Resolving Event Coreference with Supervised Representation Learning and Clustering-Oriented Regularization

Kian Kenyon-Dean, Jackie Chi Kit Cheung, & Doina Precup

McGill University, School of Computer Science;
McGill Reasoning and Learning Lab (RL-Lab);
Montreal Institute for Learning Algorithms (MILA)

Primary Contact:
Kian Kenyon-Dean

kian.kenyon-dean@mail.mcgill.ca
https://kiankd.github.io
https://github.com/kiankd/events

The Event Coreference Problem

What are the "events" in these documents?

Which events "**corefer**"?

Consider the following excerpts from two "hypothetical" news articles:

- The president's speech shocked the audience. He announced several new controversial policies.
- The policies proposed by the president will not surprise those who followed his campaign.

The Event Coreference Problem

What are the "events" in these documents? ✓

Which events "**corefer**"?

Event detection - other work (e.g., Rospocher et al. 2016) :

- The president's **speech**$_{m1}$ **shocked**$_{m2}$ the audience. He **announced**$_{m3}$ several new controversial policies.
- The policies **proposed**$_{m4}$ by the president will not **surprise**$_{m5}$ those who **followed**$_{m6}$ his **campaign**$_{m7}$.

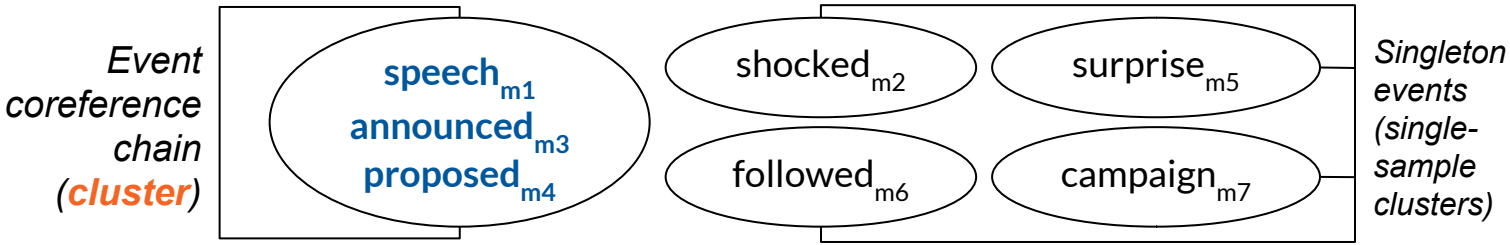
The Event Coreference Problem

~~**What** are the "events" in these documents?~~ ✓

Which events "**corefer**"?

Event Coreference Resolution:

- The president's **speech**$_{m1}$ **shocked**$_{m2}$ the audience. He **announced**$_{m3}$ several new controversial policies.
- The policies **proposed**$_{m4}$ by the president will not **surprise**$_{m5}$ those who **followed**$_{m6}$ his **campaign**$_{m7}$.

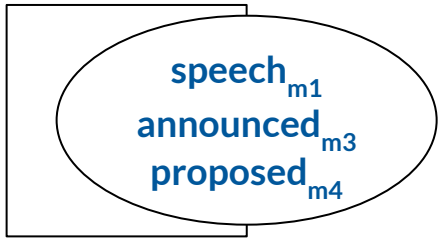


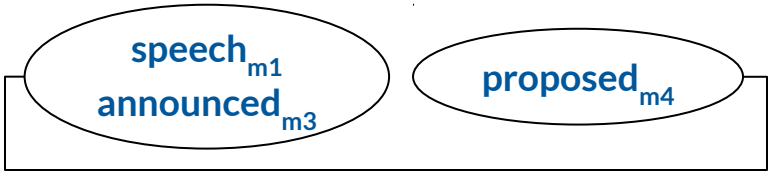
The Event Coreference Problem

Within- and cross-document coreference resolution:

- The president's **speech**$_{m1}$ **shocked**$_{m2}$ the audience. He **announced**$_{m3}$ several new controversial policies.
- The policies **proposed**$_{m4}$ by the president will not **surprise**$_{m5}$ those who **followed**$_{m6}$ his **campaign**$_{m7}$.

within- ***and*** **cross-document** resolution



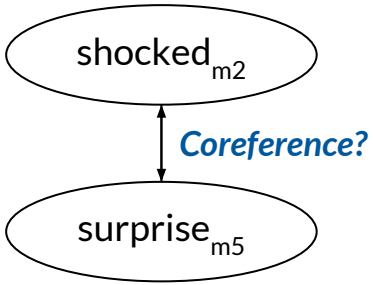


only **within-document** resolution

Summarizing the Event Coreference Problem

D1. The president's **speech**$_{m1}$ **shocked**$_{m2}$ the audience. He **announced**$_{m3}$ several new controversial policies.

