Distantly Supervised Relation Extraction with Sentence Reconstruction and Knowledge Base Priors

Fenia Christopoulou   Makoto Miwa   Sophia Ananiadou

The University of Manchester
The National Centre for Text Mining
National Institute of Advanced Industrial Science and Technology AIST
AIRC
Toyota Technological Institute

Distantly Supervised Relation Extraction (DSRE)

Automatically annotate corpora with relation pairs using a Knowledge Base (KB) as source

Distantly Supervised Relation Extraction (DSRE)

Automatically annotate corpora with relation pairs using a Knowledge Base (KB) as source

Relaxed assumption [Riedel et al., 2010]

There is at least one sentence expressing the relation of a pair in a KB

Entity 1	*Entity 2*	*Relation*
Steve Jobs	Apple	/business/company/founders
Ray Nagin	New Orleans	/people/person/place_of_birth

Bag annotated with the relation "founders"

Among other reasons , **Apple** 's chief executive , **Steve Jobs** , …
About **Apple** 's **Steve Jobs** , who bought out …
bag 1

Bag annotated with the relation "place_of_birth"

Mayor **Ray Nagin** born in **New Orleans** has already …
C. **Ray Nagin** , the mayor of **New Orleans** , …
bag 2

Distantly Supervised Relation Extraction (DSRE)

Automatically annotate corpora with relation pairs using a Knowledge Base (KB) as source

Goal: Identify the relation of the bag from a *predefined set of relations*

➜ *Multi-label classification* problem (one bag can have multiple relations)

Entity 1	*Entity 2*	*Relation*
Steve Jobs	Apple	/business/company/founders
Ray Nagin	New Orleans	/people/person/place_of_birth

Bag annotated with the relation "founders"

Among other reasons , **Apple** 's chief executive , **Steve Jobs** , …
About **Apple** 's **Steve Jobs** , who bought out …
bag 1

Bag annotated with the relation "place_of_birth"

Mayor **Ray Nagin** born in **New Orleans** has already …
C. **Ray Nagin** , the mayor of **New Orleans** , …
bag 2

Prior Work

- Advantages of Distantly Supervised Relation Extraction (DSRE)
    - Automatically annotate raw data with relations
    - Use distantly annotated data for KB augmentation [Ji and Grishman, 2011]

Prior Work

- Advantages of Distantly Supervised Relation Extraction (DSRE)
  - Automatically annotate raw data with relations
  - Use distantly annotated data for KB augmentation [Ji and Grishman, 2011]
- Disadvantages
  - Noisy instances → The relation is not expressed in any of the sentences
  - Long tail relations → Very few occurrences of certain relation categories
  - Unbalanced bag size → Most bags include only 1 sentence

Prior Work

- Advantages of Distantly Supervised Relation Extraction (DSRE)
  - Automatically annotate raw data with relations
  - Use distantly annotated data for KB augmentation [Ji and Grishman, 2011]
- Disadvantages
  - Noisy instances → The relation is not expressed in any of the sentences
  - Long tail relations → Very few occurrences of certain relation categories
  - Unbalanced bag size → Most bags include only 1 sentence

Existing approaches use

- Attention mechanisms [Lin et al., 2016; Ye and Ling, 2019]
- Reinforcement learning [Qinet al., 2018b; Wu et al., 2019]
- Relation type hierarchies, Entity descriptors [She et al., 2018; Zhang et al., 2019; Hu et al., 2019]
- Information from KBs (e.g. entity types, relation aliases) [Vashishth et al., 2018]
- Additional training data [Beltagy et al., 2019], Pre-trained Language Models [Alt et al., 2019]

Prior Work

- Advantages of Distantly Supervised Relation Extraction (DSRE)
  - Automatically annotate raw data with relations
  - Use distantly annotated data for KB augmentation [Ji and Grishman, 2011]
- Disadvantages
  - Noisy instances → The relation is not expressed in any of the sentences
  - Long tail relations → Very few occurrences of certain relation categories
  - Unbalanced bag size → Most bags include only 1 sentence

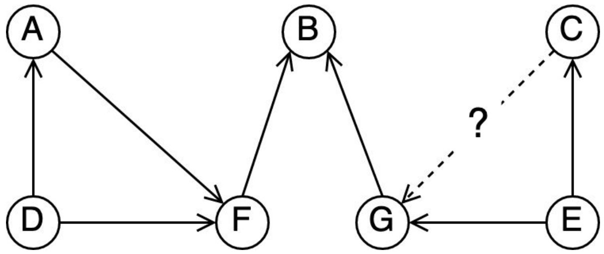
This work

Existing approaches use

- Attention mechanisms [Lin et al., 2016; Ye and Ling, 2019]
- Reinforcement learning [Qinet al., 2018b; Wu et al., 2019]
- Relation type hierarchies, Entity descriptors [She et al., 2018; Zhang et al., 2019; Hu et al., 2019]
- Information from KBs (e.g. entity types, relation aliases) [Vashishth et al., 2018]
- Additional training data [Beltagy et al., 2019], Pre-trained Language Models [Alt et al., 2019]

Incorporating KB Information

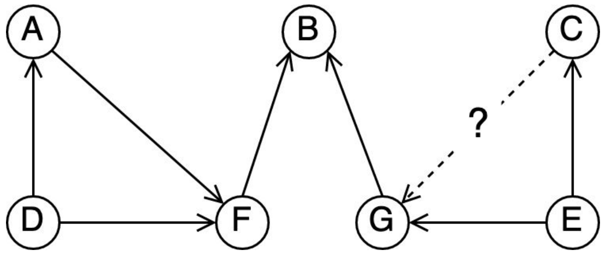
Take advantage of Link Prediction (find missing relations in Knowledge Graphs)



Incorporating KB Information

Take advantage of Link Prediction (find missing relations in Knowledge Graphs)

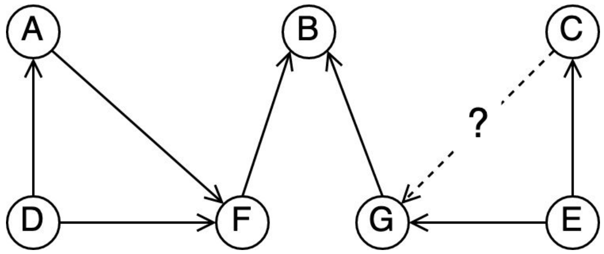
- Explicit agreement of sentence- and KB-level classifications [Weston et al., 2013; Xu and Barbosa, 2019]
- KB embeddings as attention queries [Han et al.,2018; Hu et al., 2019]
- Minimise the distance between KB and sentence representations [Wang et al., 2018]



Incorporating KB Information

Take advantage of Link Prediction (find missing relations in Knowledge Graphs)

- Explicit agreement of sentence- and KB-level classifications [Weston et al., 2013; Xu and Barbosa, 2019]
- KB embeddings as attention queries [Han et al.,2018; Hu et al., 2019]
- Minimise the distance between KB and sentence representations [Wang et al., 2018]



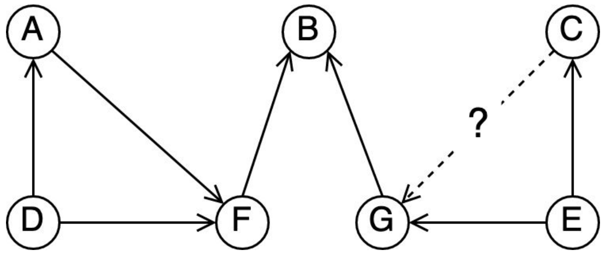
➜ Rigid connection between context-agnostic (KB) and context-aware (sentences) pairs

➜ Need representations of entities on the test set → Poor generalisation to unseen examples

Incorporating KB Information

Take advantage of Link Prediction (find missing relations in Knowledge Graphs)

- Explicit agreement of sentence- and KB-level classifications [Weston et al., 2013; Xu and Barbosa, 2019]
- KB embeddings as attention queries [Han et al.,2018; Hu et al., 2019]
- Minimise the distance between KB and sentence representations [Wang et al., 2018]



➜ Rigid connection between context-agnostic (KB) and context-aware (sentences) pairs

