Structured to Unstructured and Back: Integrated KG and NLP

Alexander Kalinowski

A proposed framework for fully unsupervised alignment between information in unstructured documents and elements of a knowledge graph for information extraction, fact validation and assisted ontology development.

Structured to Unstructured and Back

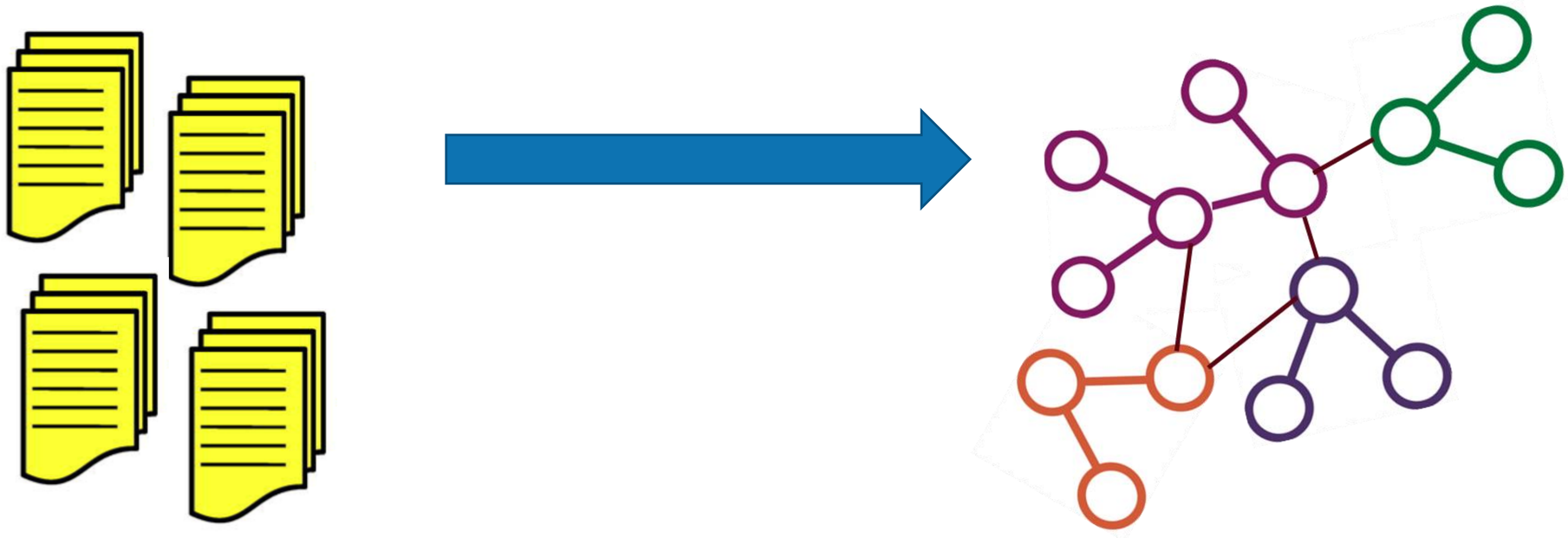
Motivation	Problems and use cases
Embedding Text	Transforming words to vectors
Embedding Graphs	Transforming graphs to vectors
Learn Alignments	Learning correspondences between graph triples and natural language

Motivating Questions and Problems

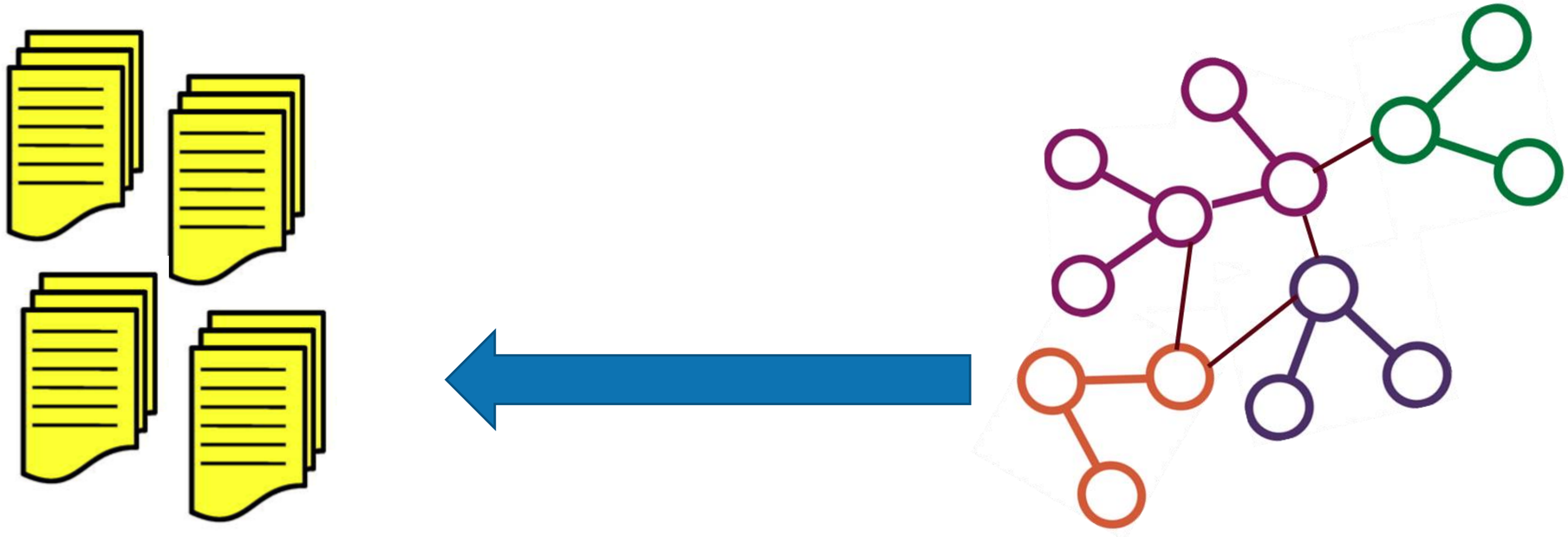
Strategy 1: Create next-generation capabilities by leveraging emerging Big Data foundations, techniques, and technologies. Continued, increasing investments in the next generation of large-scale data collection, management, and analysis will allow agencies to adapt to and manage the ever-increasing scales of data being generated, and leverage the data to create fundamentally new services and capabilities. Advances in computing and data analytics will provide new abstractions to deal with complex data, and simplify programming of scalable and parallel systems while achieving maximal performance. Fundamental advances in computer science, machine learning, and statistics will enable future data-analytics systems that are flexible, responsive, and predictive. Innovations in deep learning will be needed to create knowledge bases of interconnected information from unstructured data. Research into social computing such as crowdsourcing, citizen science, and collective distributed tasks will help develop techniques to enable humans to mediate tasks that may be beyond the scope of

Significant efforts are needed to curate datasets—to record the context as well as semantics associated with the data, and with the analyses performed on the data. Effective and proper reuse of data demands that the data context be properly registered and that data semantics be extracted and represented. While automation will be essential to accomplish this for Big Data, tasks related to curation, context, and semantics will also require a human-in-the-loop approach. Tools and ecosystems are needed to assist in this task, such as entity identification that utilizes global persistent identifiers and the use of domain ontologies for knowledge representation. Research in metadata modeling, automated metadata generation and registration, semantic technologies, ontologies, linked data, data provenance, and data citation will be important.