D2. The policies **proposed**$_{m4}$ by the president will not **surprise**$_{m5}$ those who **followed**$_{m6}$ his **campaign**$_{m7}$.



- *Given:* **set of documents** with **event mentions** detected.
- *Predict:* which event mentions **both across and within documents** "corefer":
  - i.e., which event mentions are discussing **the real-world phenomenon** that occured at the same time and place.

Summarizing the Event Coreference Problem

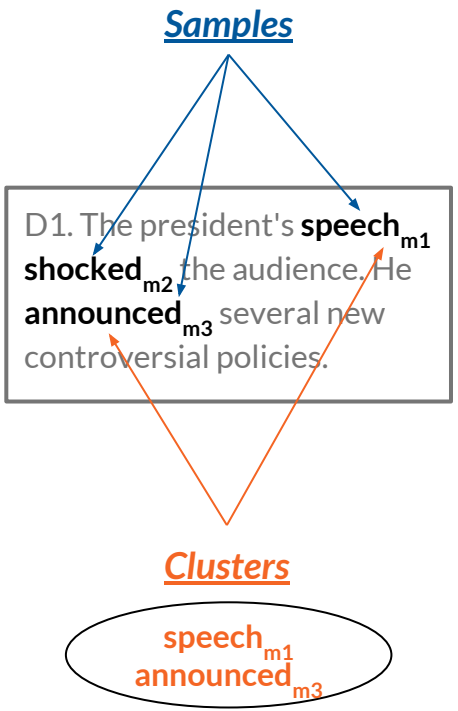
- *Given:* **set of documents** with **event mentions** detected.
- *Predict:* which event mentions **both across and within documents** "corefer":
  - i.e., which event mentions are discussing **the real-world phenomenon** that occured at the same time and place.
- *Note:* the events **will be completely different** depending on the document sets:
  - i.e., the events in the training set do not corefer with any from the validation and test sets.

Document set 1: Presidential press conferences, basketball games.

Totally distinct!

Document set 2: Celebrity rehab, New Orleans politics.

Abstracting the Event Coreference Problem



*A **non-parametric** **clustering** problem with **training data***

- **Non-parametric**: we do not know the number of clusters in the documents
- **Clustering**: we do not know the semantic identity of the clusters in advance
- **Training data**: samples and clusters are labelled in a training (and validation) set

Related Work

- **Unsupervised methods**, such as non-parametric **Bayesian clustering**; Bejan and Harabajiu 2010, 2014; Yang et al. 2015.
- **Graph partitioning**; Chen and Ji 2009.
- **Pairwise decision aggregation**; Bagga and Baldwin 1999; Chen et al. 2009; Cybulska and Vossen 2015.
- **Representation learning**; Krause et al. 2016; Choubey and Huang 2017.
- **Event templates**; Cybulska and Vossen 2015, 2016.

Uses extra pre-detected or pre-annotated info.

Event Coreference Dataset

ECB+ Statistics (full dataset, within + cross-document)

Topics	43
Documents	982
Event mentions (**samples**)	6833
Coreference chains (**clusters**)	1982
Average cluster size	3.45 samples

Dataset - **the Event Coreference Bank Plus, ECB+** (Cybulska and Vossen, 2014)

- To our knowledge, the only event coreference dataset with *both* **within-** and **cross-document** coreference labelled.
- Relatively small!

Solving the Problem

Straightforward (but not very fun) solutions:

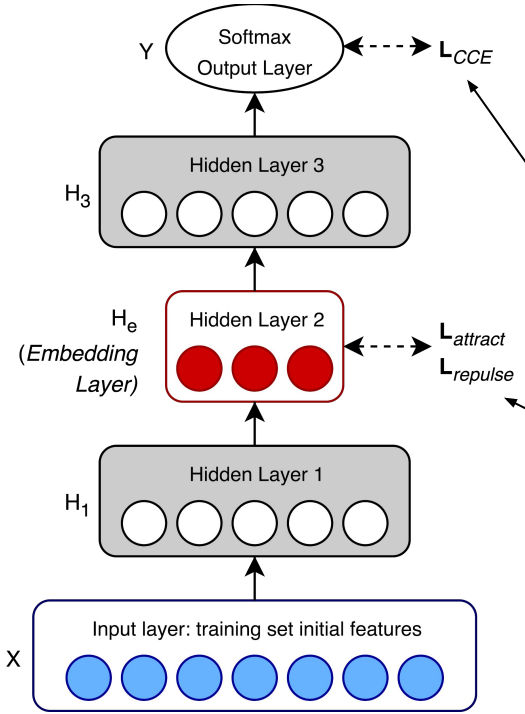
- *Lemma*: deterministic; return that coreference occurs between event mentions *if* they have **the same head lemma**.
- *Lemma-δ* (Upadhyay et al. 2016): same as above, except only return that coreference occurs *if* the event mentions come from documents with TF-IDF similarity > δ.
- *Unsupervised*: **extract features** from event mentions, then tune a **clustering algorithm** on the training/validation sets.