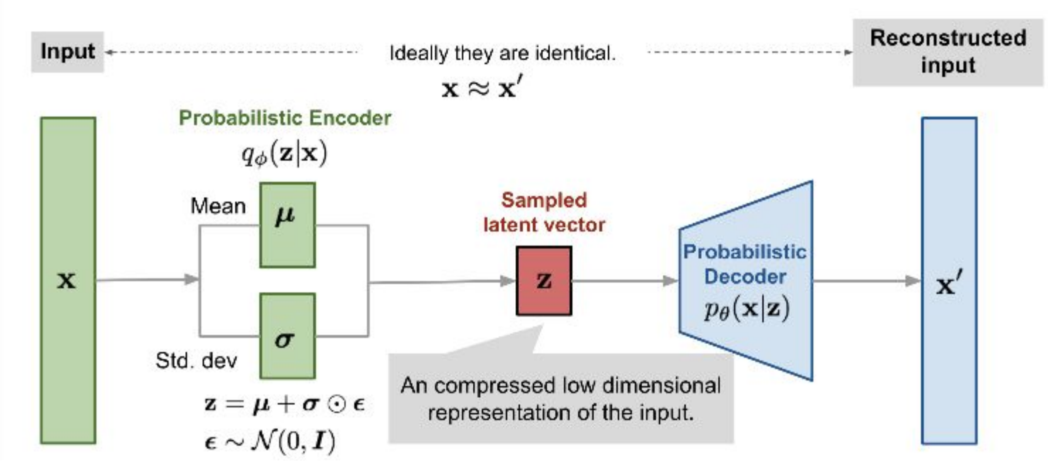
➜ Need representations of entities on the test set → Poor generalisation to unseen examples

Use KB signals to promote generalisation to unseen entity pairs via a probabilistic approach

Bring closer sentences containing the same KB pairs

Proposed Approach: Main Idea

1. Variational Autoencoders (VAEs) [Kingma and Welling, 2013]
   - Latent variable encoder-decoder models
   - Parameterise posterior distributions using neural networks
   - Learn an effective latent space influenced by a prior distribution
   - Sentence reconstruction helps sentence expressivity by learning semantic or syntactic similarities in the sentence space

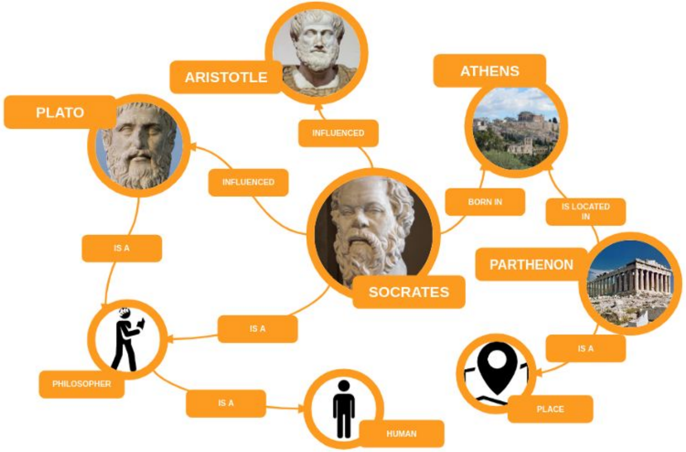


Source: lilianweng

Proposed Approach: Main Idea

1. Variational Autoencoders (VAEs) [Kingma and Welling, 2013]
    - Latent variable encoder-decoder models
    - Parameterise posterior distributions using neural networks
    - Learn an effective latent space influenced by a prior distribution
    - Sentence reconstruction helps sentence expressivity by learning semantic or syntactic similarities in the sentence space

2. Information from Knowledge Graphs
    - Detection of factual relations



Souce: towardsdatascience

Proposed Approach: Main Idea

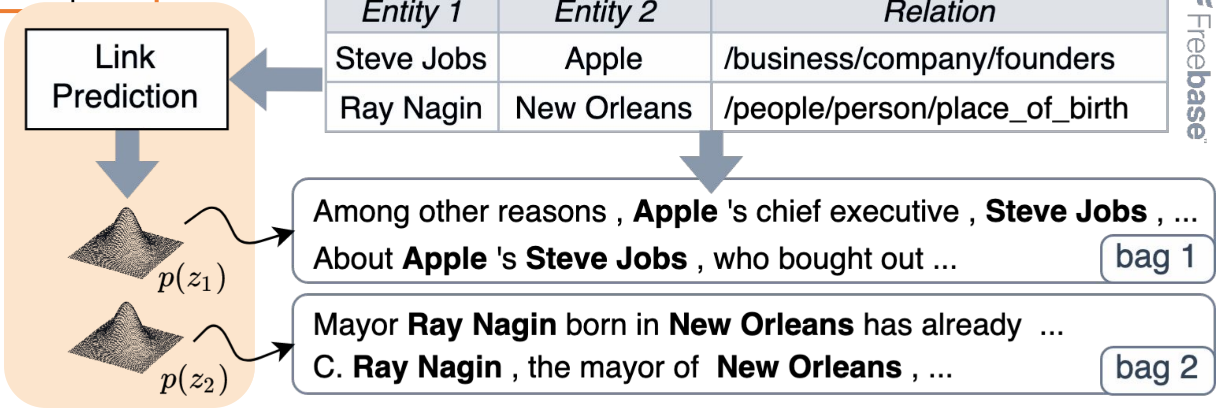
1. Variational Autoencoders (VAEs) [Kingma and Welling, 2013]
   - Latent variable encoder-decoder models
   - Parameterise posterior distributions using neural networks
   - Learn an effective latent space influenced by a prior distribution
   - Sentence reconstruction helps sentence expressivity by learning semantic or syntactic similarities in the sentence space

Combination in a multi-task learning setting

2. Information from Knowledge Graphs
   - Detection of factual relations

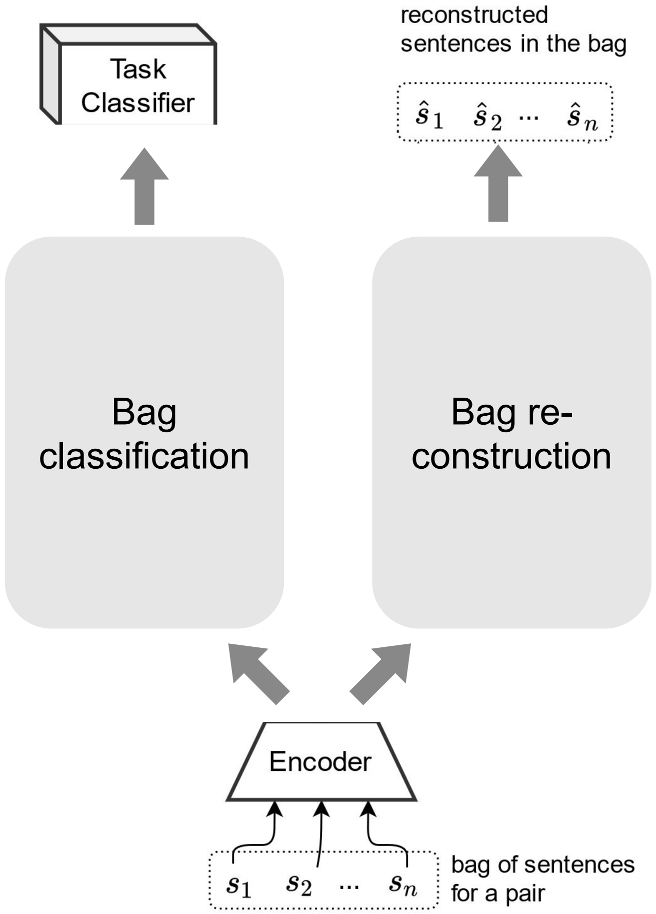
Create informative priors to assist bag classification

Entity 1	*Entity 2*	*Relation*
Steve Jobs	Apple	/business/company/founders
Ray Nagin	New Orleans	/people/person/place_of_birth



Methodology

- Model input:
  - An entity pair e_1, e_2
  - A bag of sentences $B = \{s_1, s_2, \ldots, s_n\}$ that contain the pair
- Model output:
  - Predicted relations for the given pair
  - Reconstructed sentences in the bag
- 2 Branches
  - Left: Classifier with selective attention
  - Right: VAE

