Interpreting KG Relation Representation from Word Embeddings

Carl Allen[1*] Ivana Balažević[1*] Timothy Hospedales[1,2]

{carl.allen, ivana.balazevic, t.hospedales}@ed.ac.uk

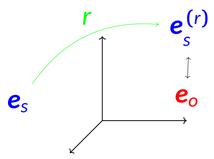
[1] School of Informatics, University of Edinburgh, UK [2] Samsung AI Centre, Cambridge, UK

Abstract: How do KG Embeddings Capture Semantics?

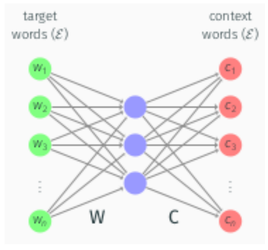
- There are many knowledge graph (KG) representation models, yet **little is understood of the latent structure they learn**.
- Building on recent **word embedding** theory (Allen and Hospedales, 2019; Allen et al., 2019), we derive **geometric properties** of knowledge graph relation representations (**relation conditions**) that map word embeddings of a subject entity to those of related object entities.
- We show that the better a model's architecture satisfies a relation's conditions, the better its performance at link prediction.

Background: Knowledge Graph Representation

- KGs store facts: binary **relations** between **entities** (e_s, r, e_o).
- Enable computational reasoning over KGs, e.g. question answering and inferring new facts (**link prediction**).
- Requires representation, typically:
  - each entity by a vector **embedding** $\boldsymbol{e} \in \mathbb{R}^d$,
  - each relation by a **transformation** from subject entity to object entity,
- A **score function** measures proximity between transformed entity embeddings.



Simplify: Consider Word Embeddings



$\mathcal{E}$ = dictionary of all words

Word2vec (Mikolov et al., 2013) loss function is minimised when:

- **Low-rank case** (Levy and Goldberg, 2014):

$$\mathbf{w}_i^\top \mathbf{c}_j = \underbrace{\log \frac{p(w_i, c_j)}{p(w_i)p(c_j)}}_{\mathbf{PMI}(w_i,\, c_j)} - \log k \doteq \mathbf{S}_{i,j}$$

- **General case** (Allen et al., 2019): Word embeddings $\mathbf{w}_i$ are a (non-linear) projection of **PMI vectors** ($\mathbf{p}^i$).

PMI vectors: $\mathbf{p}^i = \left\{\log \frac{p(c_j|w_i)}{p(c_j)}\right\}_{c_j \in \mathcal{E}} = \log \frac{p(\mathcal{E}|w_i)}{p(\mathcal{E})}$

PMI Vector Interactions

Similarity: similar words induce similar distributions over context words. Subtraction of PMI vectors finds similarity (e.g. synonyms):

$$\mathbf{p}^i - \mathbf{p}^j = \log \frac{p(\mathcal{E}|w_i)}{p(\mathcal{E}|w_j)} = \rho^{i,j}$$

Paraphrases: word sets with similar aggregate semantic meaning, e.g. {man, royal} ≈ king. Addition of PMI vectors finds paraphrases:

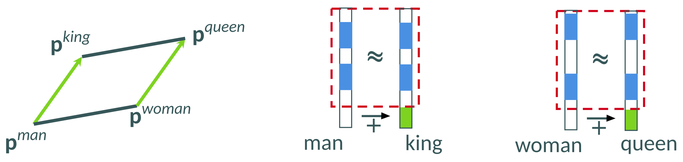
$$\begin{aligned}\mathbf{p}^i + \mathbf{p}^j &= \log \frac{p(\mathcal{E}|w_i)}{p(\mathcal{E})} + \log \frac{p(\mathcal{E}|w_j)}{p(\mathcal{E})} \\ &= \mathbf{p}^k + \underbrace{\log \frac{p(\mathcal{E}|w_i,w_j)}{p(\mathcal{E}|w_k)}}_{\rho^{\{i,j\},k}} - \underbrace{\log \frac{p(w_i,w_j|\mathcal{E})}{p(w_i|\mathcal{E})p(w_j|\mathcal{E})}}_{\sigma^{i,j}} + \underbrace{\log \frac{p(w_i,w_j)}{p(w_i)p(w_j)}}_{\tau^{i,j}}\end{aligned}$$

