VERNA DANKERS

MULTITASK LEARNING

ATCS Lecture, April 20 2020

Contents

Introduction Motivation

- IMPROVE MAIN TASK THROUGH AUXILIARY TASKS

  E.g. Improve dependency parsing through POS labelling.

- MOVE TOWARDS A UNIFIED NLP ARCHITECTURE

  E.g. Frame any NLP task as question answering task
  - DecaNLP model of McCann et al. (2018).

Introduction Motivation

- IMPROVE MAIN TASK THROUGH AUXILIARY TASKS

E.g. Improve dependency parsing through POS labelling.

- MOVE TOWARDS A UNIFIED NLP ARCHITECTURE

E.g. Frame any NLP task as question answering task - DecaNLP model of McCann et al. (2018).

Examples

Question	**Context**	**Answer**
What is a major importance of Southern California in relation to California and the US?	...Southern California is a major economic center for the state of California and the US....	major economic center
What is the translation from English to German?	Most of the planet is ocean water.	Der Großteil der Erde ist Meerwasser
What is the summary?	Harry Potter star Daniel Radcliffe gains access to a reported £320 million fortune...	Harry Potter star Daniel Radcliffe gets £320M fortune...
Hypothesis: Product and geography are what make cream skimming work. Entailment, neutral, or contradiction?	Premise: Conceptually cream skimming has two basic dimensions – product and geography.	Entailment
Is this sentence positive or negative?	A stirring, funny and finally transporting re-imagining of Beauty and the Beast and 1930s horror film.	positive

Introduction Inductive Biases

How can MTL improve performance on the main task (Caruana, 1993)?

1 DATA AMPLIFICATION

Introducing an auxiliary task means adding data and introducing regularisation.

2 REPRESENTATION BIAS

Introducing an auxiliary task may lead to finding different local minima, i.e. lead to finding different representations in the hypothesis space.

3 ATTRIBUTE SELECTION

Introducing the auxiliary task can help the main task focus on the most relevant input features.

4 EAVESDROPPING

Features useful for both tasks may be easier to learn on the auxiliary task.

Introduction Inductive Biases

1 DATA AMPLIFICATION & REPRESENTATION BIAS

E.g. language modelling and autoencoding (Rei, 2017).

2 ATTRIBUTE SELECTION

E.g. use gaze prediction (auxiliary task) to allow other NLP tasks to focus on relevant input words (Barrett et al., 2018).

3 EAVESDROPPING

E.g. Cheng et al. (2015) perform name error detection (main task)
and include sentence-level name detection (auxiliary task).

Reference	my name is captain rodriguez
Hypothesis	my name is captain road radios

Approach

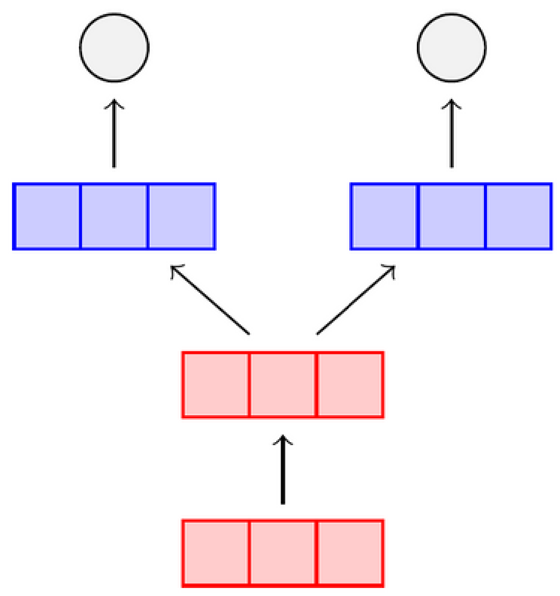
- NETWORK ARCHITECTURE

  Develop network based on the task hierarchy;
  Select hard or soft parameter sharing.