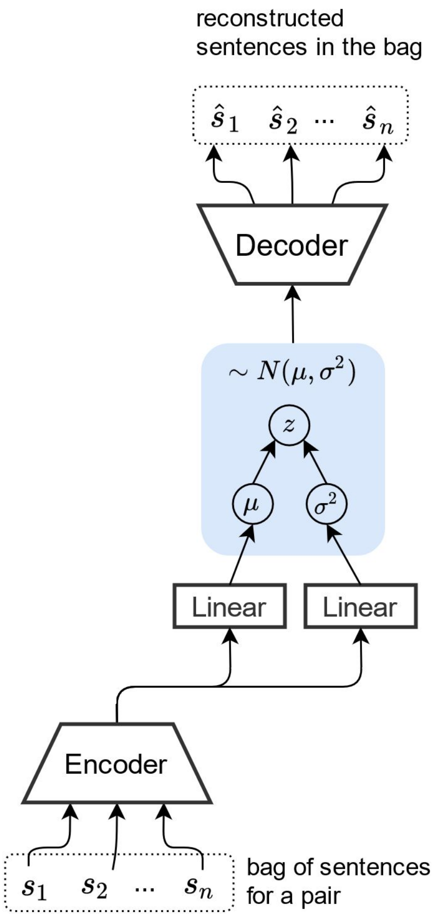


Bag Reconstruction

- *Encoder*: BiLSTM [Hochreiter et al., 1997]
- The last hidden and cell states of the encoder are used to construct the parameters of a multivariate Gaussian

$$\boldsymbol{\mu} = \mathbf{W}_{\mu}[\mathbf{h};\mathbf{c}] + \mathbf{b}_{\mu}, \quad \boldsymbol{\sigma}^2 = \mathbf{W}_{\sigma}[\mathbf{h};\mathbf{c}] + \mathbf{b}_{\sigma}$$

representing the feature space of the sentence



Bag Reconstruction

- *Encoder*: BiLSTM [Hochreiter et al., 1997]
- The last hidden and cell states of the encoder are used to construct the parameters of a multivariate Gaussian

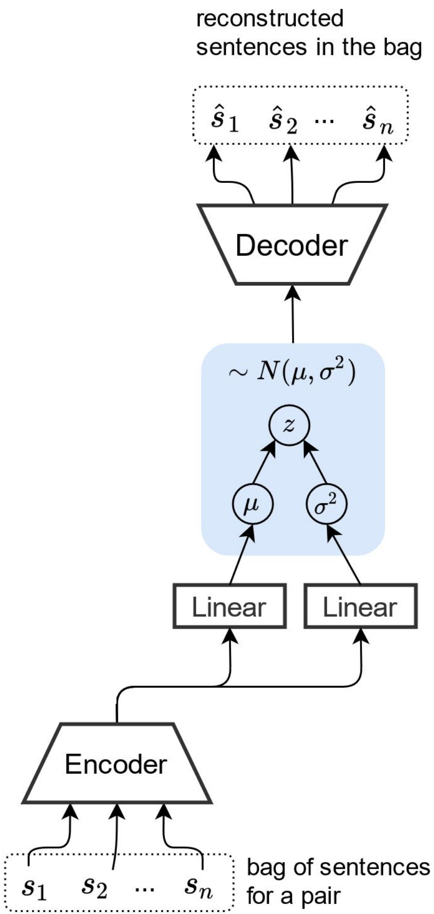
$$\boldsymbol{\mu} = \mathbf{W}_{\mu}[\mathbf{h}; \mathbf{c}] + \mathbf{b}_{\mu}, \quad \boldsymbol{\sigma}^2 = \mathbf{W}_{\sigma}[\mathbf{h}; \mathbf{c}] + \mathbf{b}_{\sigma}$$

representing the feature space of the sentence

- Re-parameterisation trick [Kingma and Welling, 2013]

$$\mathbf{z} = \boldsymbol{\mu} + \boldsymbol{\sigma} \odot \boldsymbol{\epsilon}, \text{ where } \boldsymbol{\epsilon} \sim \boxed{\mathcal{N}(\mathbf{0}, \mathbf{I})}$$

Prior is assumed the Normal Distribution



Bag Reconstruction

- *Encoder*: BiLSTM [Hochreiter et al., 1997]
- The last hidden and cell states of the encoder are used to construct the parameters of a multivariate Gaussian

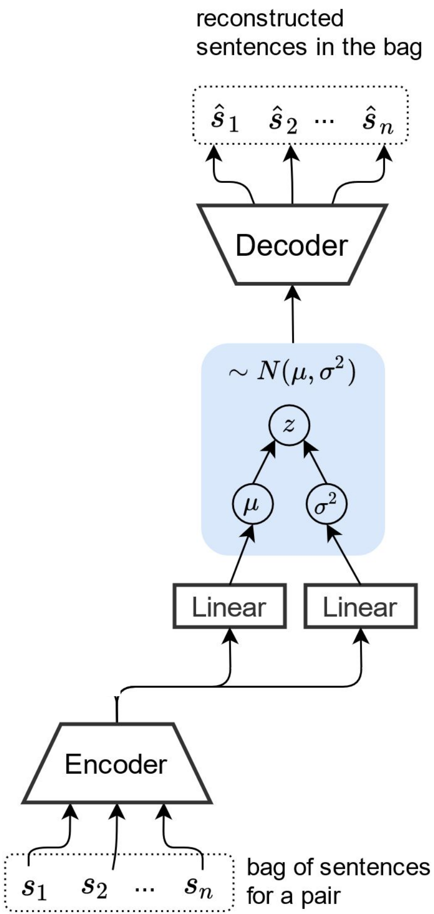
$$\boldsymbol{\mu} = \mathbf{W}_{\mu}[\mathbf{h};\mathbf{c}] + \mathbf{b}_{\mu}, \quad \boldsymbol{\sigma}^2 = \mathbf{W}_{\sigma}[\mathbf{h};\mathbf{c}] + \mathbf{b}_{\sigma}$$

  representing the feature space of the sentence

- Re-parameterisation trick [Kingma and Welling, 2013]

$$\mathbf{z} = \boldsymbol{\mu} + \boldsymbol{\sigma} \odot \boldsymbol{\epsilon}, \text{ where } \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}).$$

- *Decoder*: Unidirectional LSTM
  - Fed the latent code z following Bowman et al. (2016)



Bag Reconstruction

- *Encoder*: BiLSTM [Hochreiter et al., 1997]
- The last hidden and cell states of the encoder are used to construct the parameters of a multivariate Gaussian

$$\boldsymbol{\mu} = \mathbf{W}_{\mu}[\mathbf{h};\mathbf{c}] + \mathbf{b}_{\mu}, \quad \boldsymbol{\sigma}^2 = \mathbf{W}_{\sigma}[\mathbf{h};\mathbf{c}] + \mathbf{b}_{\sigma}$$

representing the feature space of the sentence

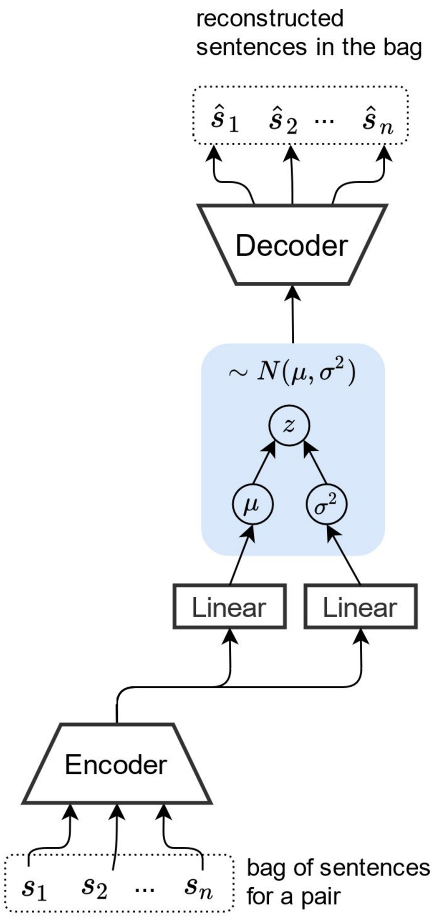
- Re-parameterisation trick [Kingma and Welling, 2013]

$$\mathbf{z} = \boldsymbol{\mu} + \boldsymbol{\sigma} \odot \boldsymbol{\epsilon}, \text{ where } \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}).$$

- *Decoder*: Unidirectional LSTM
  - Fed the latent code z following Bowman et al. (2016)