Analogies: word pairs that share a similar semantic difference, e.g. {man, king} and {woman, queen}. A linear combination of PMI vectors identifies analogies (subject to similar error terms, Allen and Hospedales (2019)):

$$\mathbf{p}^{b^*} - \mathbf{p}^b \approx \mathbf{p}^{a^*} - \mathbf{p}^a$$

All interactions *linear*
$\implies$ preserved in **word embeddings** under (≈) linear projection

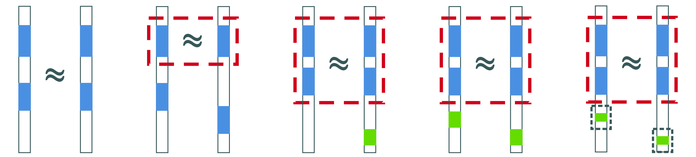
From Analogies to KG Relations



Analogy Relation

- Analogies contain common **binary word relations**, similar to KGs.
- For certain analogies ("specialisations"), the associated "vector offset" gives an **affine transformation** that represents the relation.
- Not all relations fit this semantic pattern, but we can now consider geometric aspects (**relation conditions**) of other relation types.

Categorising Relations: Semantics ⟶ Relation Conditions



Similarity **R**elatedness **S**pecialisation **C**ontext-shift Gen. cont.-shift

Categorical completeness: view PMI vectors as **sets** of word features and relation types as **set operations**.

- similarity ⇒ set equality
- relatedness ⇒ subset equality (relation-specific)
- (gen.) context-shift ⇒ set difference (relation-specific)

Relation Conditions ⟶ Mappings between Embeddings

R: $\mathcal{S}$-relatedness requires $\boldsymbol{e}_s$ and $\boldsymbol{e}_o$ to share a subspace component $\mathbb{V}_S$

- project onto $\mathbb{V}_S$ (multiply by matrix $\boldsymbol{P}_r \in \mathbb{R}^{d\times d}$) and compare.
- Dot product: $(\boldsymbol{P}_r\boldsymbol{e}_s)^\top(\boldsymbol{P}_r\boldsymbol{e}_o) = \boldsymbol{e}_s^\top \boldsymbol{P}_r^\top \boldsymbol{P}_r \boldsymbol{e}_o = \boldsymbol{e}_s^\top \boldsymbol{M}_r \boldsymbol{e}_o$
- Euclidean distance: $\|\boldsymbol{P}_r\boldsymbol{e}_s - \boldsymbol{P}_r\boldsymbol{e}_o\|^2 = \|\boldsymbol{P}_r\boldsymbol{e}_s\|^2 - 2\boldsymbol{e}_s^\top \boldsymbol{M}_r \boldsymbol{e}_o + \|\boldsymbol{P}_r\boldsymbol{e}_o\|^2$

S/C: requires $\mathcal{S}$-relatedness and relation-specific component(s) ($\boldsymbol{v}_r^s$, $\boldsymbol{v}_r^o$).

- project onto a subspace corresponding to $\mathcal{S}$, $\boldsymbol{v}_r^s$ and $\boldsymbol{v}_r^o$ (i.e. test $\mathcal{S}$-relatedness while preserving relation-specific components);
- add relation-specific $\boldsymbol{r} = \boldsymbol{v}_r^o - \boldsymbol{v}_r^s \in \mathbb{R}^d$ to transformed embeddings.
- Dot product: $(\boldsymbol{P}_r\boldsymbol{e}_s + \boldsymbol{r})^\top \boldsymbol{P}_r\boldsymbol{e}_o$
- Euclidean distance: $\|\boldsymbol{P}_r\boldsymbol{e}_s + \boldsymbol{r} - \boldsymbol{P}_r\boldsymbol{e}_o\|^2$

References

Carl Allen and Timothy Hospedales. Analogies explained: Towards understanding word embeddings. In *International Conference on Machine Learning*, 2019.

Carl Allen, Ivana Balažević, and Timothy Hospedales. What the Vec? Towards Probabilistically Grounded Embeddings. In *Advances in Neural Information Processing Systems*, 2019.

Omer Levy and Yoav Goldberg. Neural word embedding as implicit matrix factorization. In *Advances in Neural Information Processing Systems*, 2014.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, 2013.