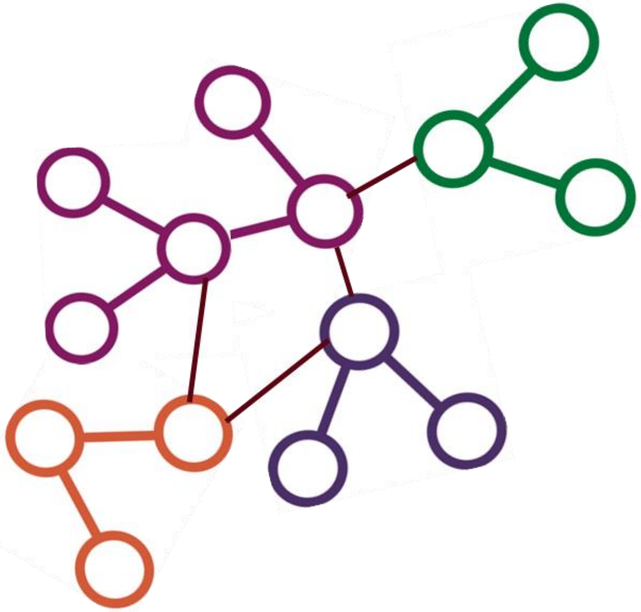
Information Extraction



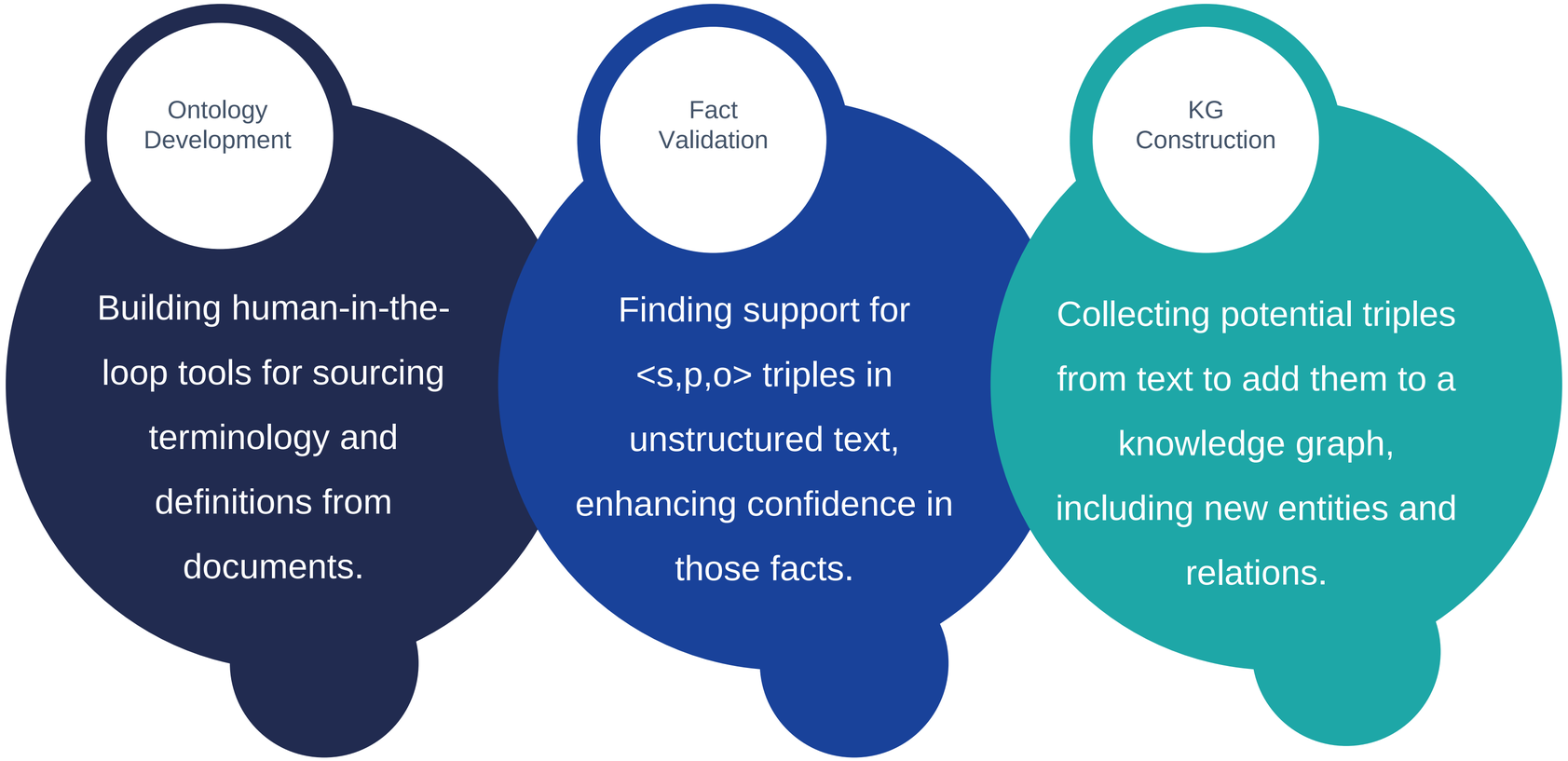
Automated Metadata Enrichment



Why Not Both?



Use Cases


Ontology Development
Building human-in-the-loop tools for sourcing terminology and definitions from documents.
Fact Validation
Finding support for <s,p,o> triples in unstructured text, enhancing confidence in those facts.
KG Construction
Collecting potential triples from text to add them to a knowledge graph, including new entities and relations.

Machinery Part I: Embedding Text

Word Vectors

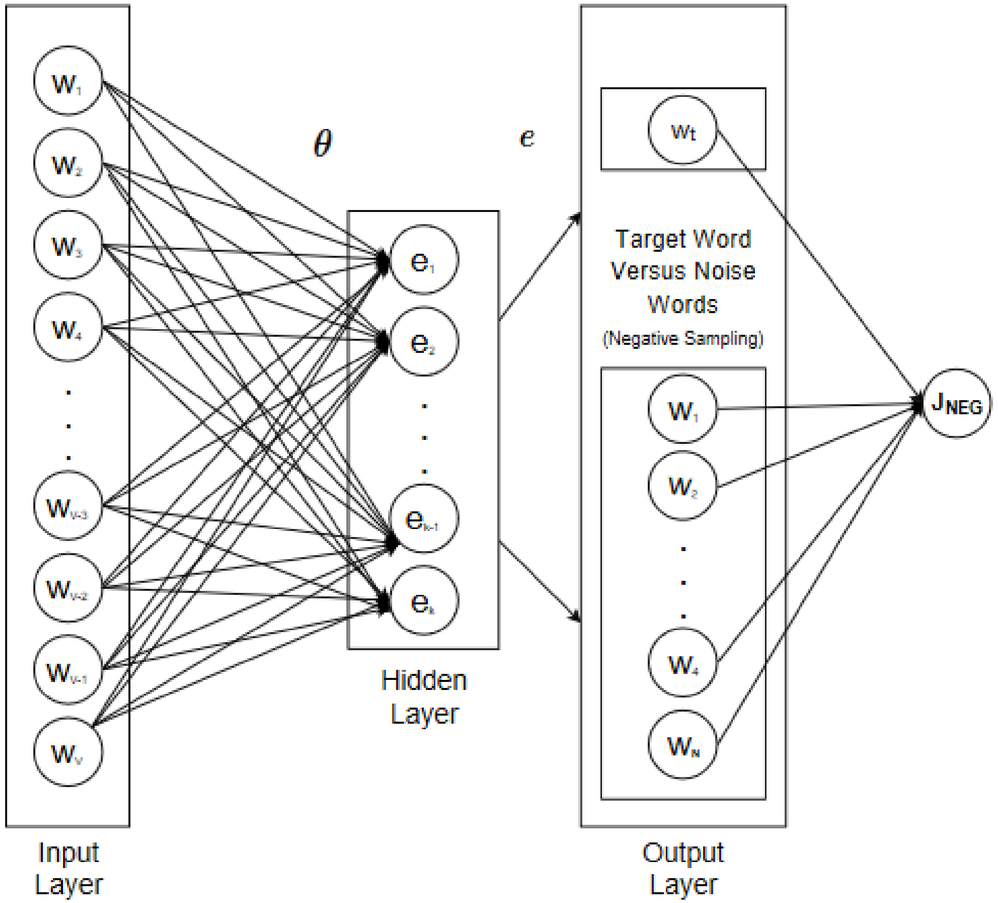
Representing Words as Dense Vectors

Assume that our text documents have been processed, parsed and tokenized – i.e. each individual word can be identified as a distinct element. How can we represent these words as features for machine learning?