Solving the Problem

Fundamental questions:

- How can we **take advantage of the training data** - the labelled event coreference chains - in a way that **generalizes to novel documents and events?**
- Moreover, since we have to perform clustering to build the final event coreference chains, **how can we optimize training for the clustering task**?

Solving the Problem - Overview

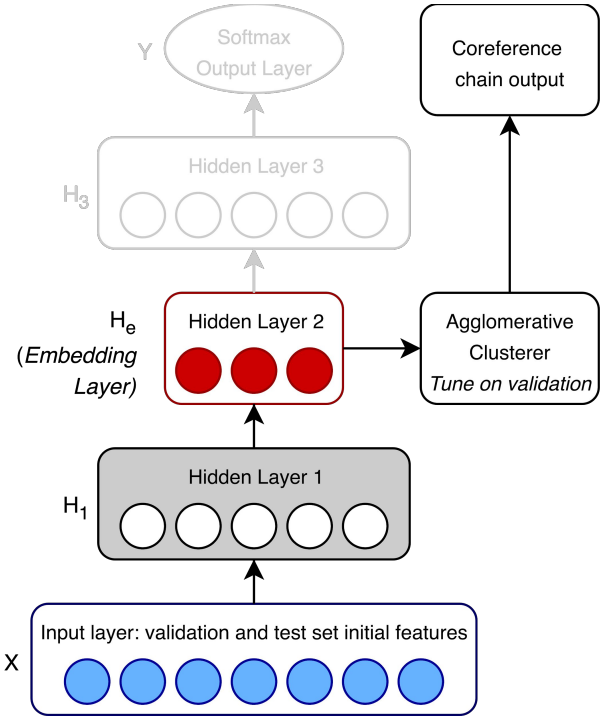


Supervised representation learning: train **neural network** to learn **bottleneck features** to represent event mentions;

- Define a **classification** task **on the training set**;
- Use **clustering-oriented regularization** in training objective to make **clusterable representations**;

Solving the Problem - Overview

- Apply **non-parameteric agglomerative clustering** algorithm on validation representations and tune;
- Predict final test set coreference chains with **trained model** and **tuned clustering algorithm.**



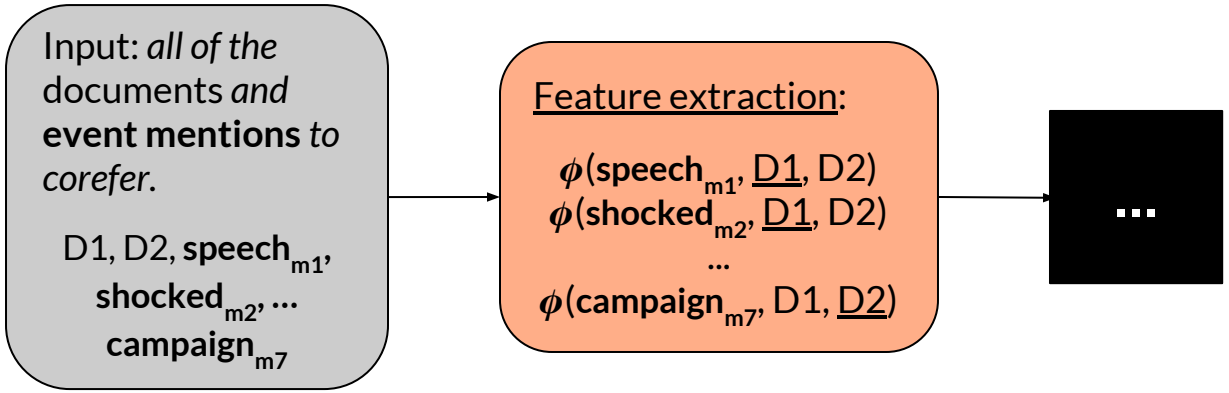
Solving the Problem - Overview

- **Supervised representation learning**: train **neural network** to learn **bottleneck features** to represent event mentions;
  - Define a **classification** task **on the training set**;
  - Use **clustering-oriented regularization** in training objective to make **clusterable representations**;
- Apply **non-parameteric agglomerative clustering** algorithm on validation representations and tune;
- Predict final test set coreference chains with **trained model** and **tuned clustering algorithm.**

