Cross-Topic Distributional Semantic Representations via Unsupervised Mappings

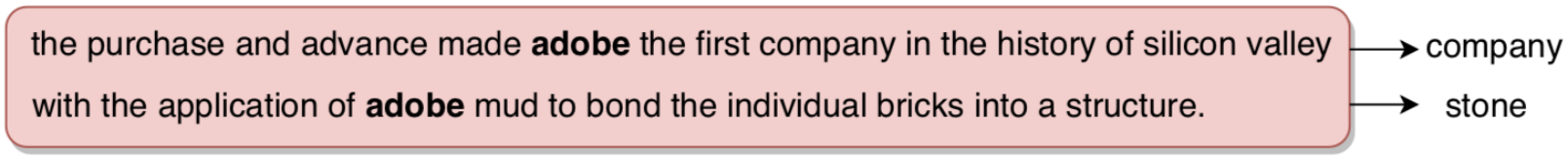
Eleftheria Briakou[1,2] Nikos Athanasiou[2]
Alexandros Potamianos[2,3]

[1]University of Maryland, College Park, MD
[2]**ECE, National Technical University of Athens, Athens, Greece**
[3]Signal Analysis and Interpretation Laboratory, USC, Los Angeles, USA

Why multiple embeddings for a single word?

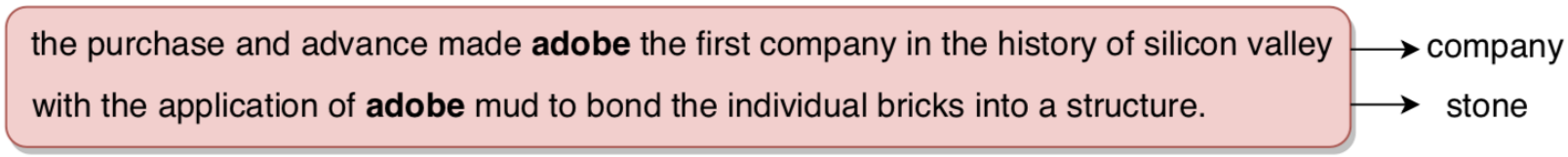
Why multiple embeddings for a single word?

Words change their meaning based on the *context* they reside in.



Why multiple embeddings for a single word?

Words change their meaning based on the *context* they reside in.

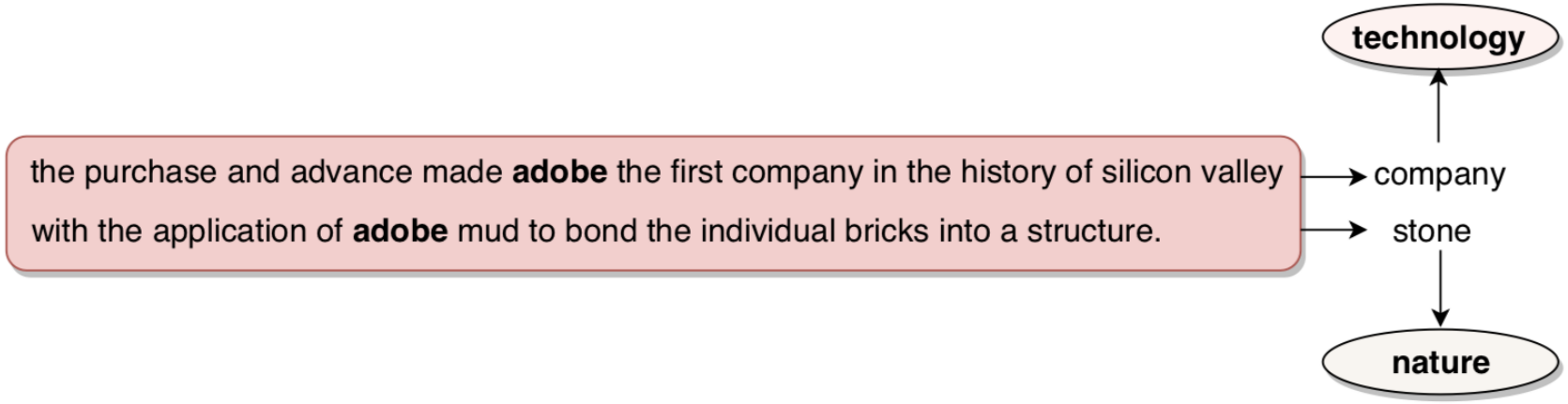


We need Distributional Semantic Models (DSMs) that:

- go beyond the representation of words by one point in the semantic space
- are able to capture the distinct meanings of polysemous words

Why topic-based embeddings?

Different *contexts* can be found under different *topic domains*.
Topic-embeddings can capture variations in word semantics

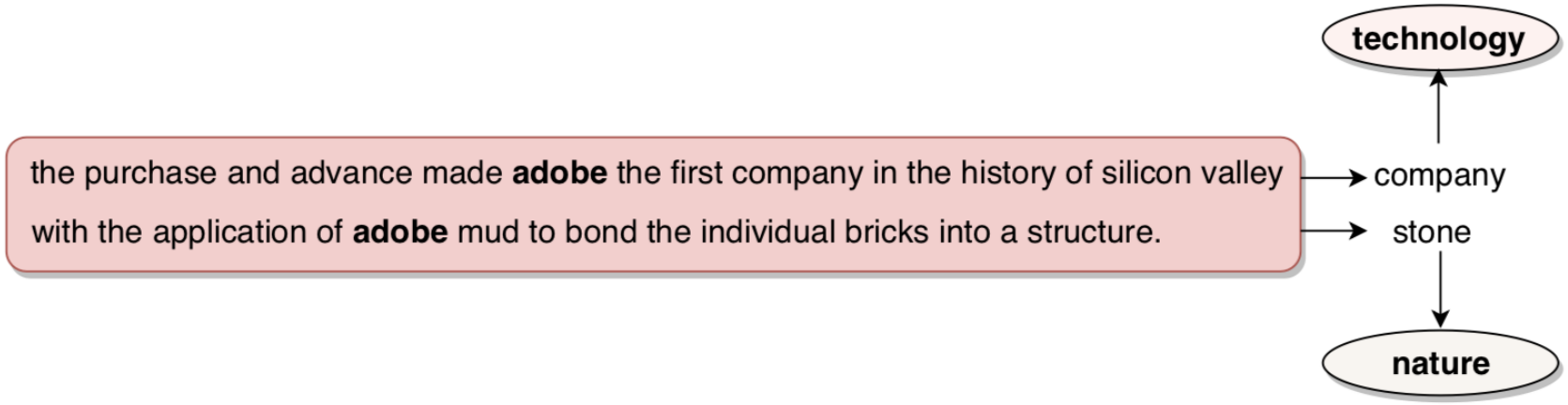


Hypothesis

Polysemous words may change their semantics under different topics
Monosemous words share the same semantics regardless of their topic

Why topic-based embeddings?