“You shall know a word by the company it keeps”

- J.R. Firth, 1957

W2V



Setting up the Model:

0) We choose the model parameters

- Embedding dimension K
- Negative sampling rate N

1) Each word in the vocabulary V is input to the **input layer** (these are the one-hot encodings)

2) A matrix **θ** (V rows by K columns) represents the edge weights, initialized randomly

3) A matrix **e** (K rows by V columns) represents the embeddings, initialized randomly

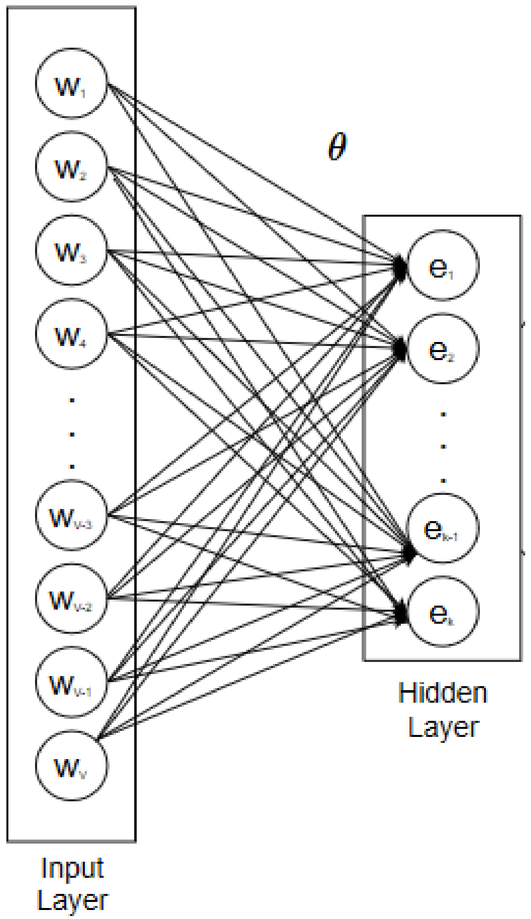
4) Words are **sampled** according to the negative sampling rate

5) **Probabilities** are computed – the output of which is J_{NEG}

$$J_{NEG} = \log Q_\theta(D = 1|w_t, e) + N\mathbb{E}_{\tilde{w}}[\log Q_\theta(D = 0|\bar{w}, e)]$$

Goal: We aim to change the model parameters θ in order to minimize J. We do this through gradient descent and backpropagation

W2V

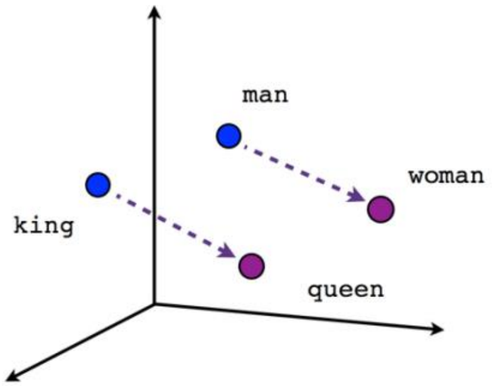


We keep the results of this hidden layer and use them to represent each word as a vector of floats.

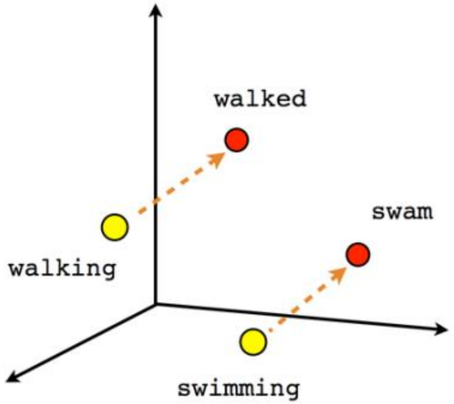
Word Vectors

We then throw away the final layers of the network and just keep the matrix θ – these are the final word vectors

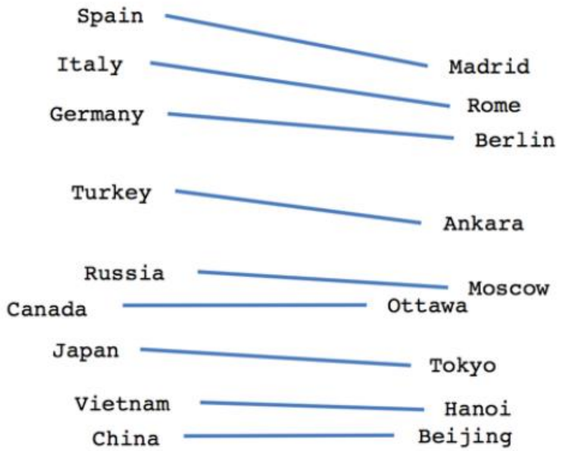
Semantics Captured



Male-Female



Verb tense



Country-Capital

https://www.tensorflow.org/tutorials/text/word2vec

Increased Complexity

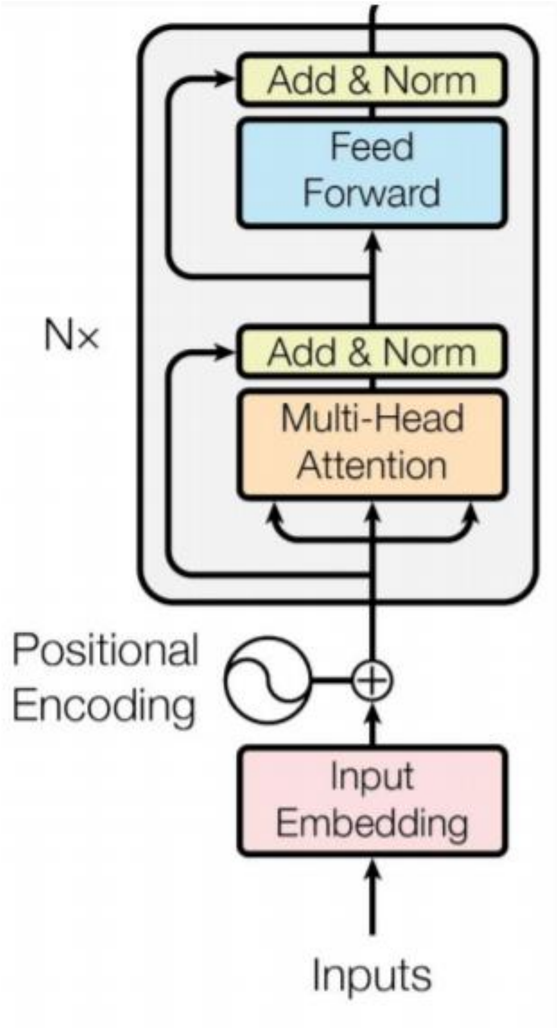
Words, Sentences, Documents

Not only are we interested in embedding words, but also entire sentences and documents.

- GloVe
- FastText

Spurred major research advances and new models:

- Recurrent Neural Networks
  - InferSent
  - QuickThought
  - SkipThought
- Transformer Methods
  - ELMo
  - BERT



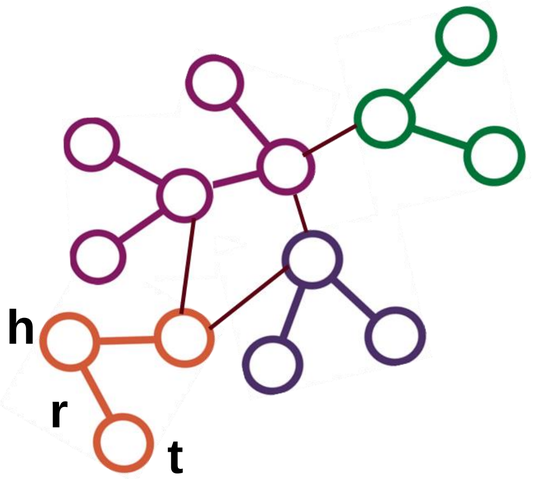
Attention Is All You Need. Vaswani et. Al. Google Brain. 2017.