Solving the Problem - Step 1

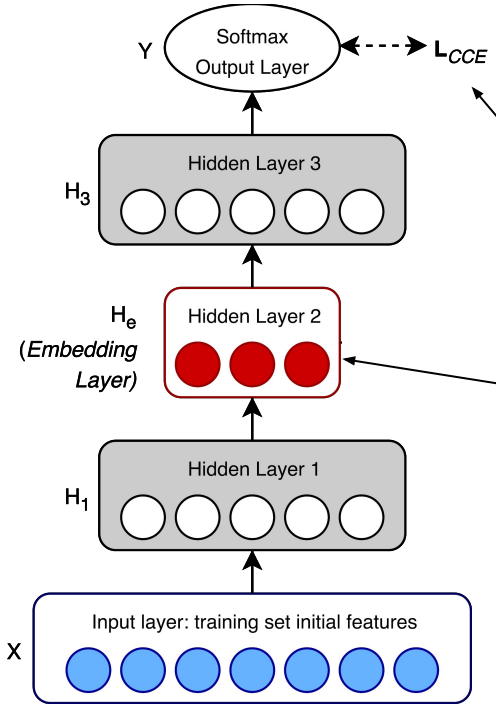
D1: The president's **speech**$_{m1}$ **shocked**$_{m2}$ the audience. He **announced**$_{m3}$ several new controversial policies.

D2: The policies **proposed**$_{m4}$ by the president will not **surprise**$_{m5}$ those who **followed**$_{m6}$ his **campaign**$_{m7}$.



Extract features to represent: **context** (averaged word embeddings), the **document** (TF-IDF vector), **comparisons** (features *relating* the event mention to the others in the documents).

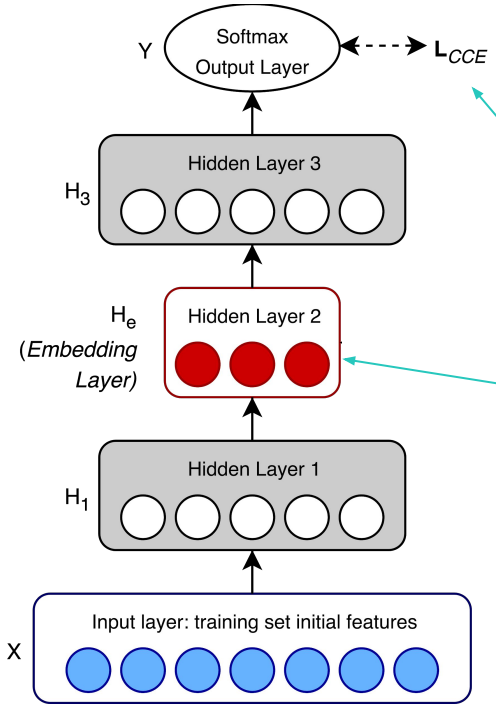
Solving the Problem - Step 2



Supervised Representation Learning

- Train on the **training set** data like a classification problem:
  - Objective: **predict the cluster a sample belongs to**, out of C total possible clusters (*categorical cross-entropy*).
  - Select an **embedding layer** as the event **embedding**.
  - For singleton events, assign them to a final class C+1.

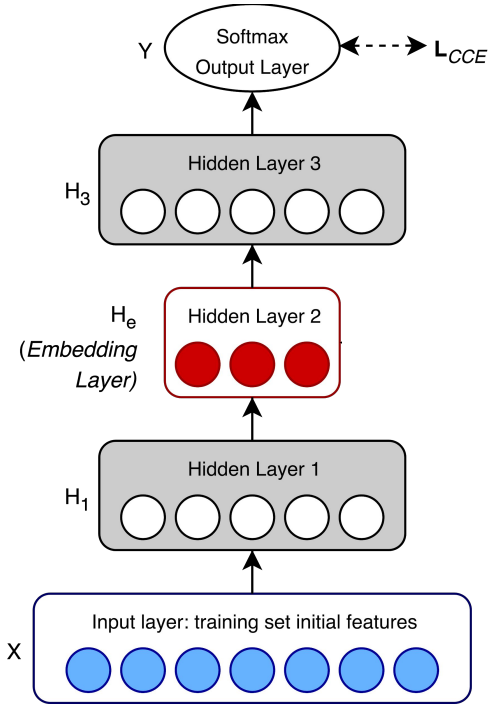
Solving the Problem - Step 2



Supervised Representation Learning

- Train on the **training set** data like a classification problem:
  - Objective: **predict the cluster a sample belongs to**, out of C total possible clusters (*categorical cross-entropy*).
  - Select an **embedding layer** as the event **embedding**.
  - For singleton events, assign them to a final class C+1.
- **Hypothesis**: *the model will learn to represent the **general distributional relationships** between **samples** and **clusters** in the bottleneck embedding layer and extract those relevant features.*