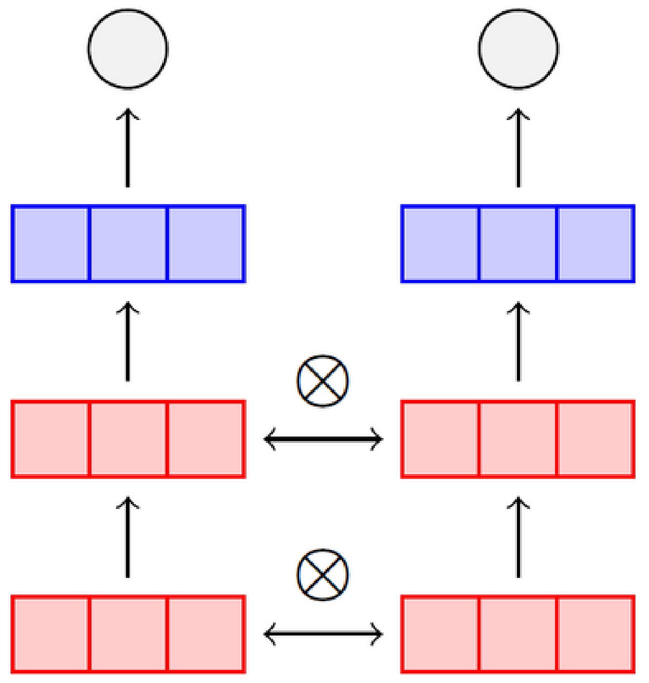
- TASK PRIORITISATION

  Prioritisation in parameter update frequencies;
  Prioritisation through task weighting.