Different *contexts* can be found under different *topic domains*.
Topic-embeddings can capture variations in word semantics

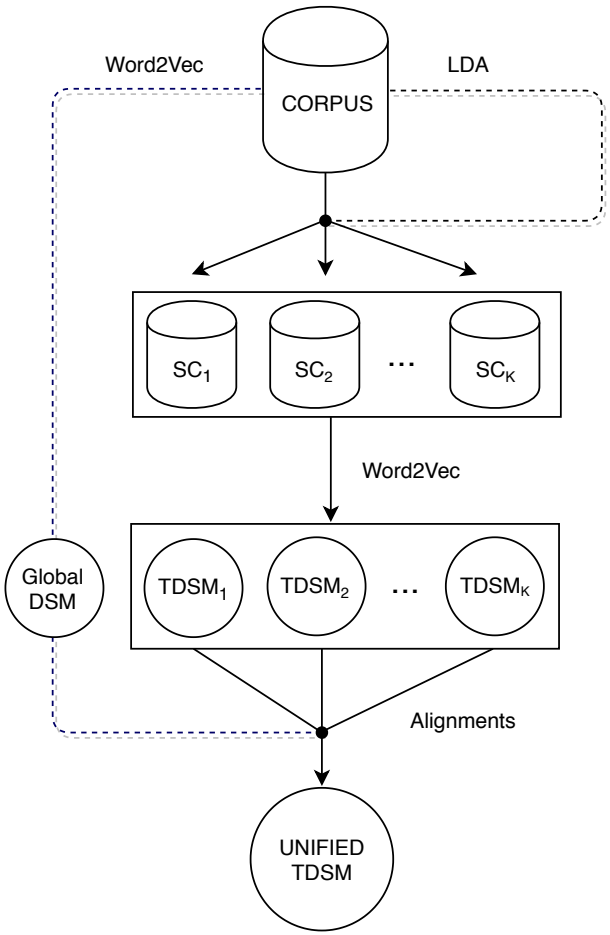


Hypothesis

Polysemous words may change their semantics under different topics
Monosemous words share the same semantics regardless of their topic

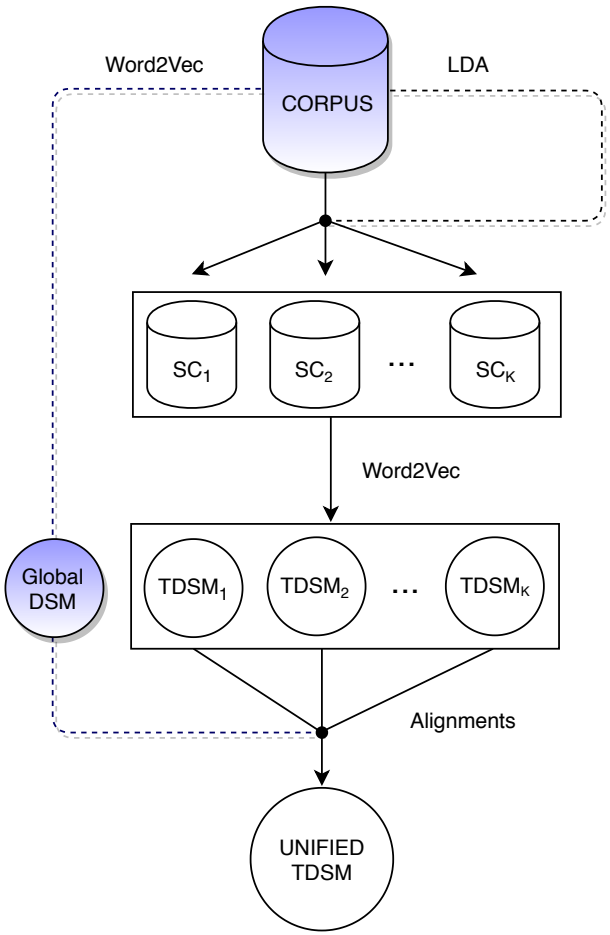
How could we create a unified space of multiple topic representations per word?