- Learning: Minimize Evidence LOwer Bound (ELBO)

$$L_{\text{ELBO}} = \mathbb{E}_{z \sim q_{\phi}(z|h)} \left[\log(p_{\theta}(\mathbf{h}|\mathbf{z}))\right] - D_{\text{KL}}\left(q_{\phi}(\mathbf{z}|\mathbf{h})||p_{\theta}(\mathbf{z})\right)$$



Bag Reconstruction

- *Encoder*: BiLSTM [Hochreiter et al., 1997]
- The last hidden and cell states of the encoder are used to construct the parameters of a multivariate Gaussian

$$\boldsymbol{\mu} = \mathbf{W}_{\mu}[\mathbf{h};\mathbf{c}] + \mathbf{b}_{\mu}, \quad \boldsymbol{\sigma}^2 = \mathbf{W}_{\sigma}[\mathbf{h};\mathbf{c}] + \mathbf{b}_{\sigma}$$

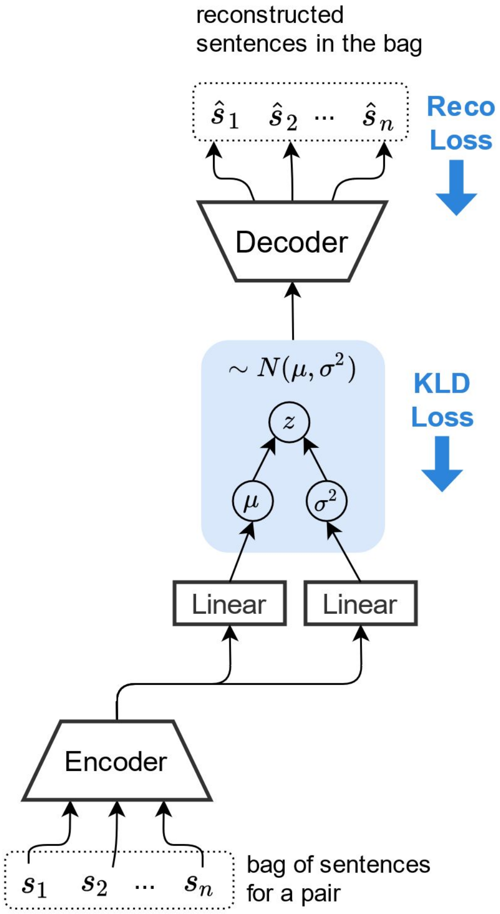
  representing the feature space of the sentence

- Re-parameterisation trick [Kingma and Welling, 2013]

$$\mathbf{z} = \boldsymbol{\mu} + \boldsymbol{\sigma} \odot \boldsymbol{\epsilon}, \text{ where } \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$$

- *Decoder*: Unidirectional LSTM
  - Fed the latent code z following Bowman et al. (2016)
- Learning: Minimize Evidence LOwer Bound (ELBO)

$$L_{\text{ELBO}} = \boxed{\mathbb{E}_{z\sim q_{\phi}(z|h)}\left[\log(p_{\theta}(\mathbf{h}|\mathbf{z}))\right]} \quad \text{Reconstruction Loss}$$
$$\boxed{- D_{\text{KL}}\left(q_{\phi}(\mathbf{z}|\mathbf{h}) || p_{\theta}(\mathbf{z})\right)} \quad \text{Kullback-Leibler divergence}$$

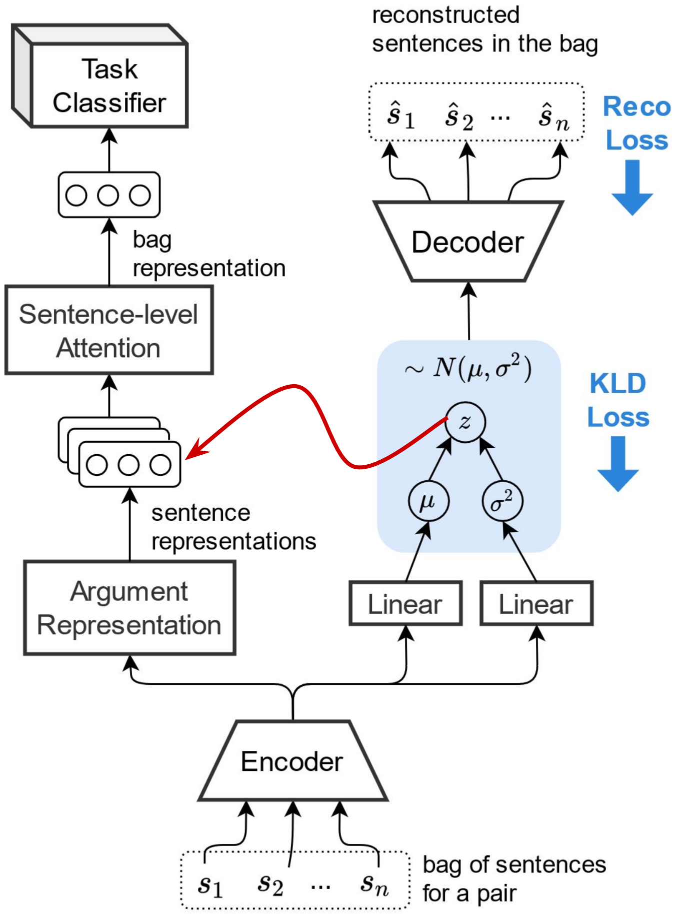


Bag Classification

Sentence Representation

- Create a sentence representation s using the latent code z and each entity of the pair

$$\mathbf{s} = \mathbf{W}_v[\mathbf{z}; \mathbf{e}_1; \mathbf{e}_2]$$



Bag Classification

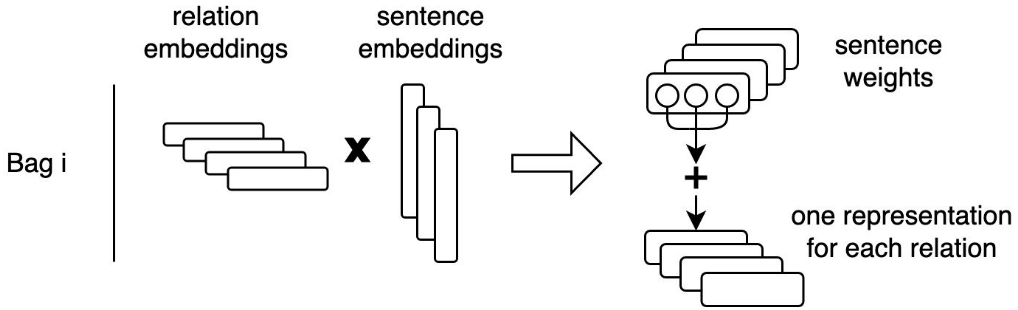
Sentence Representation

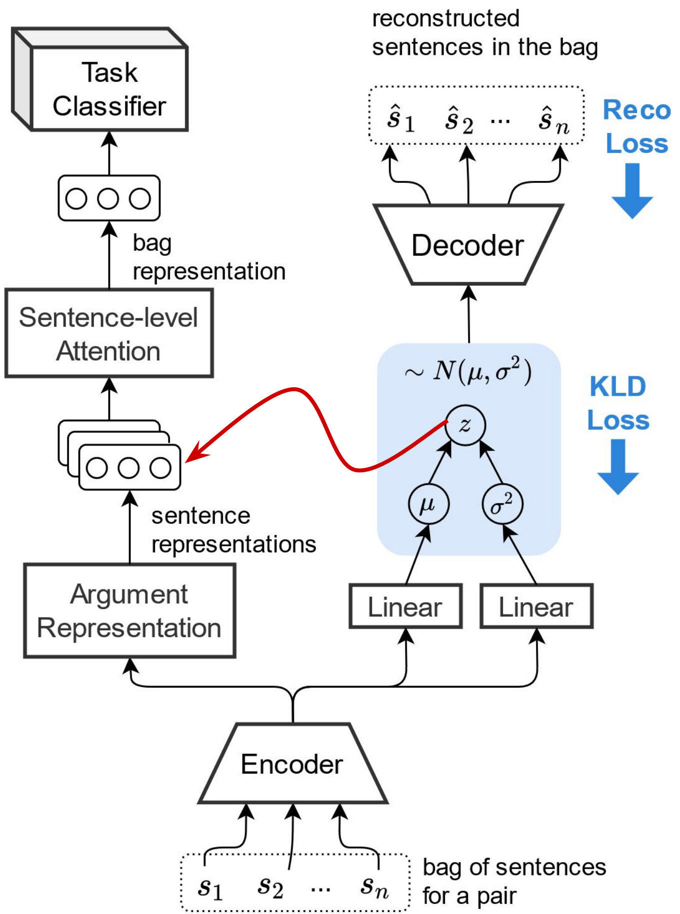
- Create a sentence representation s using the latent code z and each entity of the pair

$$\mathbf{s} = \mathbf{W}_v[\mathbf{z};\mathbf{e}_1;\mathbf{e}_2]$$

Bag Representation

- Use selective attention from Lin et al. (2016)