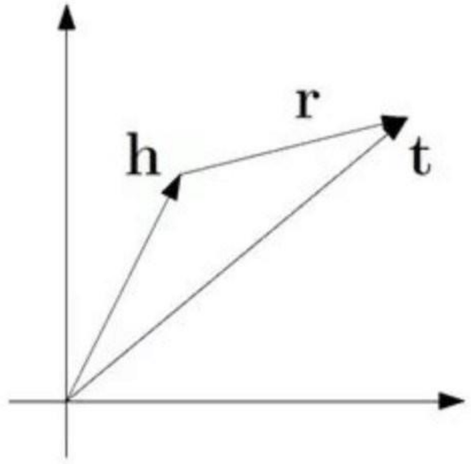
Machinery Part II:
Embedding Graphs

Graph Vectors

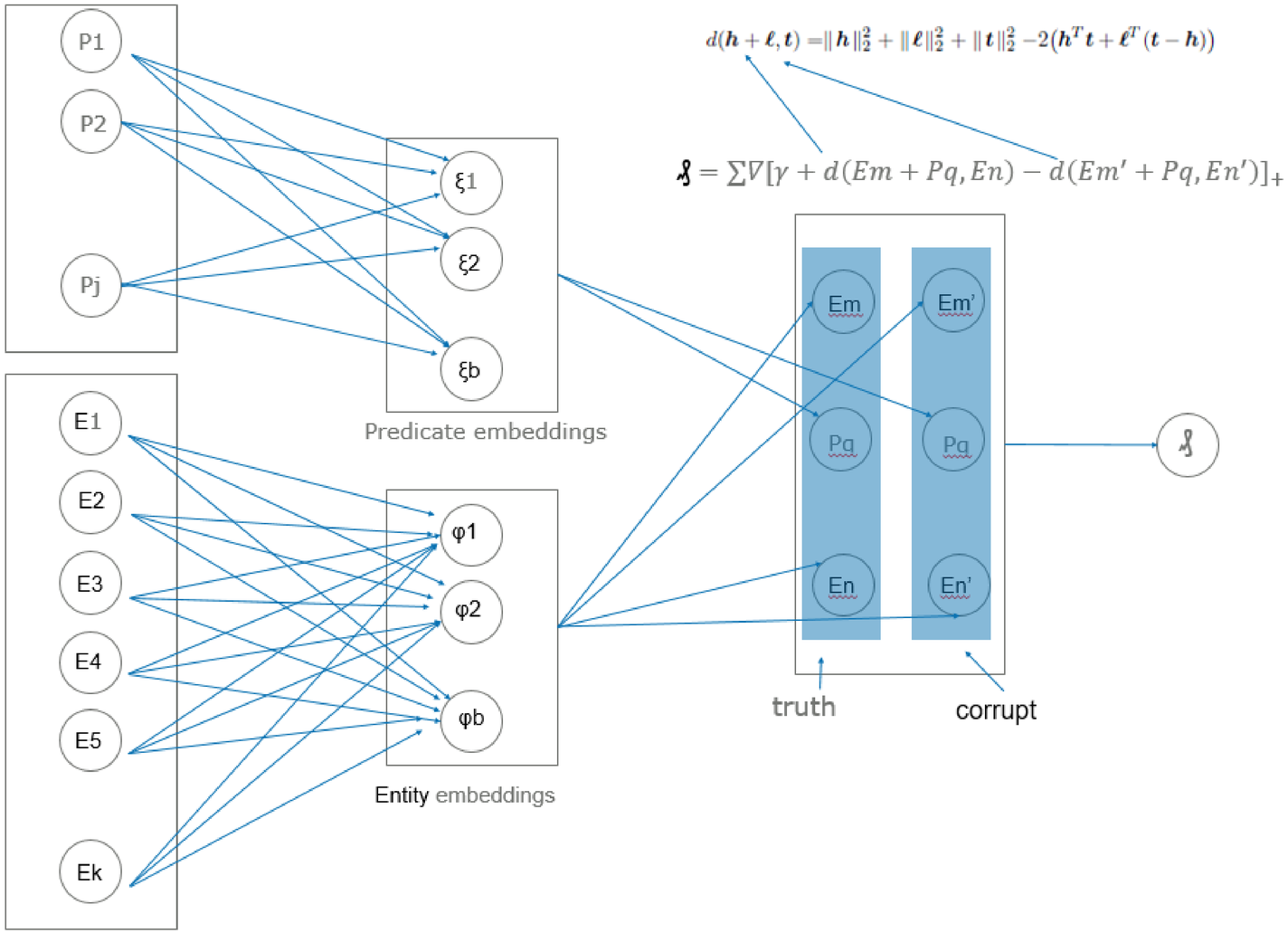
Using the same distributional hypothesis that underpinned word vectors, can we leverage the same types of approaches to encode the nodes and edges of a knowledge graph?

Use the same ‘semantic translation’ results from word vectors to treat relations as operations between entities.