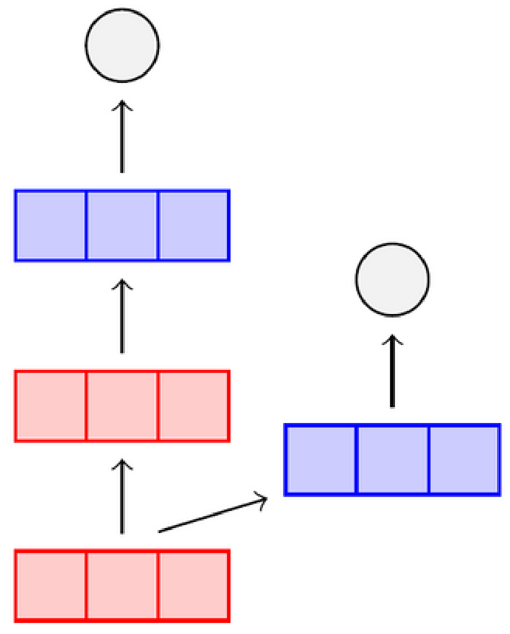
Approach Network Architecture



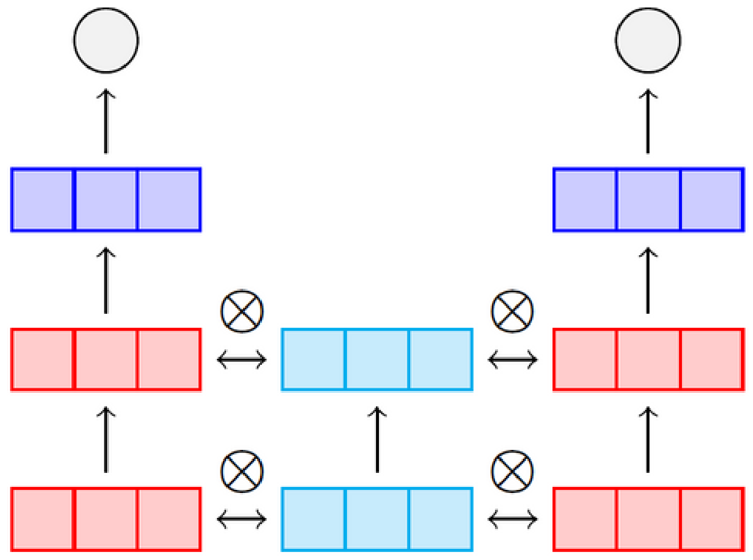
- HARD SHARING



- SOFT SHARING

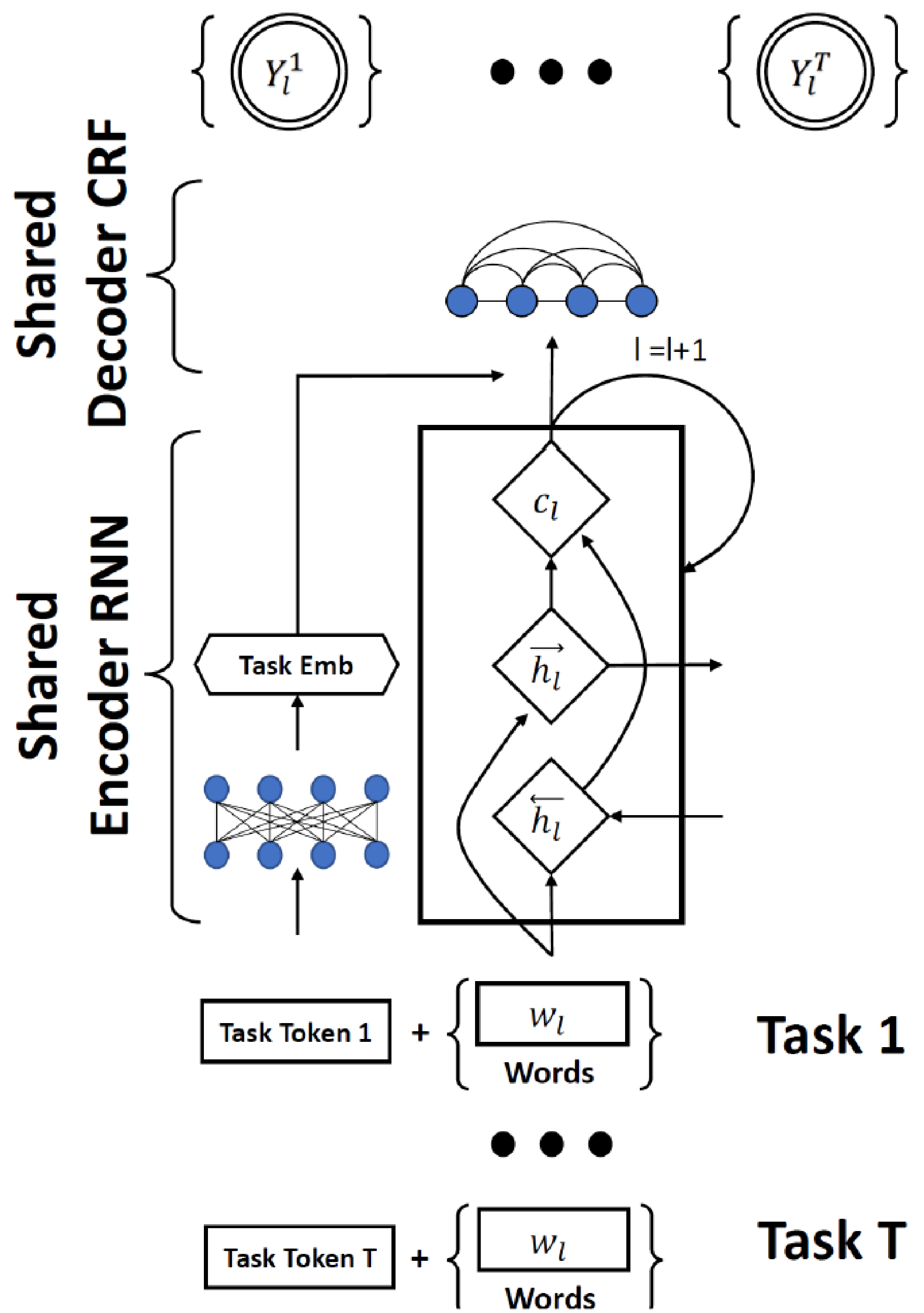


- HIERARCHICAL SHARING