Bag Classification

Learning

- Use the respective bag relation embedding
- Binary cross entropy loss

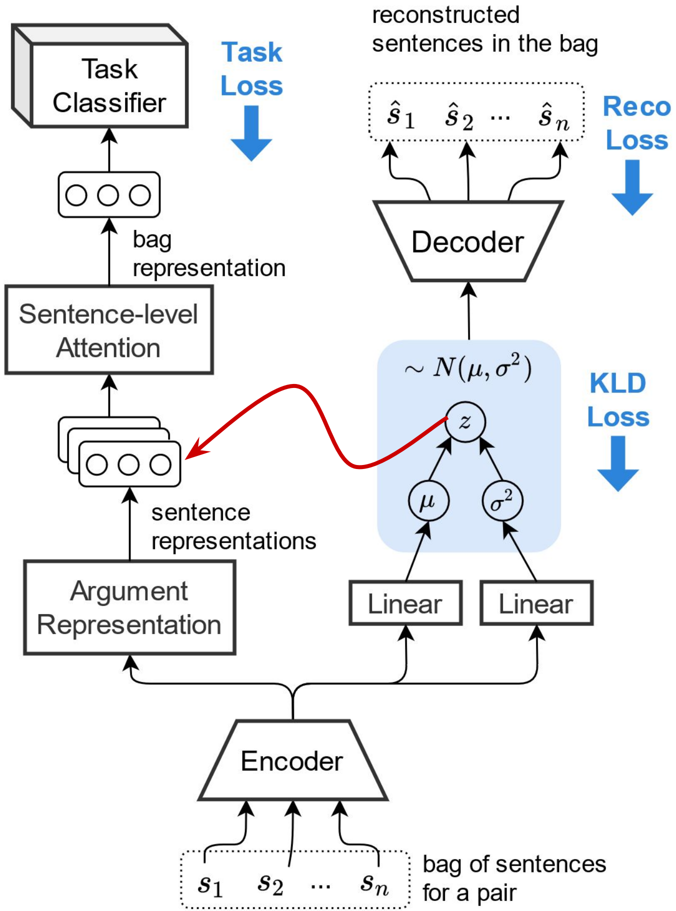
$$p(r=1|B) = \sigma(\mathbf{W}_c\, \mathbf{B}_r + \mathbf{b}_c)$$

$$L_{\text{BCE}} = -\sum_r y_r \log p(r|B) + (1-y_r)\log(1-p(r|B))$$

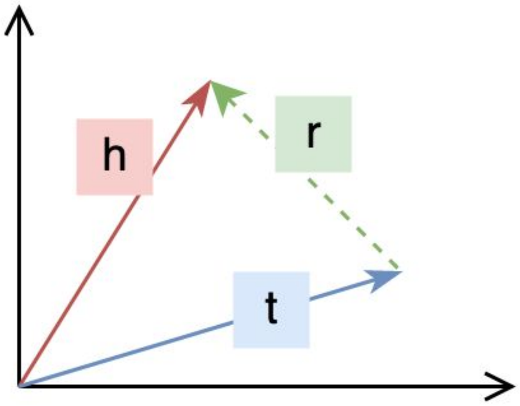
Training Objective

- Linear combination of VAE loss and task loss

$$L = \lambda\, L_{\text{BCE}} + (1-\lambda) L_{\text{ELBO}}$$



Knowledge Base Priors

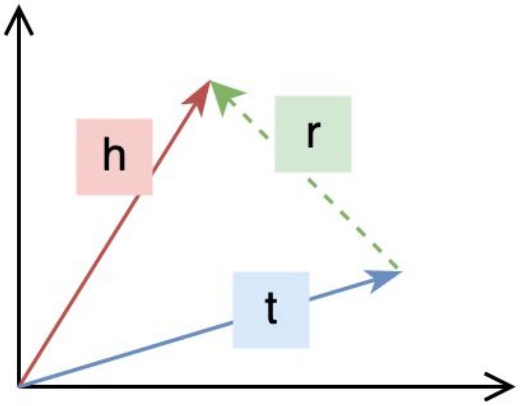


- Inject KB information into the model
- KB Priors:
  - Another Gaussian distribution
  - Mean value ~ KB pair representation
  - Covariance equal to the Identity Matrix
- TransE Link Prediction algorithm [Bordes et al., 2013]
  - Relations are represented as translations in the embedding space

$$p_\theta(\mathbf{z}) \sim \mathcal{N}(\boldsymbol{\mu}_{\mathrm{KB}}, \mathbf{I}), \text{ with } \boldsymbol{\mu}_{\mathrm{KB}} = \mathbf{e}_h - \mathbf{e}_t$$

Identity Covariance   Entity embeddings from TransE

Knowledge Base Priors



- Inject KB information into the model
- KB Priors:
  - Another Gaussian distribution
  - Mean value ~ KB pair representation
  - Covariance equal to the Identity Matrix
- TransE Link Prediction algorithm [Bordes et al., 2013]
  - Relations are represented as translations in the embedding space

Expect the sentence latent space to become similar to that of the KG

$$p_\theta(\mathbf{z}) \sim \mathcal{N}(\boldsymbol{\mu}_{\mathrm{KB}}, \mathbf{I}), \text{ with } \boldsymbol{\mu}_{\mathrm{KB}} = \mathbf{e}_h - \mathbf{e}_t$$

Identity Covariance   Entity embeddings from TransE

Experimental Settings

- Two distantly supervised datasets NYT-10 [Riedel et al., 2010], WikiDistant [Han et al., 2020]
- NYT-10:
  - 570K instances: Containing overlaps between train and test pairs
  - 520K instances: Clean data, no overlaps

Evaluation of both settings

- Knowledge Graphs used with TransE:
  - Freebase 3M entities [Xu et al., 2019], Wikidata 5M entities [Wang et al., 2019]

Dataset	Split	Instances	Bags	NA (%)
NYT10 # Relations: 53	Train	469,290	252,044	93.4
	Val.	53,321	28,109	93.5
	Test	172,448	96,678	97.9
WIKIDISTANT # Relations: 454	Train	1,050,246	575,620	64.8
	Val.	29,145	14,748	70.6
	Test	28,897	15,509	72.0

Baselines

- *Baseline*: Simple bag classification, no VAE component at all
- $p_\theta(z) \sim \mathcal{N}(0, \mathbf{I})$: Multi-task learning with Normal priors
- $p_\theta(z) \sim \mathcal{N}(\mu_{\mathrm{KB}}, \mathbf{I})$: Multi-task learning with KB priors

Proposed Approach

Prior Works:

- *PCNN-ATT*: Simple selective attention over instances in the bag [Lin et al., 2016]
- *Intra-Inter*: Intra-Inter bag attention [Ye and Ling, 2019]
- *JointNRE*: Joint training of Link Prediction and Bag classifications [Han et al., 2018]
- *RESIDE*: Additional KB information (entity types, relation aliases) [Vashishth et al., 2018]
- *DISTRE*: GPT-2 pre-trained language model [Alt et al., 2019]

Metrics:

- Area Under the Curve (AUC) score → Area under the Precision-Recall curve
- Precision at N (P@N) → Precision of the top N most confident predictions

Results: NYT-10

Version without overlaps

Method	Encoder	NYT 520K			
		AUC (%)	P@N (%)		
			100	200	300
Baseline		34.94	74.0	67.5	67.0
$+ p_\theta(z) \sim \mathcal{N}(0, I)$	BiLSTM	38.59	74.0	74.5	71.6
$+ p_\theta(z) \sim \mathcal{N}(\mu_{\text{KB}}, I)$		42.89	83.0	75.5	73.0
PCNN-ATT (Lin et al., 2016)	PCNN	32.66	71.0	67.5	62.6
JOINT NRE (Han et al., 2018)	CNN	30.62	60.0	57.0	55.3
RESIDE (Vashishth et al., 2018)	BiGRU	35.80	80.0	69.0	65.3
INTRA-INTER BAG (Ye and Ling, 2019)	PCNN	34.41	82.0	74.0	69.0
DISTRE (Alt et al., 2019)	GPT-2	42.20	68.0	67.0	65.3