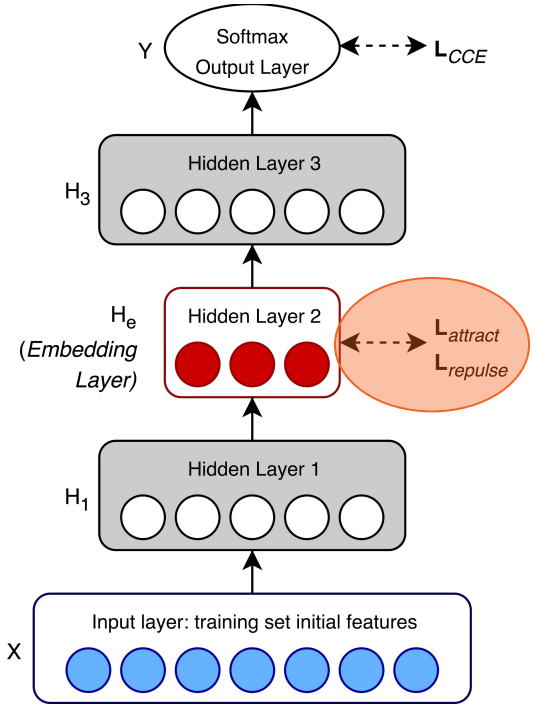
Solving the Problem - Step 2



Supervised Representation Learning

- *Wait!!*
- *How do we know that the representations will actually be useful?*
- *More precisely, what will guarantee that they will be **useful for a clustering algorithm**?*

Solving the Problem - Step 2.5

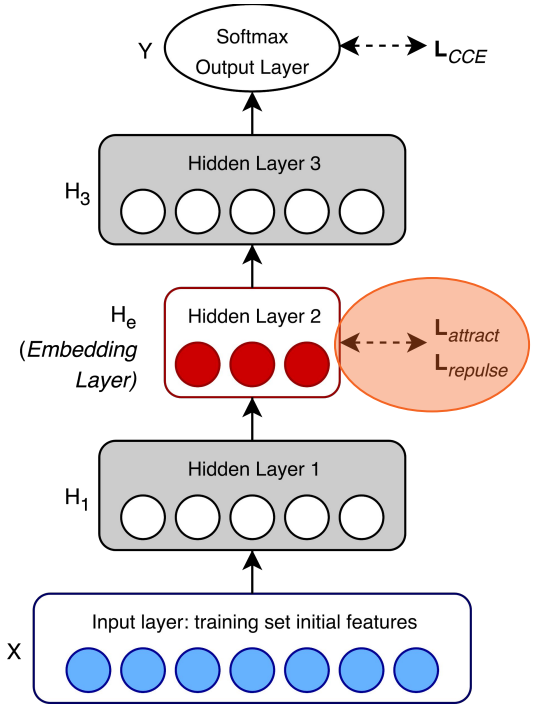


Supervised Representation Learning

with Clustering-Oriented Regularization (CORE)

- The **representation space** will be easily clusterable if:
  - *Samples belonging to the same cluster are **similar**; and,*
  - *Samples belonging to different clusters are **dissimilar**.*
- Categorical cross-entropy does not induce this geometric quality into the latent space.

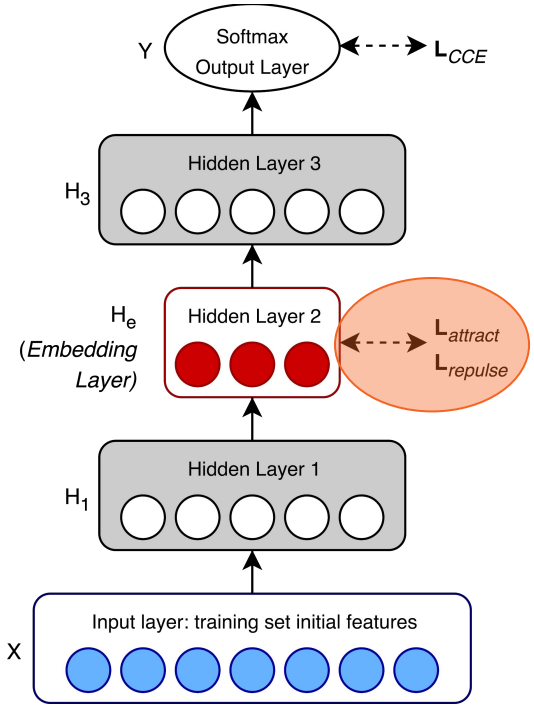
Solving the Problem - Step 2.5



Supervised Representation Learning
with Clustering-Oriented Regularization (CORE)

- Induce **clusterability** with two terms:
  - *Attractive loss*: **minimize** the **distance** between sample **embeddings** belonging to the **same cluster**.
  - *Repulsive loss*: **maximize** the **distance** between sample **embeddings** belonging to **different clusters**.