Model Overview

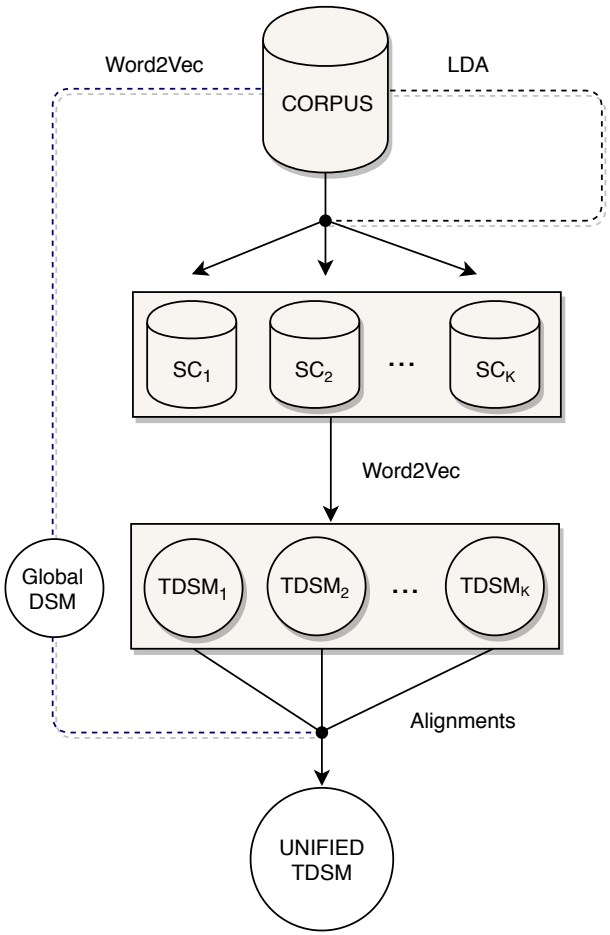


Model Overview

1. *Generic* corpus $\Rightarrow$ *Global* DSM



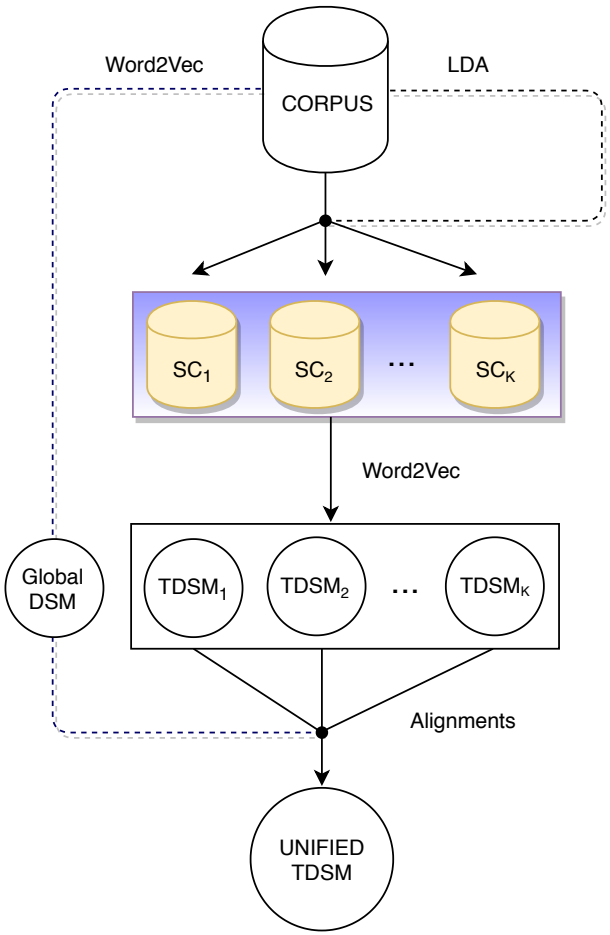
Model Overview



1. *Generic* corpus $\Rightarrow$ *Global* DSM

2. Topic-Based DSMs (TDSMs) (Christopoulou et al., 2018)

Model Overview

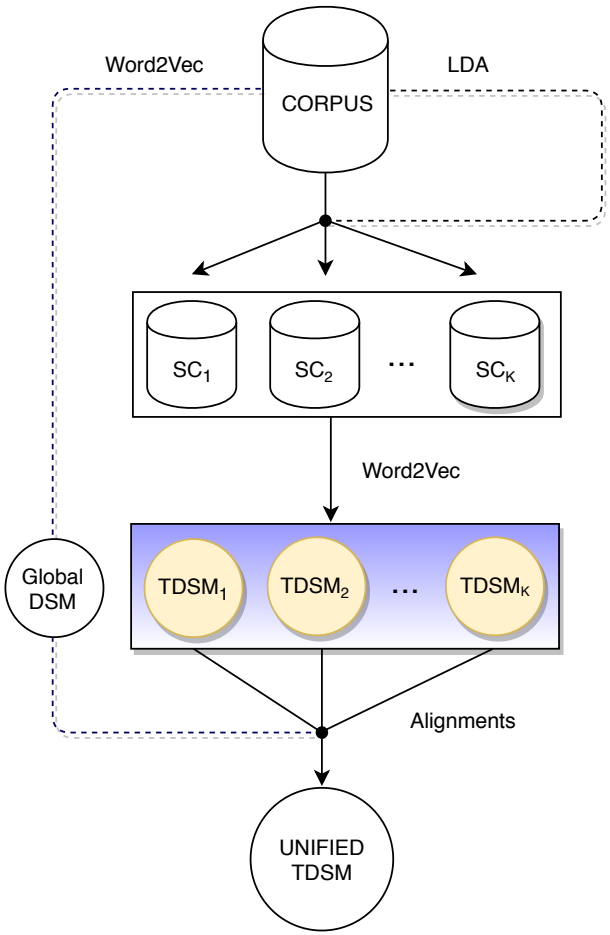


1. *Generic* corpus ⇒ *Global* DSM

2. Topic-Based DSMs (TDSMs) (Christopoulou et al., 2018)

   - Train LDA under generic corpus. *Cluster* sentences ⇒ topic sub-corpora

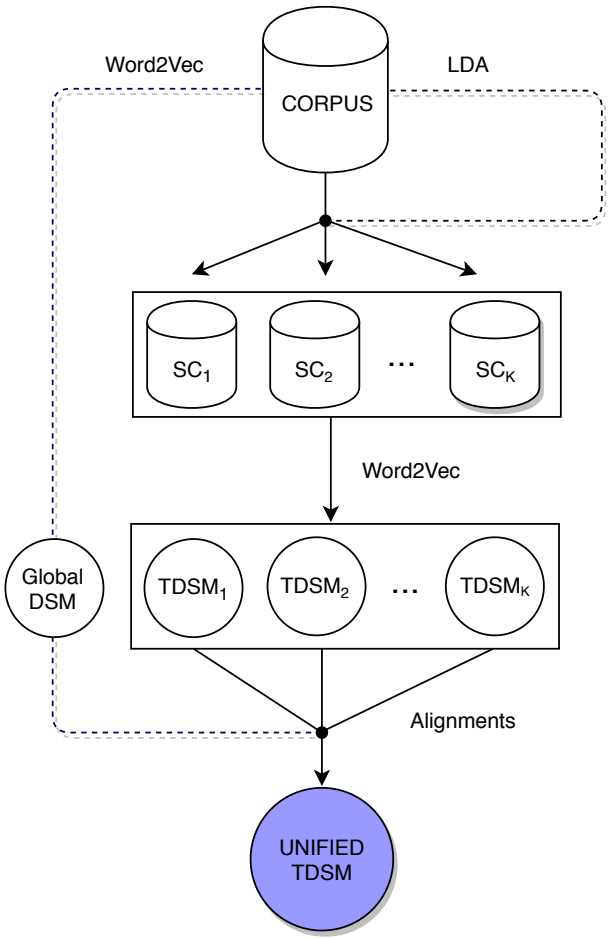
Model Overview



1. *Generic* corpus $\Rightarrow$ *Global* DSM

2. Topic-Based DSMs (TDSMs) (Christopoulou et al., 2018)
   - Train LDA under generic corpus. *Cluster* sentences $\Rightarrow$ topic sub-corpora
   - Train K TDSMs