- +4% boost in AUC over the Baseline with Normal priors
- +8% boost in AUC over the Baseline with KB priors
- Improve performance over a pre-trained language model (GPT-2)

Results: NYT-10

Version without overlaps → NYT 520K

Version with overlaps → NYT 570K

Method	Encoder	NYT 520K AUC (%)	NYT 520K P@N (%) 100	200	300	NYT 570K AUC (%)	NYT 570K P@N (%) 100	200	300
Baseline		34.94	74.0	67.5	67.0	43.59	84.0	77.0	75.3
$+ p_\theta(z) \sim \mathcal{N}(0, I)$	BiLSTM	38.59	74.0	74.5	71.6	44.64	80.0	76.0	75.6
$+ p_\theta(z) \sim \mathcal{N}(\mu_{\text{KB}}, I)$		42.89	83.0	75.5	73.0	45.52	81.0	77.5	73.6
PCNN-ATT (Lin et al., 2016)	PCNN	32.66	71.0	67.5	62.6	36.25	76.0	72.5	64.0
JOINT NRE (Han et al., 2018)	CNN	30.62	60.0	57.0	55.3	40.15	75.8	-	68.0
RESIDE (Vashishth et al., 2018)	BiGRU	35.80	80.0	69.0	65.3	41.60	84.0	78.5	75.6
INTRA-INTER BAG (Ye and Ling, 2019)	PCNN	34.41	82.0	74.0	69.0	42.20	91.8	84.0	78.7
DISTRE (Alt et al., 2019)	GPT-2	42.20	68.0	67.0	65.3	-	-	-	-

- Similar observations for the version with train-test pair overlaps
- Pair overlaps significantly benefit prior models
- Tail of the distribution is improved when including test pairs in the training set

Results: WikiDistant

Method	AUC (%)	P@N (%)		
		100	200	300
Baseline	28.54	94.0	93.0	88.3
$+ p_\theta(z) \sim \mathcal{N}(0, I)$	30.59	96.0	93.5	89.3
$+ p_\theta(z) \sim \mathcal{N}(\mu_{\mathrm{KB}}, I)$	29.54	92.0	89.0	90.0
PCNN-ATT (Han et al., 2020)	22.20	-	-	-

- KB Priors seem to not help
- We find that only 72% of training pairs are assigned a KB prior (vs 96% in NYT-10)
- Repeat experiments by removing 28% of the data

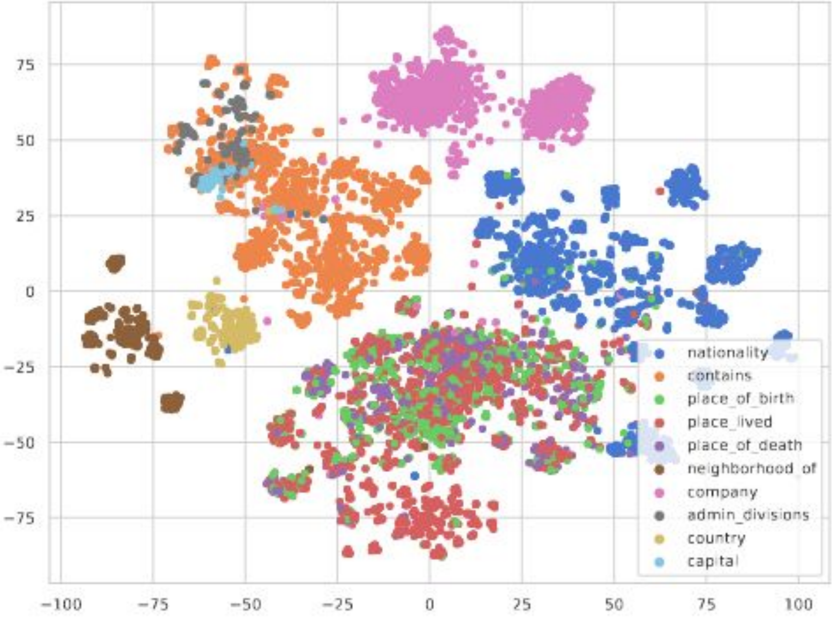
Results: WikiDistant

Method	AUC (%)	P@N (%)		
		100	200	300
Baseline	28.54	94.0	93.0	88.3
$+ p_\theta(z) \sim \mathcal{N}(0, I)$	30.59	96.0	93.5	89.3
$+ p_\theta(z) \sim \mathcal{N}(\mu_{\text{KB}}, I)$	29.54	92.0	89.0	90.0
PCNN-ATT (Han et al., 2020)	22.20	-	-	-
w/o non KB-prior pairs (72% of training pairs preserved)				
Baseline	26.16	88.0	85.0	82.6
$+ p_\theta(z) \sim \mathcal{N}(0, I)$	27.46	90.0	88.0	84.6
$+ p_\theta(z) \sim \mathcal{N}(\mu_{\text{KB}}, I)$	28.38	94.0	95.0	89.3

- KB Priors seem to not help
- We find that only 72% of training pairs are assigned a KB prior (vs 96% in NYT-10)
- Repeat experiments by removing 28% of the data
- Coverage of training pair priors is important

Analysis: Latent Space (NYT-10)

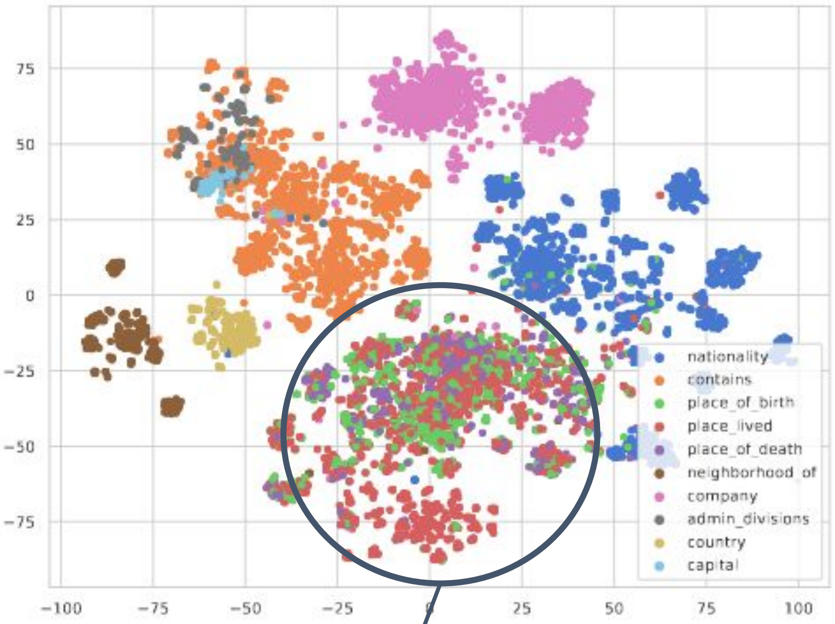
Prior Space



- t-SNE plots of TransE embeddings (prior space), VAE μ embeddings (posterior space)
- Top 10 most frequent relation categories

Analysis: Latent Space (NYT-10)

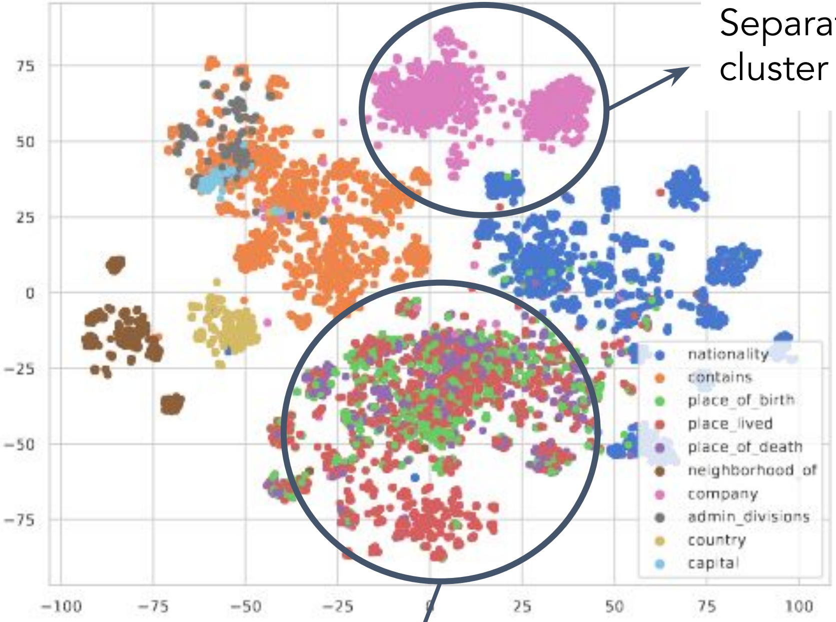
Prior Space



Overlapping region:
"place of birth", "place of death",
"placed lived"