Solving the Problem - Step 2.5



Supervised Representation Learning

with Clustering-Oriented Regularization (CORE)

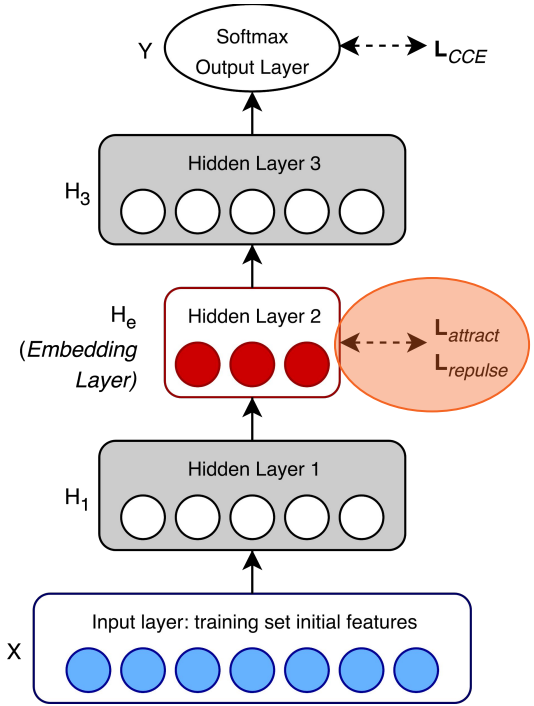
- Induce **clusterability** with two terms:

$$\mathbf{L}_{attract} = \frac{1}{|\mathcal{S}|} \sum_{(a,b)\in\mathcal{S}} \mathbf{d}(\vec{e_a}, \vec{e_b})$$

$$\mathbf{L}_{repulse} = 1 - \frac{1}{|\mathcal{D}|} \sum_{(c,d)\in\mathcal{D}} \mathbf{d}(\vec{e_c}, \vec{e_d})$$

$$\mathbf{L} = \mathbf{L}_{CCE} + \lambda_1 \mathbf{L}_{attract} + \lambda_2 \mathbf{L}_{repulse}$$

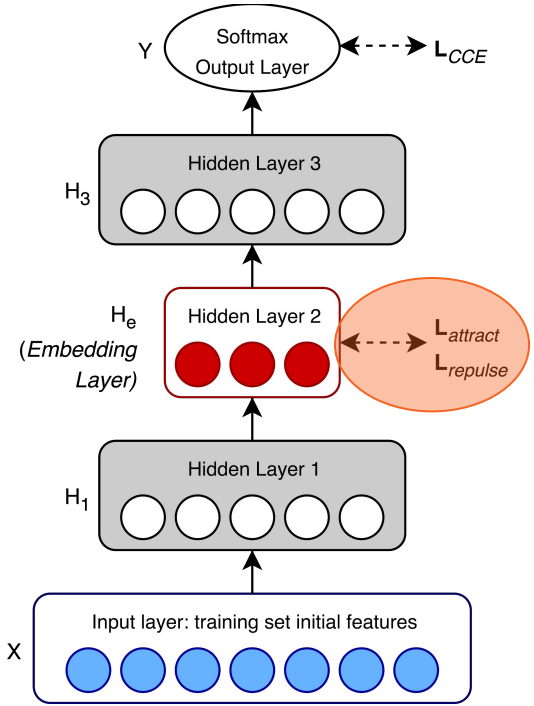
Solving the Problem - Step 2.5



Supervised Representation Learning
with Clustering-Oriented Regularization (CORE)

- Measure distance with **cosine distance** so that embedding norms don't explode.
- Apply the loss over all sample **pairs in the training mini-batch**:
  - *Not actually too expensive* - most expensive operation is the input matrix multiplied by its transpose; no noticable descrease in training time.