Model Overview



1. *Generic* corpus ⇒ *Global* DSM

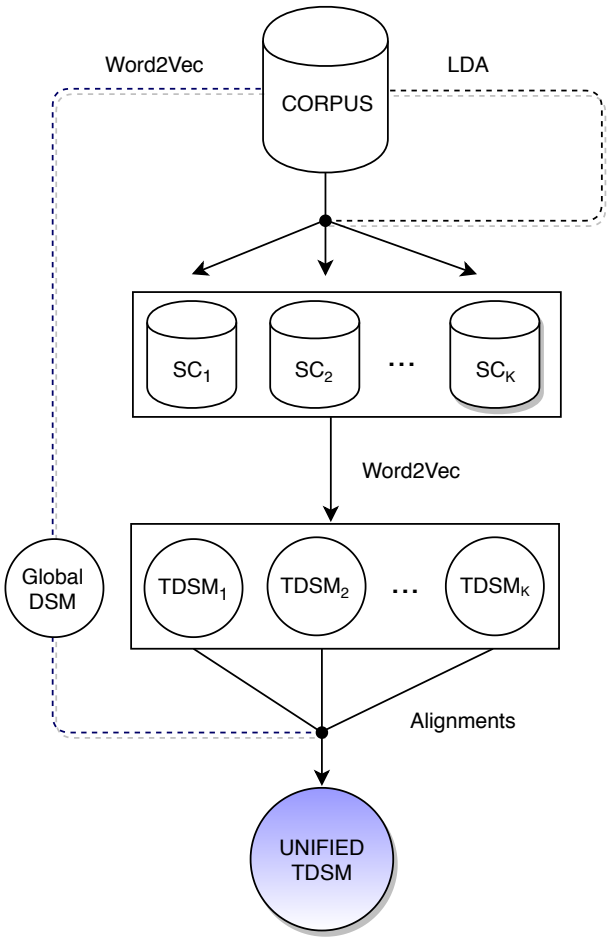
2. Topic-Based DSMs (TDSMs) (Christopoulou et al., 2018)
   - Train LDA under generic corpus. *Cluster* sentences ⇒ topic sub-corpora
   - Train K TDSMs

3. Project each TDSM to Global-DSM

   Anchor Selection

   Unified Topic-based DSM (UTDSM)

Model Overview



1. *Generic* corpus ⇒ *Global* DSM

2. Topic-Based DSMs (TDSMs) (Christopoulou et al., 2018)
   - Train LDA under generic corpus. *Cluster* sentences ⇒ topic sub-corpora
   - Train K TDSMs

3. Project each TDSM to Global-DSM
   Anchor Selection
   Unified Topic-based DSM (UTDSM)

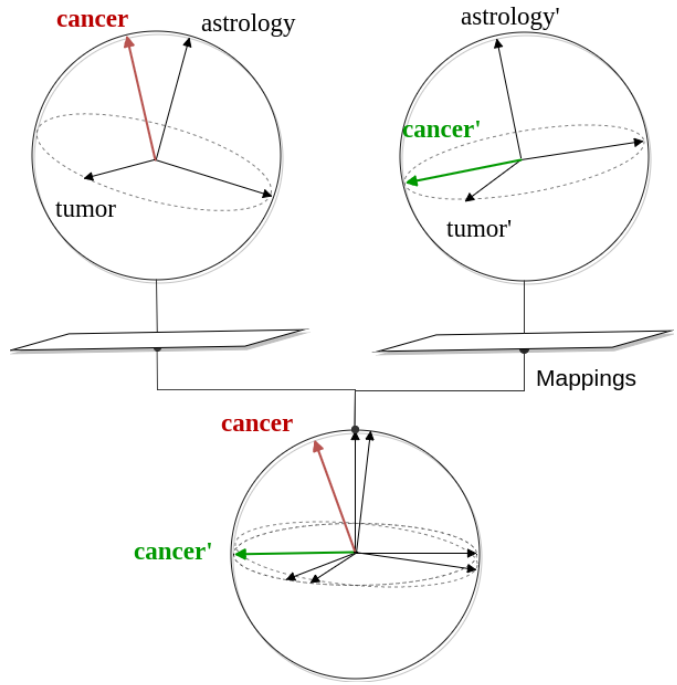
4. Smoothing approach

Anchor Selection

- *Anchors* $\Rightarrow$ define mappings
  $\Rightarrow$ stable relationships

Anchor Selection

- *Anchors* ⇒ define mappings
  ⇒ stable relationships
- Monosemous words



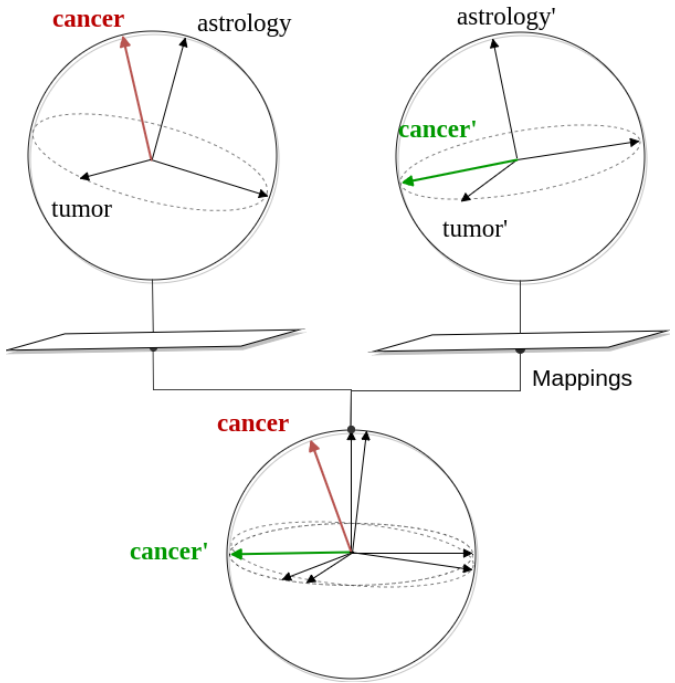
Hypothesis

Semantic relationships between monosemous words remain stable across *topic domains*

Relative distances between monosemous words are preserved across *TDSMs*!

Anchor Selection

- *Anchors* $\Rightarrow$ define mappings
  $\Rightarrow$ stable relationships
- Monosemous words
- How to retrieve monosemous words without supervision?
- Semantic similarity matrices!