Analysis: Latent Space (NYT-10)

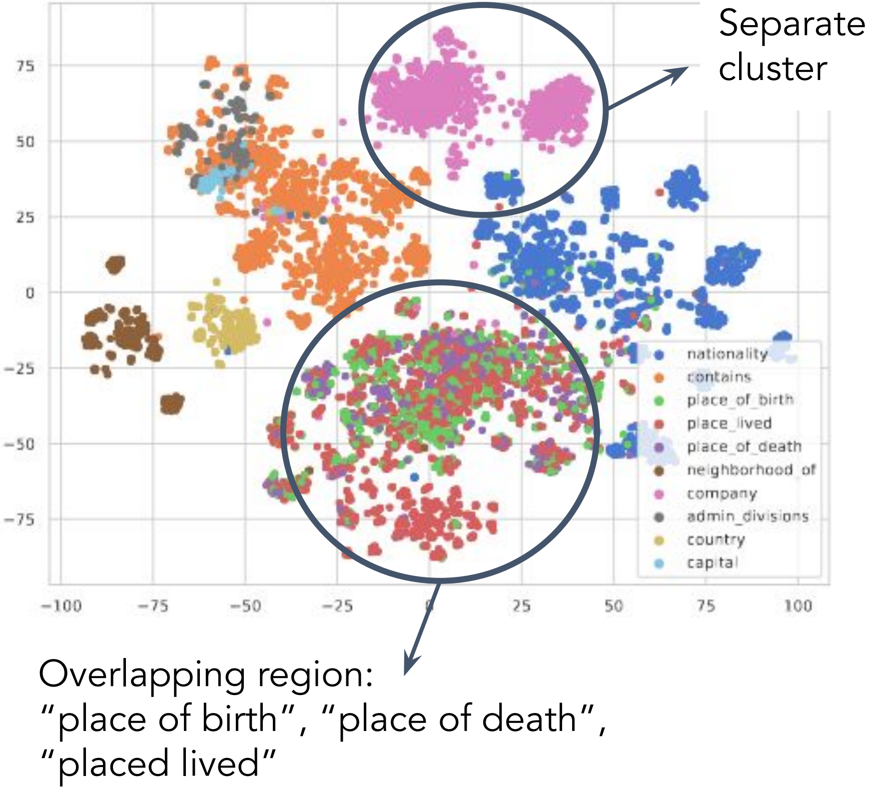
Prior Space

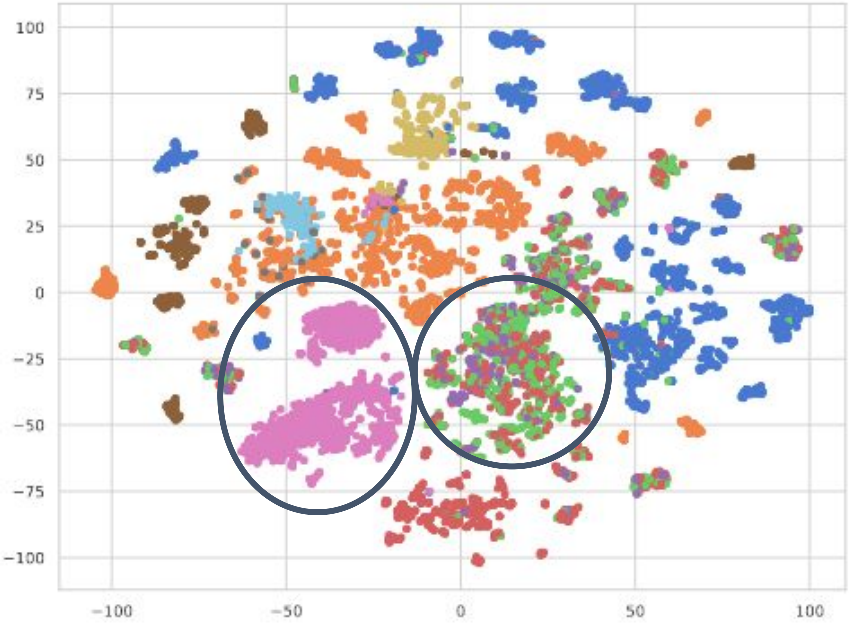


Separate cluster

Overlapping region:
"place of birth", "place of death",
"placed lived"

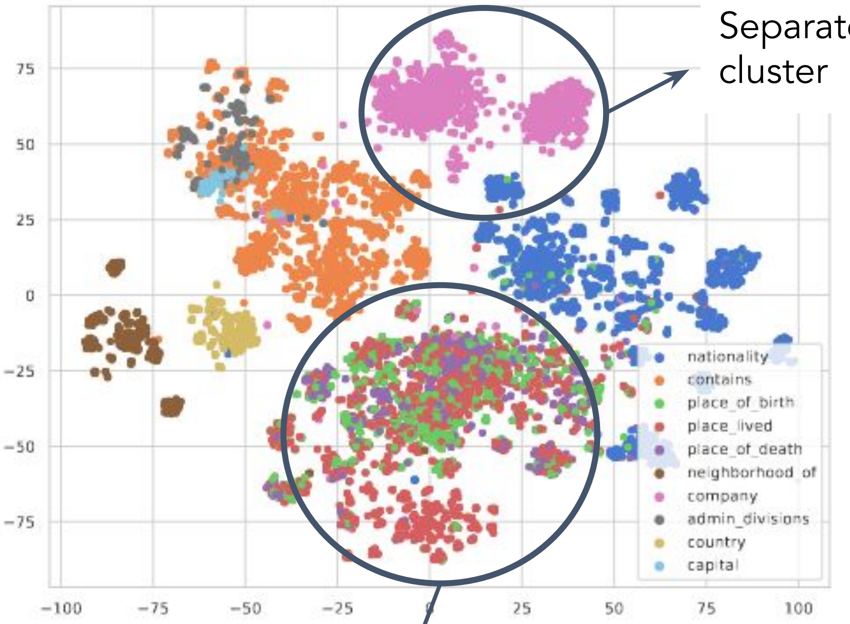
Analysis: Latent Space (NYT-10)





Analysis: Latent Space (NYT-10)