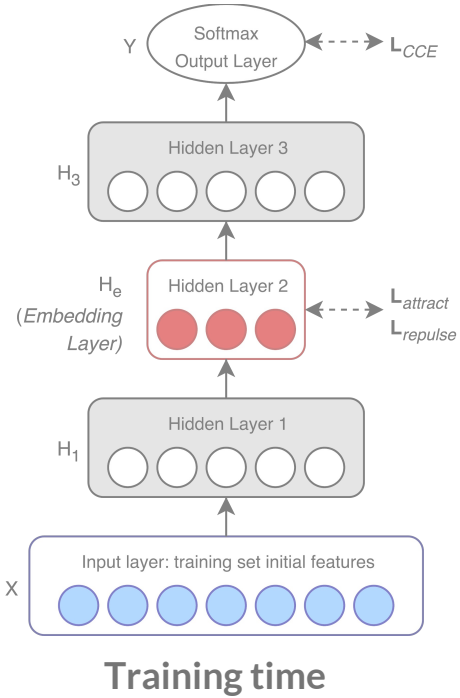
Solving the Problem - Step 2.5

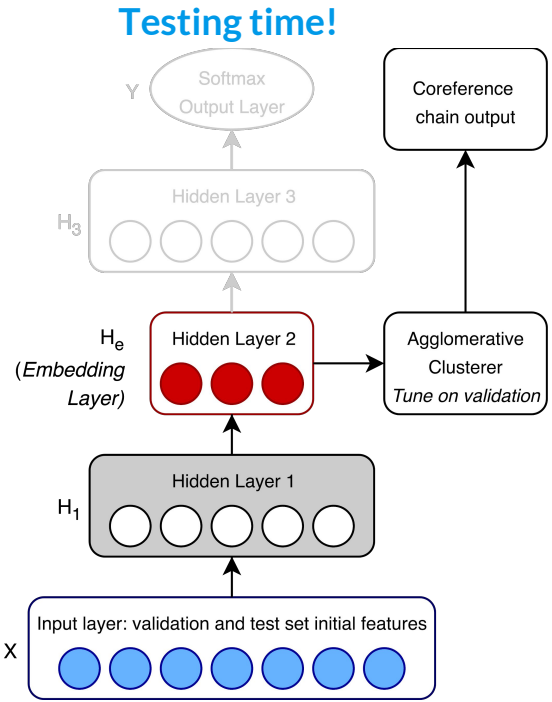


Supervised Representation Learning
with Clustering-Oriented Regularization (CORE)

- **Motivation**: imposing this geometric quality will induce the network to learn a *naturally clusterable latent space* that is easy for the agglomerative clustering algorithm.
- This will property will naturally hold on the training set **but does it generalize? Yes!**

Solving the Problem - Step 3





After training the model, use it to **build test embeddings**.

Tune an **agglomerative clustering algorithm** (with parameter **τ**, allowing for non-parametric clustering) on validation set.

After tuning, predict the final coreference chains!

Results - Combined within+cross

Model	MUC	B3	CEAF-M	CEAF-E	BLANC	CoNLL
Cybulska & Vossen [2]	55	**69**	**58**	66	63	64
Lemma	62	62	51	54	63	61
Lemma-***δ***	61	**69**	**59**	66	**67**	66
Unsupervised	48	66	51	58	58	57
CCE	65	64	50	61	59	63
CCE+CORE	66	68	56	**68**	62	67
CCE+CORE+Lemma	**69**	**69**	**58**	**69**	64	**69**

Related work

Baselines

Our models

Results - Combined within+cross

Supervised representation learning works better than just using the original features!

Model	MUC	B3	CEAF-M	CEAF-E	BLANC	CoNLL	
Cybulska & Vossen [2]	55	**69**	**58**	66	63	64	Related work
Lemma	62	62	51	54	63	61	Baselines
Lemma-δ	61	**69**	**59**	66	**67**	66	
Unsupervised	48	66	51	58	58	57	
CCE	65	64	50	61	59	63	**Our models**
CCE+CORE	66	68	56	**68**	62	67	
CCE+CORE+Lemma	**69**	**69**	**58**	**69**	64	**69**	

Results - Combined within+cross

Clustering-oriented regularization generalizes! It substantially improves the clusterability of the latent space, even in the test set!