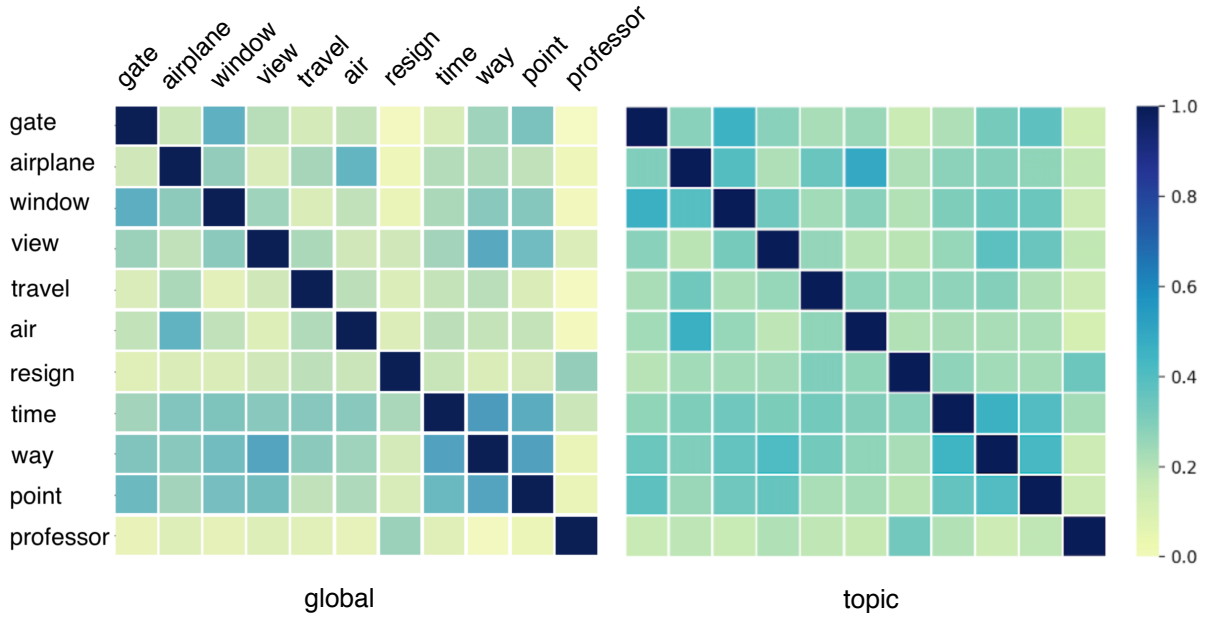


Hypothesis

Semantic relationships between monosemous words remain stable across *topic domains*

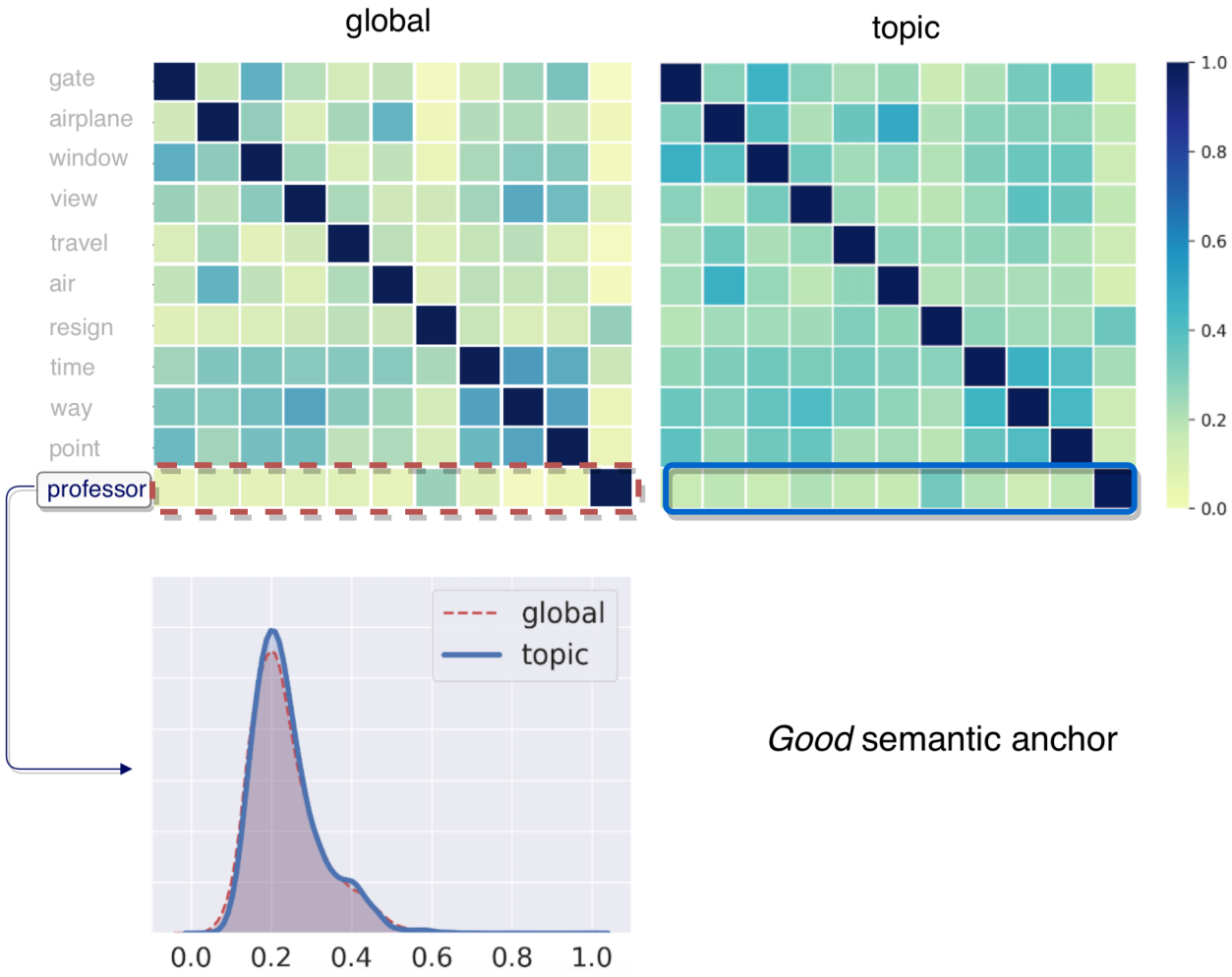
Relative distances between monosemous words are preserved across *TDSMs*!