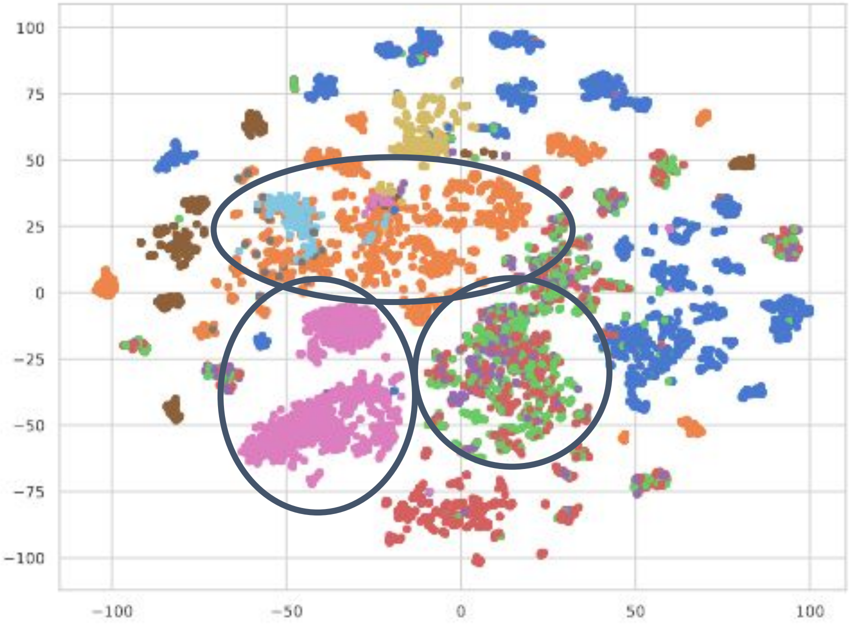
Prior Space



Separate cluster

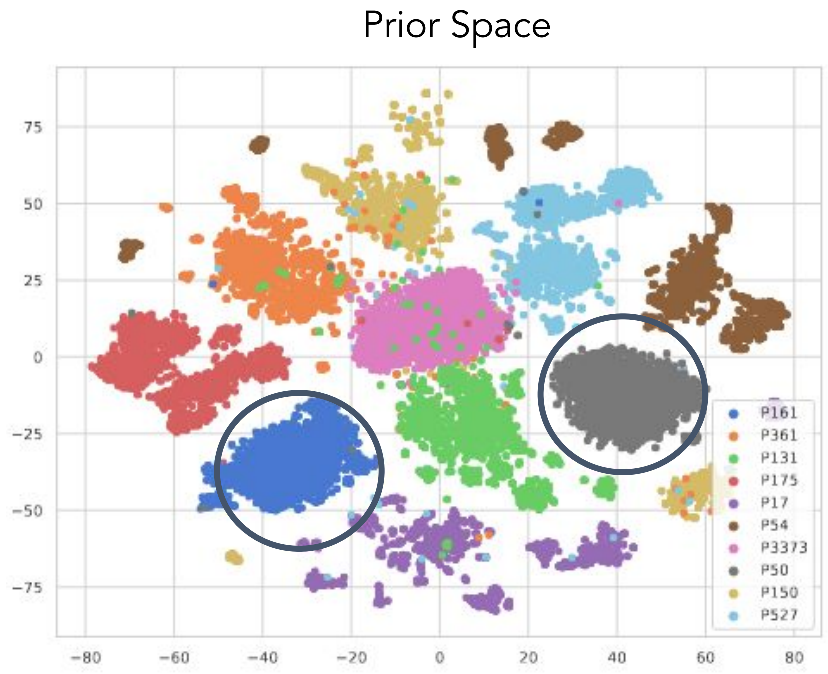
Overlapping region:
"place of birth", "place of death",
"placed lived"

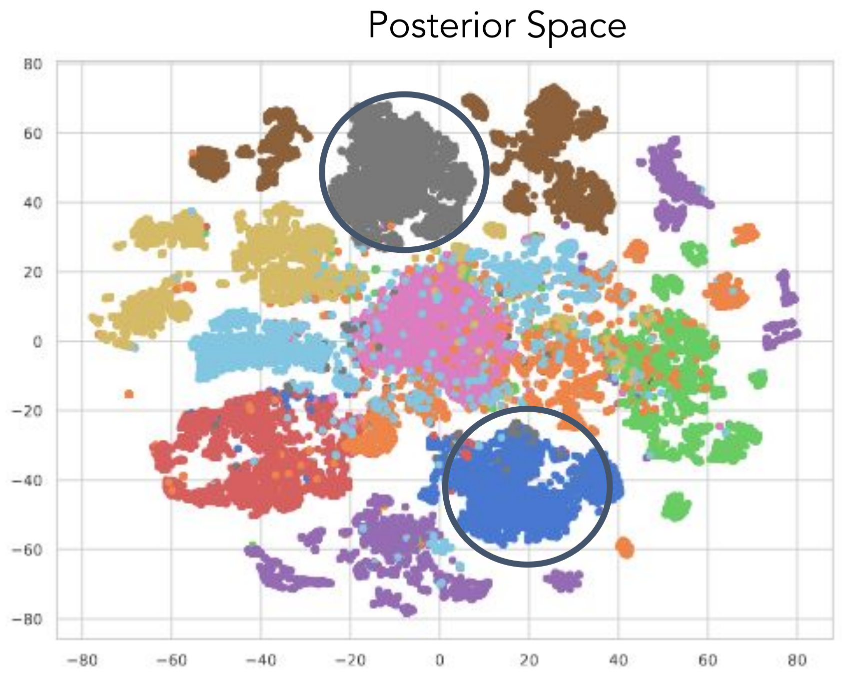
Posterior Space



- Cluster separation not so large
- Overall, very similar spaces

Analysis: Latent Space (WikiDistant)





- Similar results for WikiDistant
- “Part of” (orange), “has part” (cyan) sometimes not well separated

Conclusions

- We presented a multi-task, probabilistic approach to bring close sentences containing similar KB pairs in DSRE

\+ Combination of bag reconstruction and bag classification is proved effective
  - +4% boost in performance over the baseline when using Normal distribution priors
  - +8% boost in performance over the baseline when using KB priors

\+ The sentence latent space becomes very similar to the space of the priors

\+ Encoder-Decoder agnostic

\+ No requirement for test pair KB representations

\+ Improvement over a large pre-trained Language Model

Future Work

- Combine this method with pre-trained language models/noise reduction methods
- Investigate other ways to create priors via other Link Prediction methods

THANK YOU !

CODE

efstathia.christopoulou@manchester.ac.uk

https://fenchri.github.io

https://twitter.com/fenchri

References

- Christoph Alt, Marc Hübner, and Leonhard Hennig. Fine-tuning pre-trained transformer language models to distantly supervised relation extraction. In ACL 2019.
- Iz Beltagy, Kyle Lo, and Waleed Ammar. Combining distant and direct supervision for neural relation extraction. In NAACL 2019.
- Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. Translating embeddings for modeling multi-relational data. In NeurIPs 2013.
- Samuel R. Bowman, Luke Vilnis, Oriol Vinyals, An-drew Dai, Rafal Jozefowicz, and Samy Bengio. Generating sentences from a continuous space. In SIGNLL 2016.
- Xu Han, Tianyu Gao, Yankai Lin, Hao Peng, Yao-liang Yang, Chaojun Xiao, Zhiyuan Liu, Peng Li, Jie Zhou, and Maosong Sun. More data, morerelations, more context and more openness: A re-view and outlook for relation extraction. In AACL 2020.
- Raphael Hoffmann, Congle Zhang, Xiao Ling, Luke Zettlemoyer, and Daniel S. Weld. Knowledge-based weak supervision for information extraction of overlapping relations. In ACL 2011.
- Diederik P Kingma and Max Welling. Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114. 2013.
- Xu Han, Zhiyuan Liu, and Maosong Sun. Neural knowledge acquisition via mutual attention between knowledge graph and text. In AAAI 2018.
- Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory.Neural Computation,9(8):1735–1780, 1997.
- Guoliang Ji, Kang Liu, Shizhu He, and Jun Zhao. Distant supervision for relation extraction with sentence-level attention and entity descriptions. In AAAI 2017..
- Heng Ji and Ralph Grishman. Knowledge base population: Successful approaches and challenges. In ACL 2011..
- Yankai Lin, Shiqi Shen, Zhiyuan Liu, Huanbo Luan,and Maosong Sun.. Neural relation extraction with selective attention over instances. In ACL 2016.
- Pengda Qin, Weiran Xu, and William Yang Wang. DSGAN: Generative adversarial training for distant supervision relation extraction. In ACL 2018..
- Sebastian Riedel, Limin Yao, and Andrew McCallum. Modeling relations and their mentions without labeled text. In Machine Learning and Knowledge Discovery in Databases, 2010..
- Heng She, Bin Wu, Bai Wang, and Renjun Chi. Distant supervision for relation extraction with hierarchical attention and entity descriptions. In IJCNN 2018..
- Shikhar Vashishth, Rishabh Joshi, Sai Suman Prayaga,Chiranjib Bhattacharyya, and Partha Talukdar. RESIDE: Improving distantly-supervised neural relation extraction using side information. In EMNLP 2018.
- Jason Weston, Antoine Bordes, Oksana Yakhnenko,and Nicolas Usunier.. Connecting language and knowledge bases with embedding models for relation extraction. In EMNLP 2013.
- Peng Xu and Denilson Barbosa. Connecting lan-guage and knowledge with heterogeneous representations for neural relation extraction. In NAACL 2019.
- Zhi-Xiu Ye and Zhen-Hua Ling. Distant supervi-sion relation extraction with intra-bag and inter-bagattentions. In NAACL 2019.
- Ningyu Zhang, Shumin Deng, Zhanlin Sun, Guany-ing Wang, Xi Chen, Wei Zhang, and Huajun Chen. Long-tail relation extraction via knowledgegraph embeddings and graph convolution networks.In NAACL 2019.