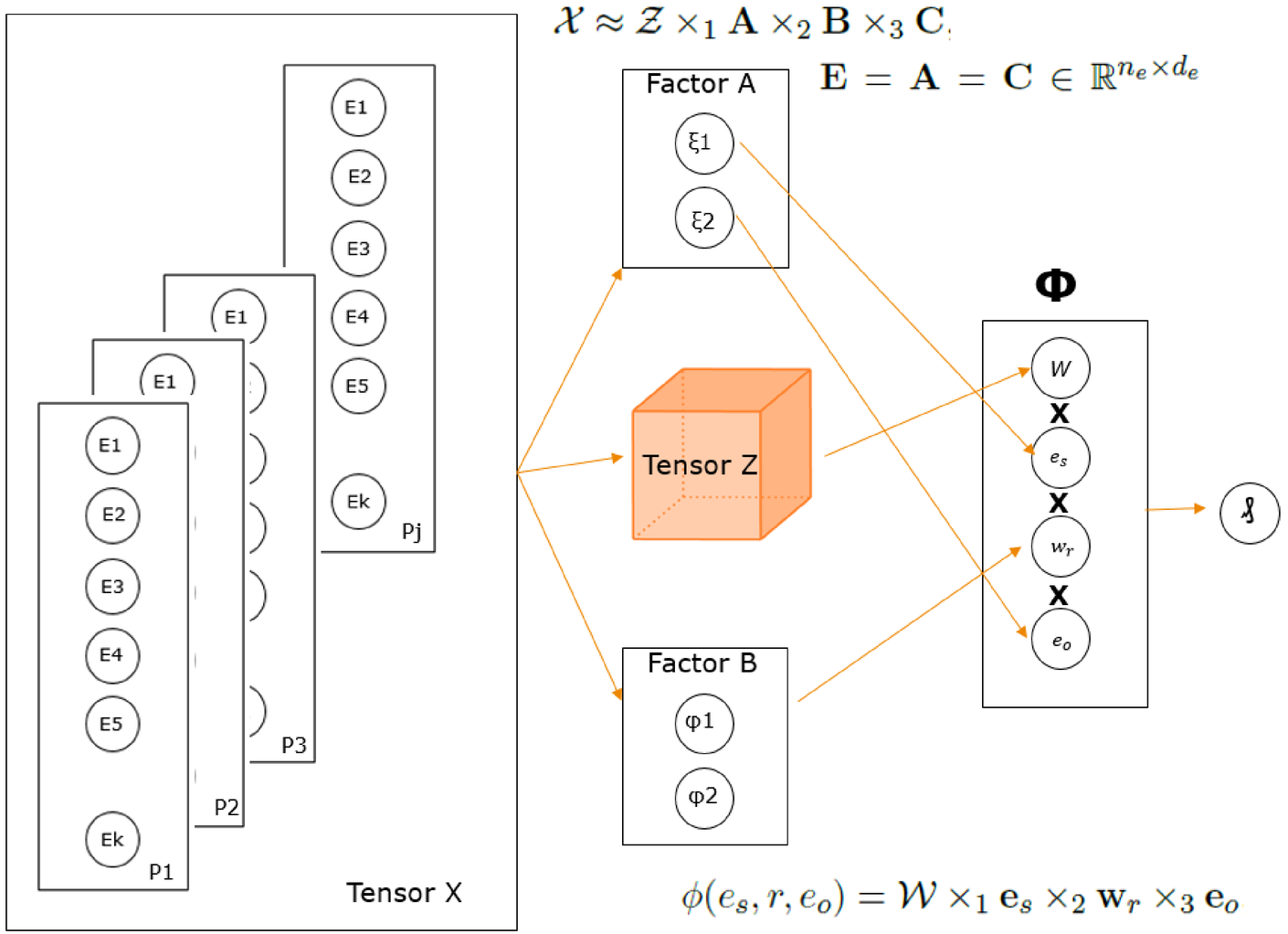
TransE



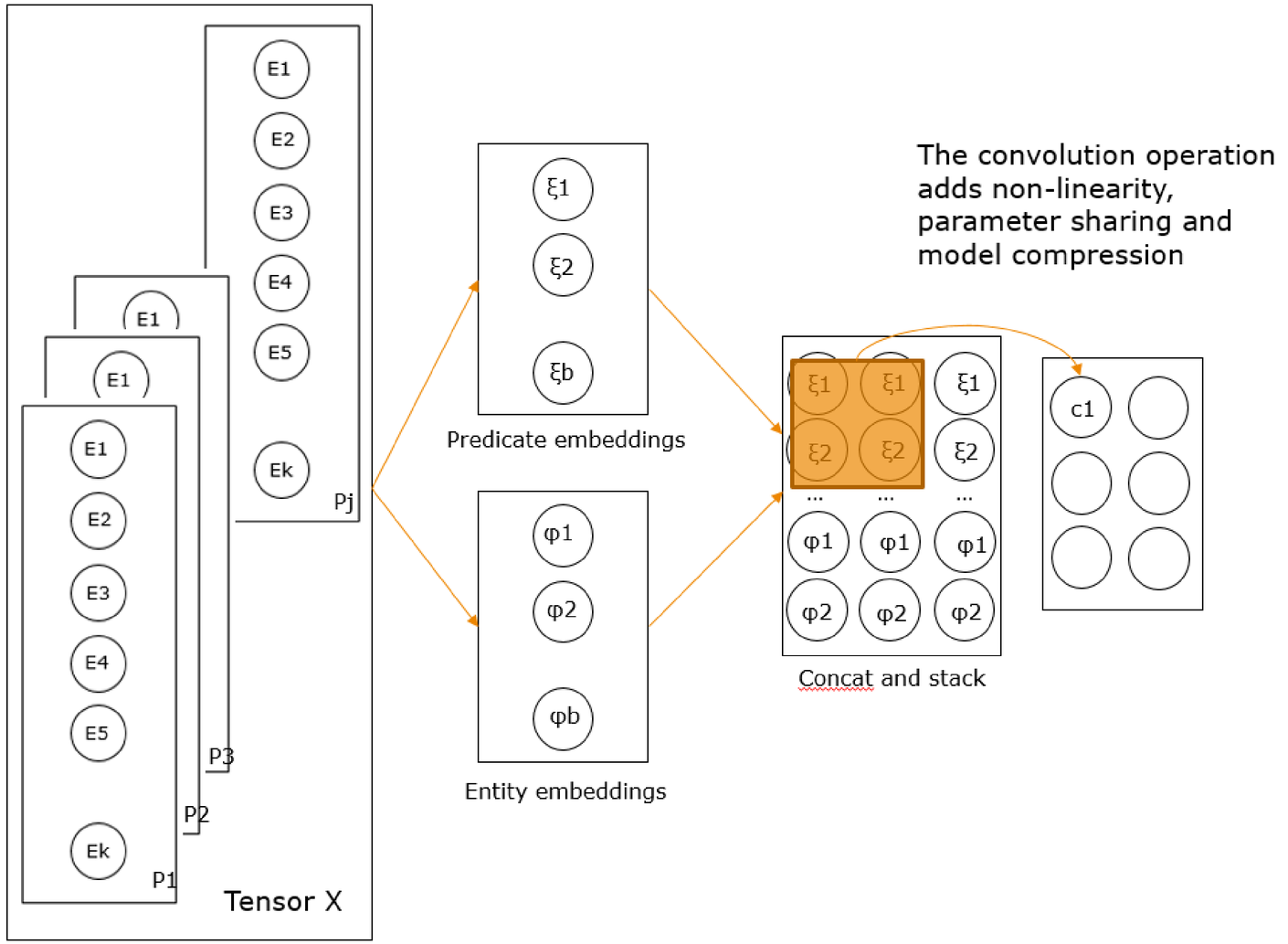
- $\mathcal{L}$ is the loss function
- γ is the margin (hyperparameter)
- b is the embedding dimension
- d is the distance function (squared Euclidean)

Used mainly for knowledge base completion (i.e. link prediction)

TuckER



ConvE


E1
E2
E3
E4
E5
Ek
Pj
P3
P2
P1
Tensor X
ξ1
ξ2
ξb
Predicate embeddings
φ1
φ2
φb
Entity embeddings
Concat and stack
c1
The convolution operation adds non-linearity, parameter sharing and model compression