Semantic Similarity Matrices



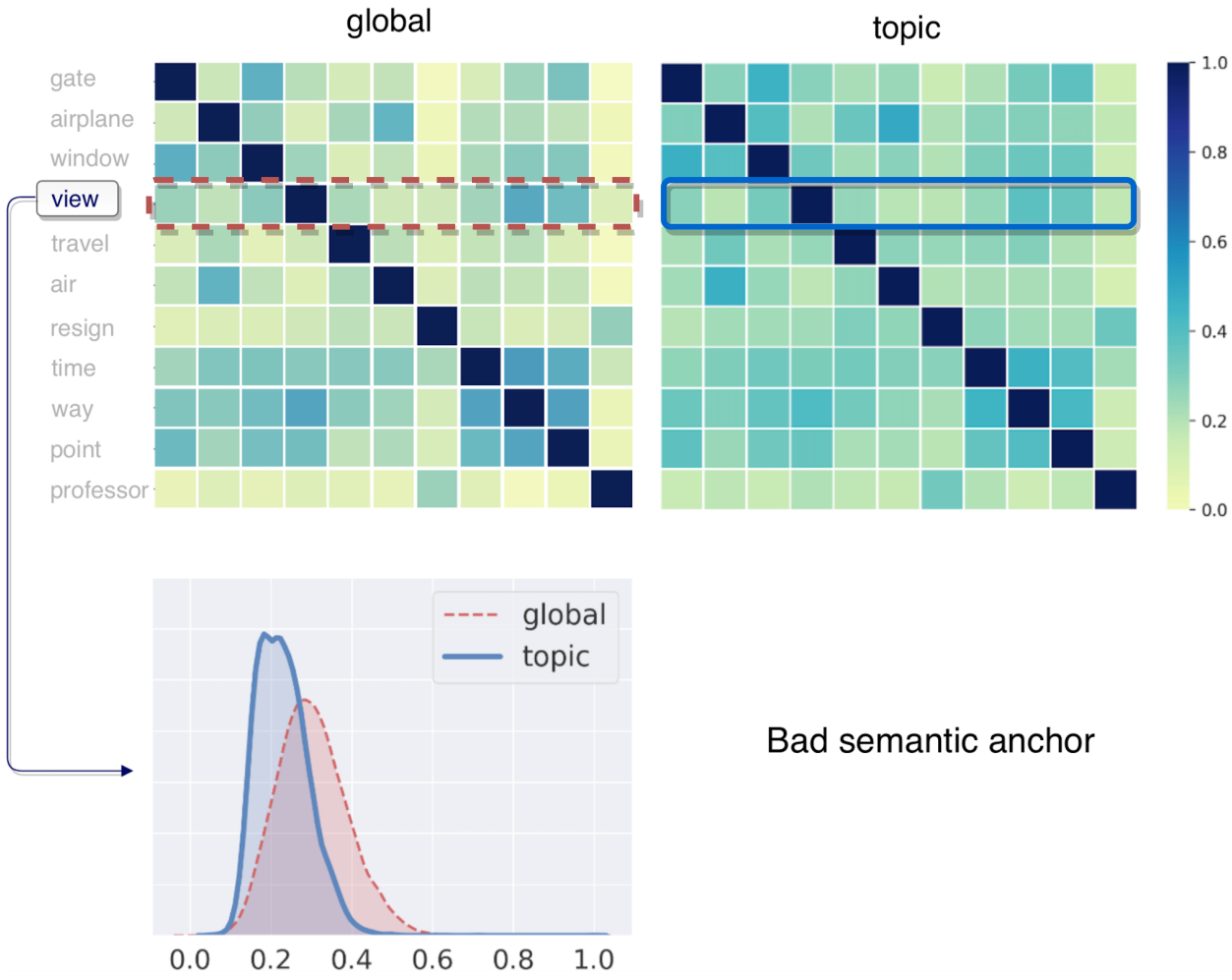
- Similarity matrices are aligned (Artetxe et al., 2018)
- Compare similarity distributions between *global* and *topic* spaces
- How? $\Rightarrow$ Euclidean distance

Semantic Similarity Matrices



Semantic anchors should have *consistent similarity distributions* regardless of the domain they exist in

Semantic Similarity Matrices



Semantic anchors should have *consistent similarity distributions* regardless of the domain they exist in

Semantic Mappings

The transformation matrix $M_k \in \mathbb{R}^{d \times d}$ that projects in the k-th topic space to the global space is learned via solving :[1]

$$\min_{M_k} \sum_{j=1}^{|A|} \|M_k \alpha_k^j - \alpha_g^j\|_2^2, \quad \text{s.t.} \quad M_k M_k^T = \mathbb{I}$$

where,
A is the list of semantic anchors
$\alpha_k^j \in \mathbb{R}^d$ is the vector of the j-th *anchor* word in the k-th topic space
$\alpha_g^j \in \mathbb{R}^d$ is the corresponding vector in the global space

[1] Orthogonal Procrustes problem Schönemann (1966)

Semantic Mappings

The transformation matrix $M_k \in \mathbb{R}^{d \times d}$ that projects in the k-th topic space to the global space is learned via solving :[1]

$$\min_{M_k} \sum_{j=1}^{|A|} \|M_k \alpha_k^j - \alpha_g^j\|_2^2, \quad \text{s.t.} \quad M_k M_k^T = \mathbb{I}$$

where,
A is the list of semantic anchors
$\alpha_k^j \in \mathbb{R}^d$ is the vector of the j-th *anchor* word in the **k-th topic space**
$\alpha_g^j \in \mathbb{R}^d$ is the corresponding vector in the global space

[1] Orthogonal Procrustes problem Schönemann (1966)

Semantic Mappings

The transformation matrix $M_k \in \mathbb{R}^{d \times d}$ that projects in the k-th topic space to the global space is learned via solving :[1]

$$\min_{M_k} \sum_{j=1}^{|A|} \|M_k \alpha_k^j - \alpha_g^j\|_2^2, \quad \text{s.t.} \quad M_k M_k^T = \mathbb{I}$$

where,
A is the list of semantic anchors
$\alpha_k^j \in \mathbb{R}^d$ is the vector of the j-th *anchor* word in the k-th topic space
$\alpha_g^j \in \mathbb{R}^d$ is the corresponding vector in the **global space**

[1] Orthogonal Procrustes problem Schönemann (1966)

Semantic Mappings

The transformation matrix $M_k \in \mathbb{R}^{d \times d}$ that projects in the k-th topic space to the global space is learned via solving :[1]

$$\min_{M_k} \sum_{j=1}^{|A|} \|M_k \alpha_k^j - \alpha_g^j\|_2^2, \quad \text{s.t.} \quad M_k M_k^T = \mathbb{I}$$

where,
A is the list of semantic anchors
$\alpha_k^j \in \mathbb{R}^d$ is the vector of the j-th *anchor* word in the k-th topic space
$\alpha_g^j \in \mathbb{R}^d$ is the corresponding vector in the global space

