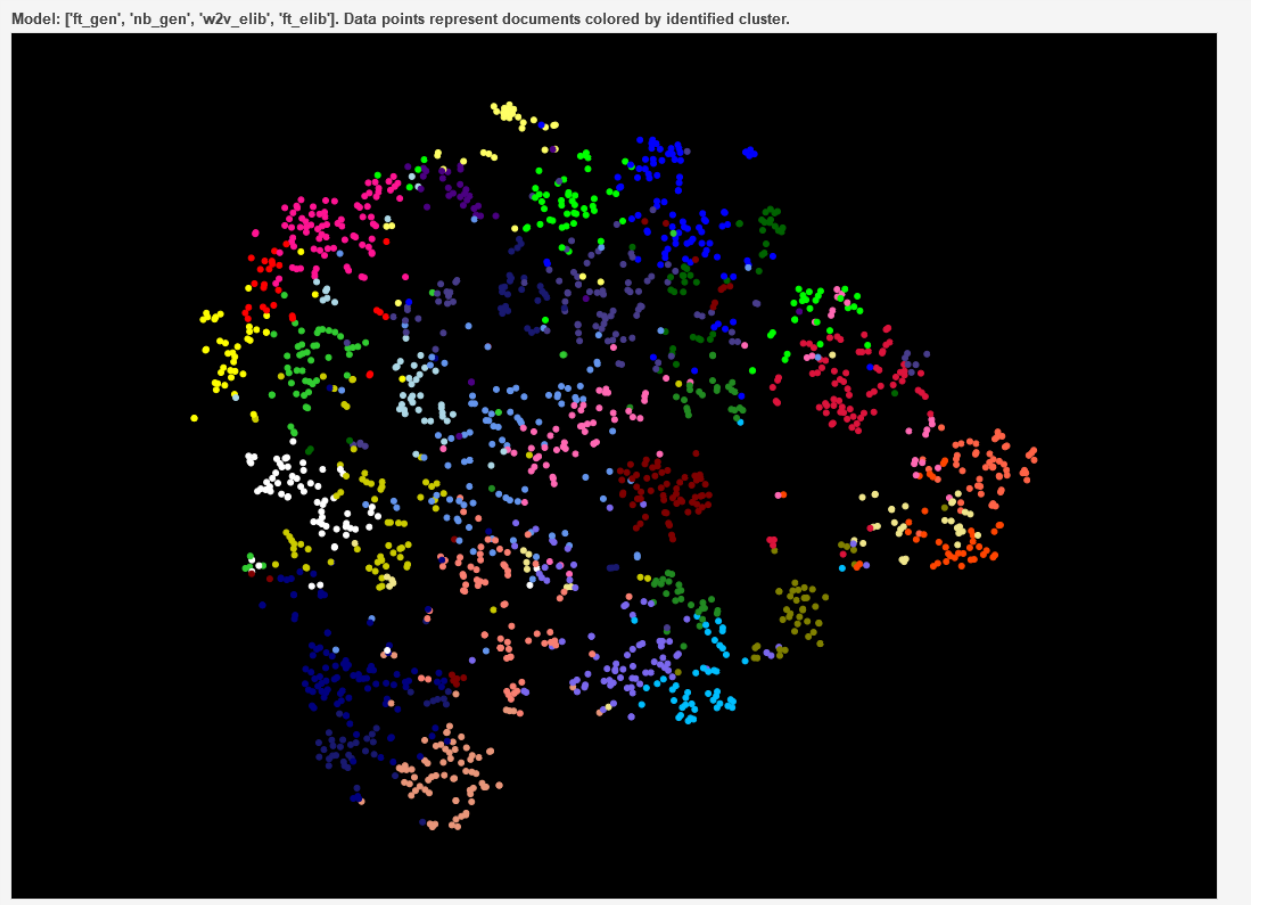
Comparing the quality of clustering in embedding space

- Domain-specific training leads to better clustering results
- Quality can be further improved by combining (concatenating) embeddings
  - Best combination: W2v_elib + fT_elib + Numberbatch + fT_gen

Score	Silhouette	Homogeneity	Completeness	V-Measure
W2v_gen	0.035	0.344	0.300	0.321
W2v_elib	**0.092**	**0.419**	**0.351**	**0.382**
fT_gen	0.032	0.367	0.311	0.336
fT_elib	0.082	0.402	0.339	0.368
Glove	0.038	0.339	0.289	0.312
Numberbatch	0.045	0.384	0.326	0.353
Combination	**0.094**	**0.425**	**0.350**	**0.384**

Pretrained fastText + NumberBatch + elib-trained fastText + elib-trained Word2Vec





Conclusions and outlook

- By combining several word embeddings we can benefit from the best aspects of the individual embeddings
  - Domain specific training increases the information on relevant details
  - Pretrained vectors provide a huge vocabulary
  - Subword-based models take into account grammatical inflections and typos
  - Knowledge-graph-enhanced methods link words to concepts
- There is room for further experiments and optimization with regard to the exact details of combination
- Here we looked only at the weighted average of word vectors in a document. More detailed insights are to be expected from a closer look at the distribution of word vectors in a document.

Combining word embeddings as a tool for subject identification

Andreas Hamm
SC-IVS
Project: MeToDiO

Thank you for your attention