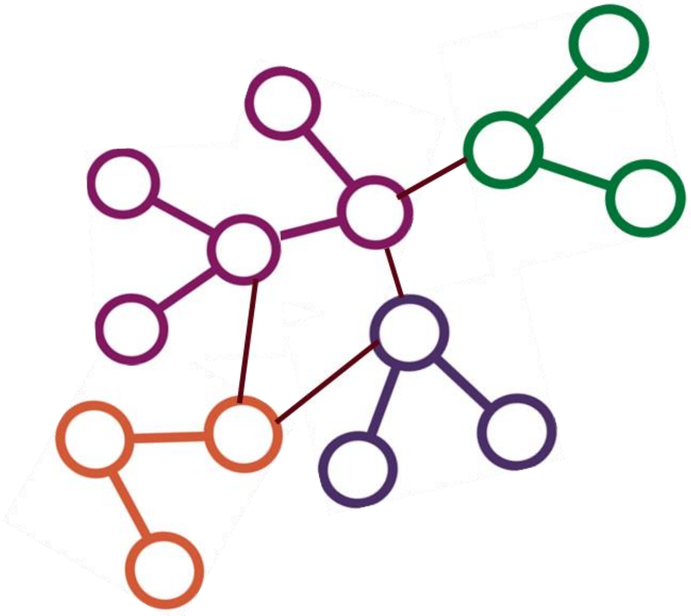
Machinery Part III: Alignment

Linear Map

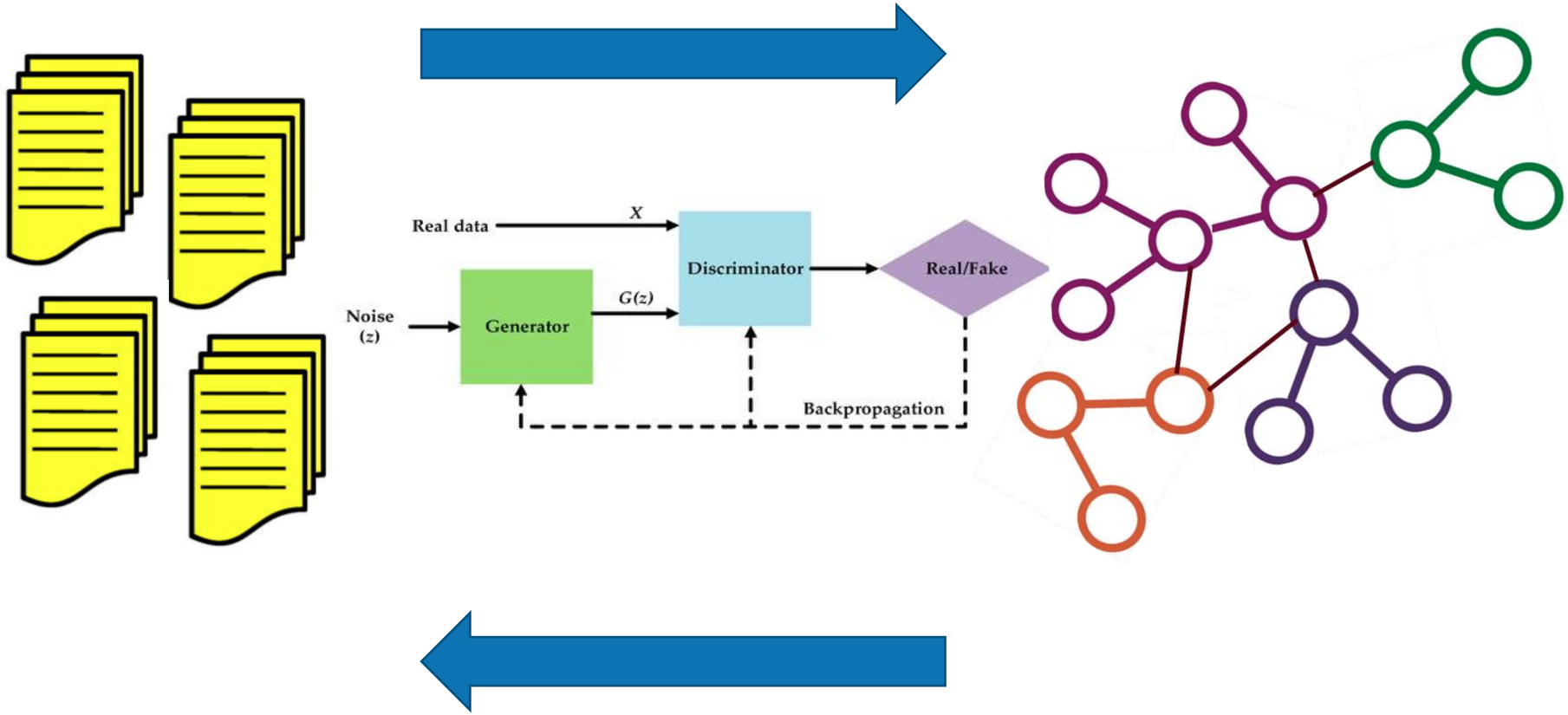
$$M: A \rightarrow B$$

$$\begin{pmatrix} a_{11} & a_{12} & \dots & a_{1m} \\ a_{21} & a_{22} & \dots & a_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \dots & a_{nm} \end{pmatrix}$$