Given a word and its k-th topic distributed representation $x_k \in \mathbb{R}^d$, we compute its projected representation $x'_k \in \mathbb{R}^d$ as follows:

$$x'_k = M_k x_k$$

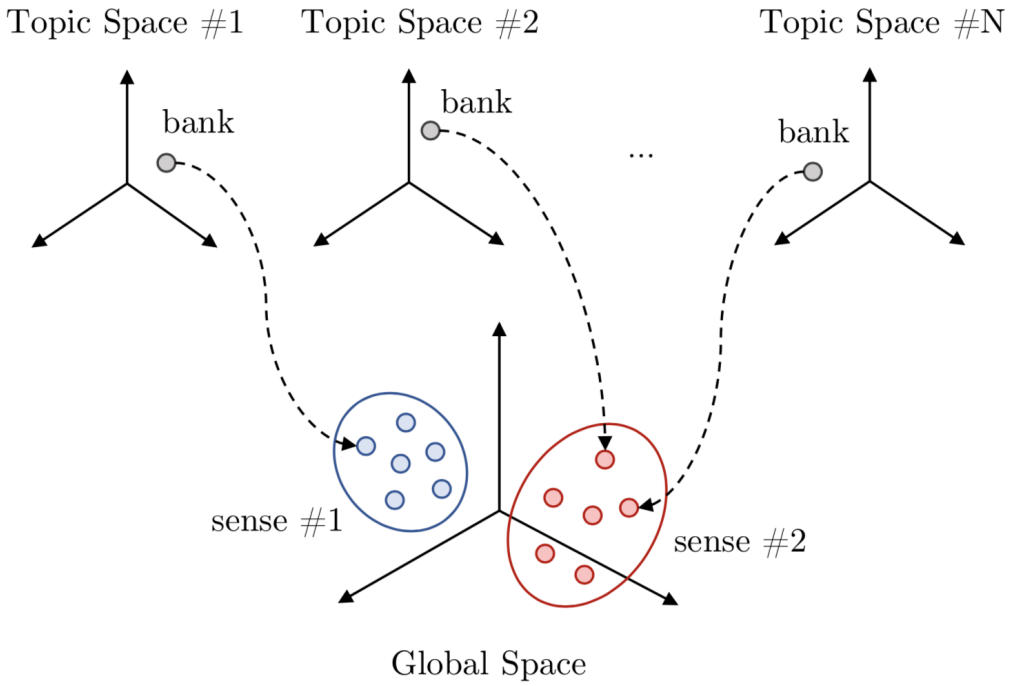
[1] Orthogonal Procrustes problem Schönemann (1966)

Smoothing

- Lessen the estimation error introduced to unified space through:
  - semantic mappings
  - sparse training data

Smoothing

- Lessen the estimation error introduced to unified space through:
  - semantic mappings
  - sparse training data
- Closely positioned vectors may correspond to the same meaning
- Smoothed representations capture finer-grained word semantics



Smoothing Approach

- Each word's topic embeddings are clustered into N Gaussian distributions via a **Gaussian Mixture Model** (GMM)
- Closely positioned topic embeddings are assigned to the **same** component
- Gaussian distribution forms a **semantically coherent unit** that corresponds to closely related semantics of the target word
- The **mean vector** of each Gaussian distribution is used as a representative vector of each component

Contextual Semantic Similarity (1/2)

- **Dataset:** Stanford's Contextual Word Similarities (SCWS)
- **Task:** Predict semantic similarity between a pair of words provided in *sentential contexts*
- **Metrics:**
  - **AvgSimC**: weighs the contribution of each topic-based word embeddings according to probability of the word belonging to that topic
  - **MaxSimC**: uses only the topic-based word embedding that corresponds to the most probable topic assignment

Contextual Semantic Similarity (2/2)

Method	AvgSimC	MaxSimC
Liu et al. (2015a)	67.3	68.1
Liu et al. (2015b)	69.5	67.9
Amiri et al. (2016)	**70.9**	-
Global-DSM	67.6	67.6
Unified-DSM	70.2	68.0
Unified-DSM + GMM	69.0	**68.5**

Contextual Semantic Similarity (2/2)

Method	AvgSimC	MaxSimC
Liu et al. (2015a)	67.3	68.1
Liu et al. (2015b)	69.5	67.9
Amiri et al. (2016)	**70.9**	-
Global-DSM	67.6	67.6
Unified-DSM	70.2	68.0
Unified-DSM + GMM	69.0	**68.5**

- Multi-topic embeddings perform better that single representations

Contextual Semantic Similarity (2/2)

Method	AvgSimC	MaxSimC
Liu et al. (2015a)	67.3	68.1
Liu et al. (2015b)	69.5	67.9
Amiri et al. (2016)	**70.9**	-
Global-DSM	67.6	67.6
Unified-DSM	70.2	68.0
Unified-DSM + GMM	69.0	**68.5**

- Multi-topic embeddings perform better that single representations
- Smoothing improves over MaxSimC—a metric sensitive to noisy word representations

Contextual Semantic Similarity (2/2)

Method	AvgSimC	MaxSimC
Liu et al. (2015a)	67.3	68.1
Liu et al. (2015b)	69.5	67.9
Amiri et al. (2016)	**70.9**	-
Global-DSM	67.6	67.6
Unified-DSM	70.2	68.0
Unified-DSM + GMM	69.0	**68.5**

- Multi-topic embeddings perform better that single representations
- Smoothing improves over MaxSimC—a metric sensitive to noisy word representations
- Results are competitive to the state-of-the-art models

Text Classification

- **Dataset:** 20NewsGroup dataset
- **Task:** Classify each document into one of the 20 different newsgroups based on its content
- **Document-level embeddings** used as features
- SVM classifier

Method	F1-score	Accuracy
Global-DSM	62.9	63.3
Unified-DSM	**64.5**	**65.5**

Document level representations extracted from multiple topic-based embeddings outperform single-prototype models.