Model	MUC	B3	CEAF-M	CEAF-E	BLANC	CoNLL	
Cybulska & Vossen [2]	55	**69**	**58**	66	63	64	Related work
Lemma	62	62	51	54	63	61	Baselines
Lemma-δ	61	**69**	**59**	66	**67**	66	
Unsupervised	48	66	51	58	58	57	
CCE	65	64	50	61	59	63	**Our models**
CCE+CORE	66	68	56	**68**	62	67	
CCE+CORE+Lemma	**69**	**69**	**58**	**69**	64	**69**	

Results - Combined within+cross

Model	MUC	B3	CEAF-M	CEAF-E	BLANC	CoNLL
Cybulska & Vossen [2]	55	**69**	**58**	66	63	64
Lemma	62	62	51	54	63	61
Lemma-***δ***	61	**69**	**59**	66	**67**	66
Unsupervised	48	66	51	58	58	57
CCE	65	64	50	61	59	63
CCE+CORE	66	68	56	**68**	62	67
CCE+CORE+Lemma	**69**	**69**	**58**	**69**	64	**69**

Related work

Baselines

Our models

Results - only within-document

Model	MUC R	MUC P	MUC F	B^3 R	B^3 P	B^3 F	CM F	CE R	CE P	CE F	BLANC R	BLANC P	BLANC F	CoNLL F
Baselines														
LEMMA-δ	41	77	53	86	97	**92**	85	92	82	87	65	86	71	77
UNSUPERVISED	32	36	34	85	86	85	74	80	78	79	65	55	57	66
Model Variants														
CCE	44	49	46	87	89	88	79	82	80	81	67	67	67	72
CORE	55	32	40	89	70	78	65	64	79	71	75	54	56	63
CORE+CCE	43	68	53	87	95	91	84	90	82	86	67	76	70	76
CORE+CCE+LEMMA	57	69	**63**	90	94	**92**	**86**	90	86	**88**	73	78	**75**	**81**

Conclusions & Future Work

- **Supervised representation learning** offers **better representations** of samples than the original features!
- **Clustering-oriented regularization (CORE)** subsantially **improves the quality of embeddings** and generalizes to improving the clusterability of the latent space overall!
- *Can we make an end-to-end system for event coreference, without feature extraction?*
- *Can CORE be used for other tasks, where a clusterable latent space would be useful?*

References

- Agata Cybulska and Piek Vossen. 2014. *Using a sledgehammer to crack a nut? Lexical diversity and event coreference resolution.* In LREC, pages 4545– 4552.
- Agata Cybulska and Piek Vossen. 2015. *Translating Granularity of Event Slots into Features for Event Coreference Resolution.* 3rd Workshop on EVENTS at the NAACL-HLT, pages 1– 10.
- Shyam Upadhyay, Nitish Gupta, Christos Christodoulopoulos, and Dan Roth. 2016. *Re- visiting the Evaluation for Cross Document Event Coreference.* Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, pages 1949–1958, Osaka, Japan, December 11-17 2016.
- Rospocher, Marco, Marieke van Erp, Piek Vossen, Antske Fokkens, Itziar Aldabe, German Rigau, Aitor Soroa, Thomas Ploeger, and Tessel Bogaard. *Building event-centric knowledge graphs from news.* Web Semantics: Science, Services and Agents on the World Wide Web 37 (2016): 132-151.
- Vossen, Piek, and Agata Cybulska. "Identity and granularity of events in text." International Conference on Intelligent Text Processing and Computational Linguistics. Springer, Cham, 2016.
- Bejan, Cosmin Adrian, and Sanda Harabagiu. Unsupervised event coreference resolution with rich linguistic features. Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, 2010.
- Bejan, Cosmin Adrian, and Sanda Harabagiu. Unsupervised event coreference resolution. Computational Linguistics 40.2 (2014): 311-347.
- Yang, Bishan, Claire Cardie, and Peter Frazier. A Hierarchical Distance-dependent Bayesian Model for Event Coreference Resolution. Transactions of the Association for Computational Linguistics 3 (2015): 517-528.

References

- Amit Bagga and Breck Baldwin. 1999. *Cross- document event coreference: Annotations, exper- iments, and observations.* In Proceedings of the Workshop on Coreference and its Applications, pages 1–8. ACL.
- Zheng Chen, Heng Ji, and Robert Haralick. 2009. *A Pairwise Event Coreference Model, Feature Impact and Evaluation for Event Coreference Resolution.* In Proceedings of the workshop on events in emerg- ing text types, pages 17–22. ACL.
- Choubey, Prafulla Kumar, and Ruihong Huang. *Event Coreference Resolution by Iteratively Unfolding Inter-dependencies among Events.* EMNLP 2017.
- Sebastian Krause, Feiyu Xu, Hans Uszkoreit, and Dirk Weissenborn. 2016. *Event linking with sentential features from convolutional neural networks.* In Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning, pages 239–249.
- Zheng Chen and Heng Ji. 2009. Graph-based event coreference resolution. In Proceedings of the 2009 Workshop on Graph-based Methods for NLP, pages 54–57. ACL.

Thank you!

Questions? Comments?

kian.kenyon-dean@mail.mcgill.ca

Code repository and results files:
https://github.com/kiankd/events