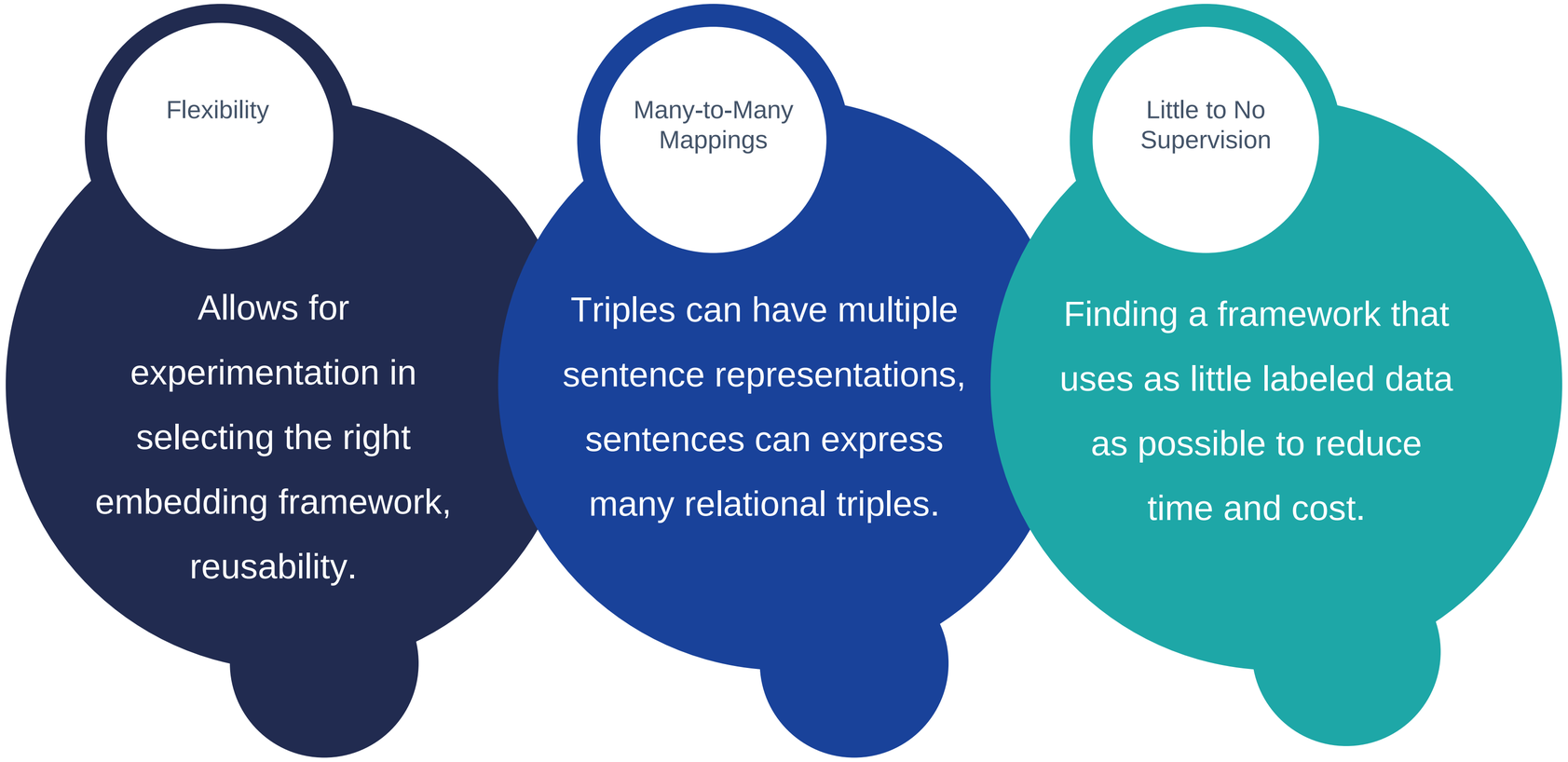
$$M^{-1}: B \rightarrow A$$



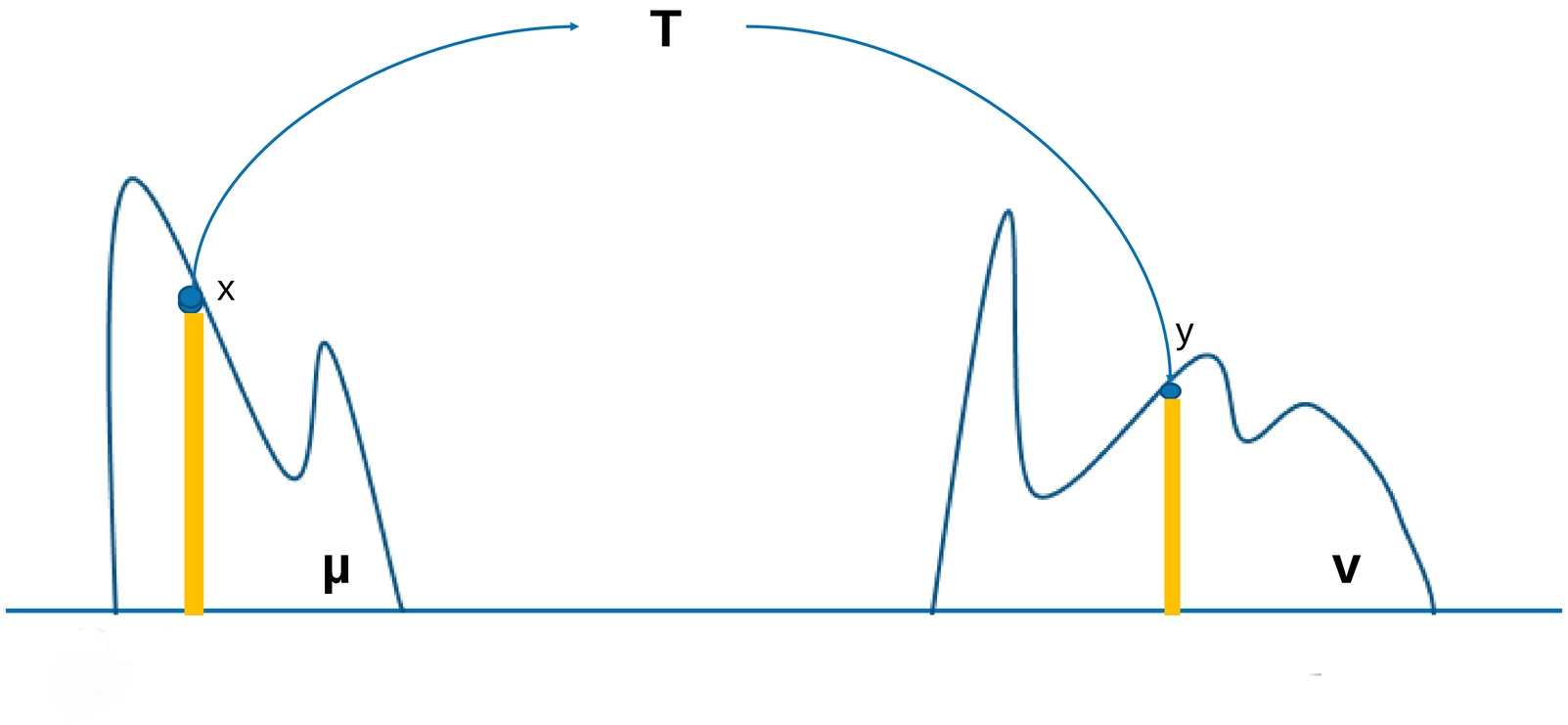
Generative Adversarial Network



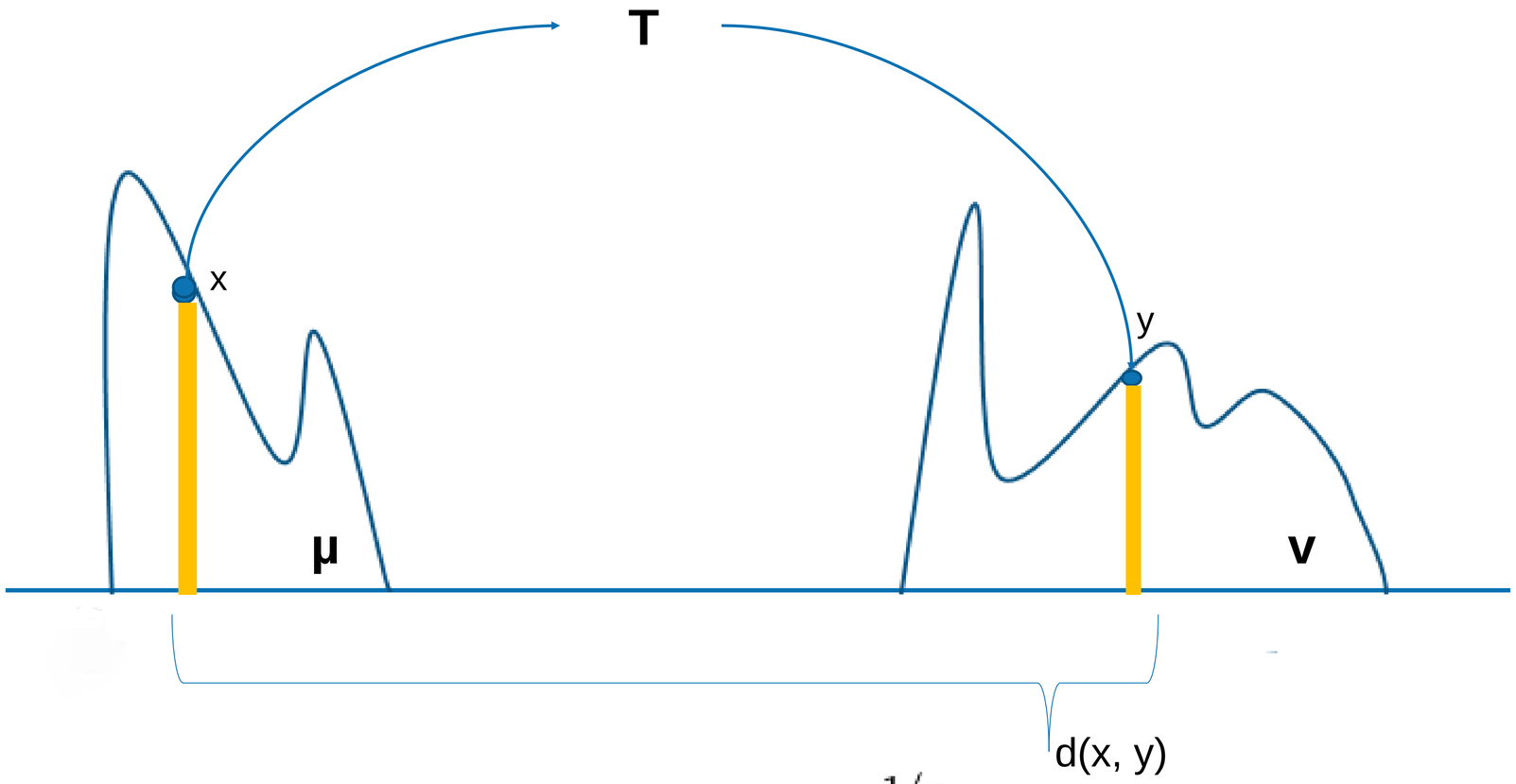
Framework Goals



Optimal Transport

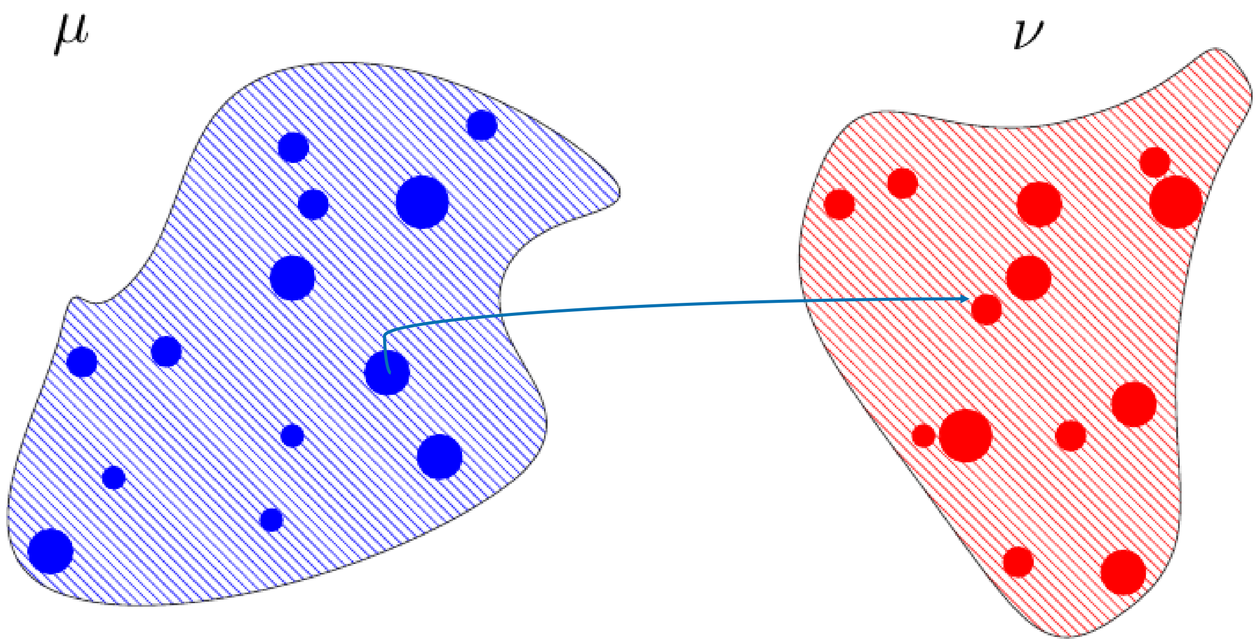


Wasserstein Distance



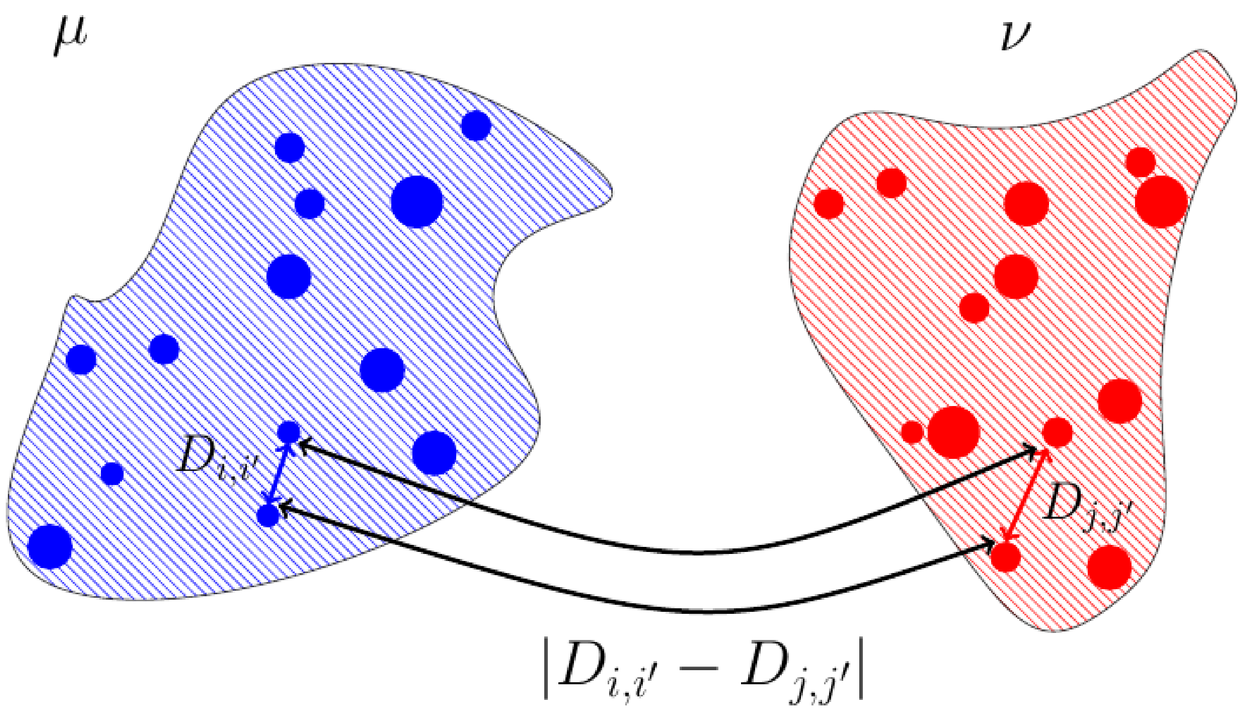
$$W_p(\mu,\nu) := \left(\inf_{\gamma \in \Gamma(\mu,\nu)} \int_{M\times M} d(x,y)^p \, \mathrm{d}\gamma(x,y) \right)^{1/p}$$

What Distance Function?



Without labeled pairs of sentences and triples, the distance between a sentence embedding space and a graph embedding space can’t be defined!

Gromov-Wasserstein Distance



Build distances within each embedding space, and use those relationships to define a cross-space distance.

Entropic Sinkhorn Algorithm

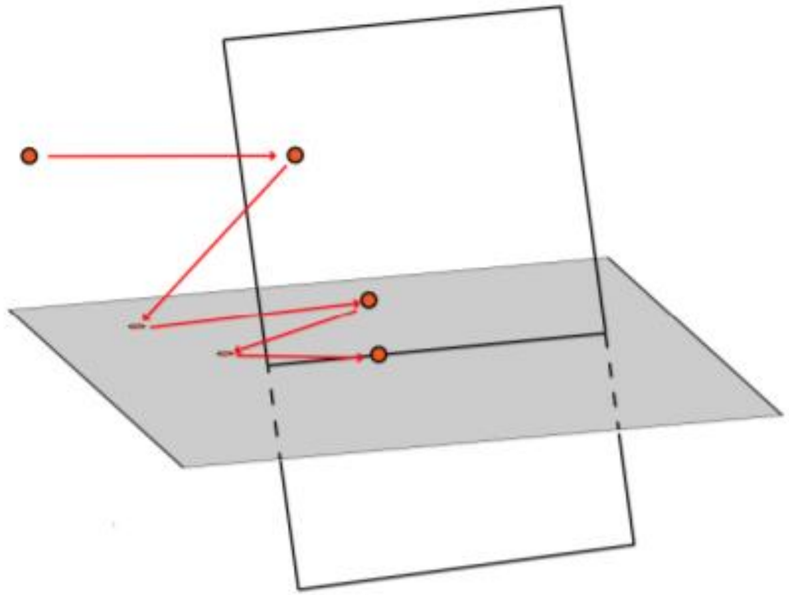


Figure 13.1: The Sinkhorn algorithm.
Starting from some initial values for u and v, we alternatingly project between the gray rectangle, representing the space of all matrices with row sums equal to a, and the white rectangle, representing the space of all matrices with column sums equal to b. The algorithm eventually converges at the intersection of the rectangles, representing the set of matrices with row sums equal to a and column sums equal to b. This would be the optimal solution.