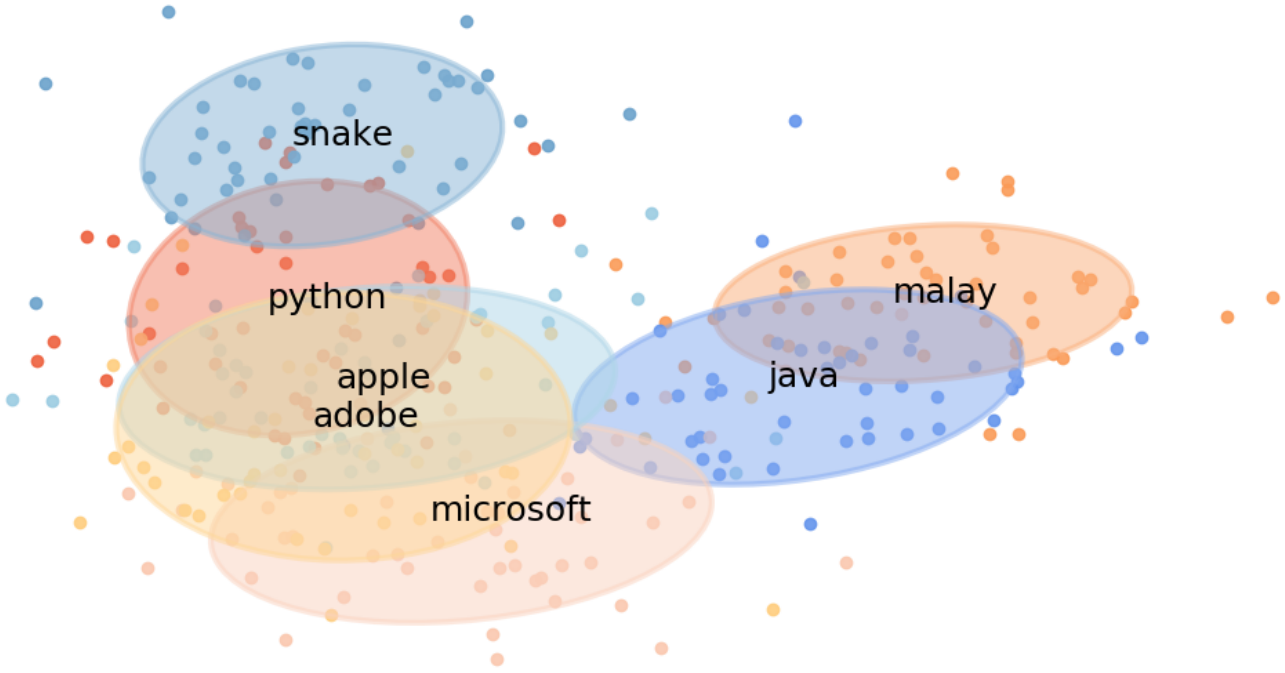
Paraphrase Identification

- **Dataset:** Microsoft Paraphrase dataset
- **Task:** Identifying whether two given sentences can be considered paraphrases or not
- **Sentence-level embeddings** used as features
- SVM classifier

Method	F1-score	Accuracy
Global-DSM	62.0	69.2
Unified-DSM	**64.0**	**69.4**

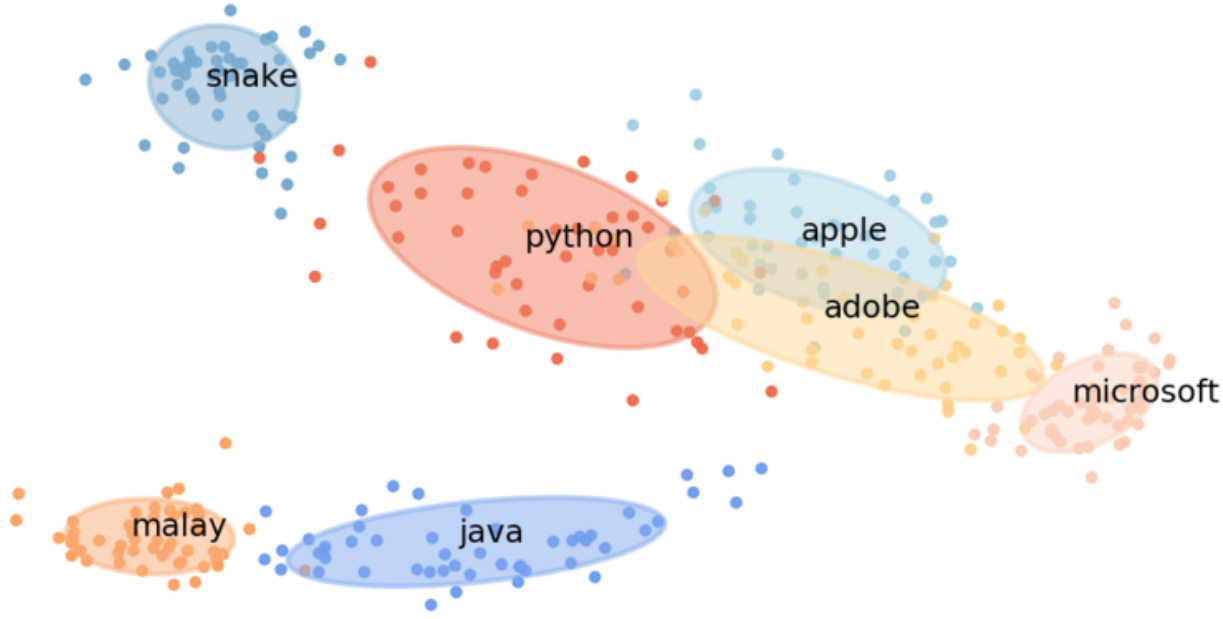
Sentence level representations extracted from multiple topic-based embeddings outperform single-prototype models.

Qualitative Results



- **Unaligned** topic sub-spaces
- Words demonstrate *similar area coverage* regardless of their polysemy

Qualitative Results



- **Aligned** topic sub-spaces
- Semantic relationships between words are better captured
- Area under a word's distribution is *indicative* of its degree of polysemy

Conclusion

- Unified space of multiple topic-based DSMs
- *unsupervised* approach for semantic anchor extraction
- projected word embeddings yield state-of-the-art results on contextual similarity
- outperform single vector representations in downstream NLP tasks
- Code at: https://github.com/Elbria/utdsm_naacl2018

References

Artetxe, M., Labaka, G., and Agirre, E. (2018). A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 789–798.

Christopoulou, F., Briakou, E., Iosif, E., and Potamianos, A. (2018). Mixture of topic-based distributional semantic and affective models. In *2018 IEEE 12th International Conference on Semantic Computing (ICSC)*, pages 203–210.

Schönemann, P. H. (1966). *A generalized solution of the orthogonal procrustes problem.*