http://www.math.ubc.ca/~geoff/courses/W2019T1/Lecture13.pdf

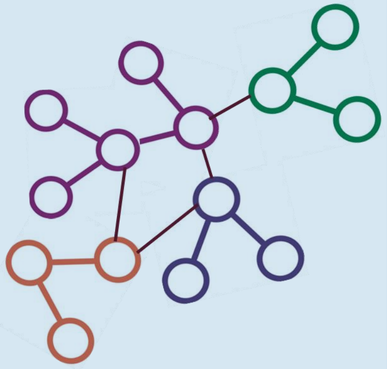
Machinery Part IV:
End-to-End System

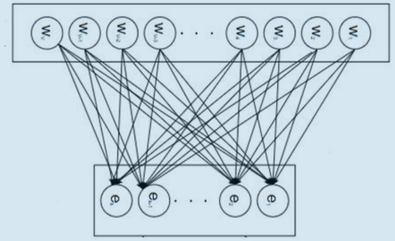
Bootstrapped data labels

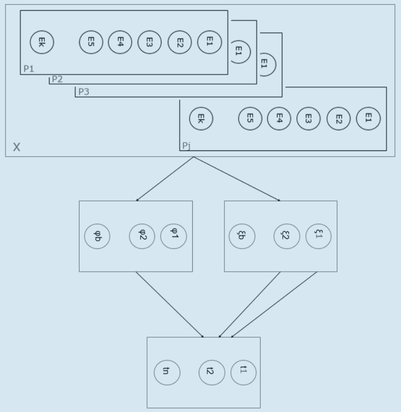
Data layer

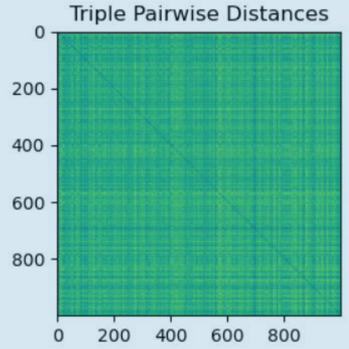
Embedding layer

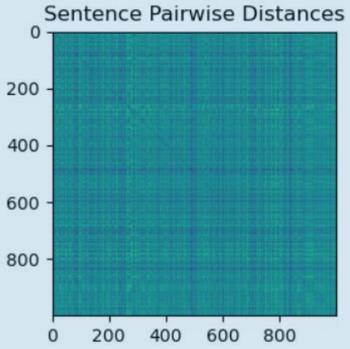
Comparison layer



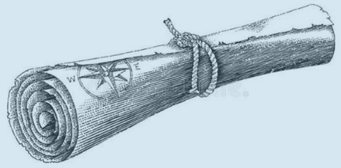


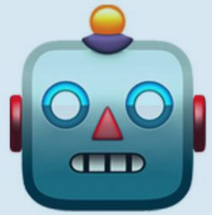






OT Model





Human-in-the-loop validation

Next Steps

Choosing representations

Which class(es) of embedding algorithms lend themselves best to aligning KG and text?

Better triple representations

Most KG embedding algorithms build separate entity and predicate representations. Are there better methods for representing a triple? Are there benchmark datasets for triple similarity?

Open Alignment Benchmarks

Development of KG/ontology and text alignment datasets are required for proper end-to-end evaluation of systems. Are there good open-source benchmarks that can be leveraged?

#KGC2021

Join the Conversation

@KGConference

linkedin.com/company/the-knowldge-graph-conference/

youtube.com/playlist?list=PLAiy7NYe9U2Gjg-600CTV1HGypiF95d_D

KGC The Knowledge Graph Conference