- SOFT LAYER SHARING

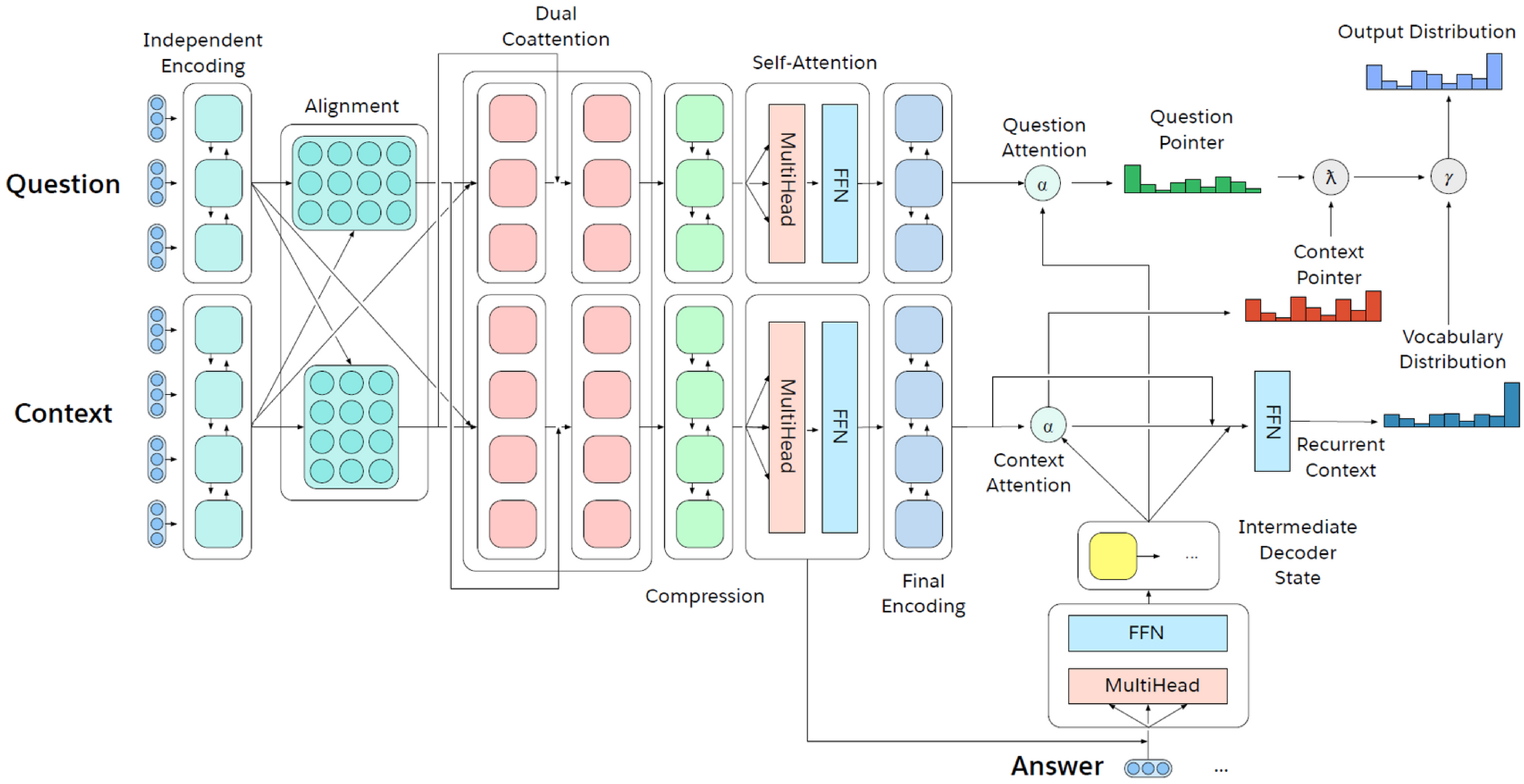
Approach Network Architecture



HARD SHARING

Changpinyo et al. (2018) share both encoder and decoder, but introduce task embeddings.

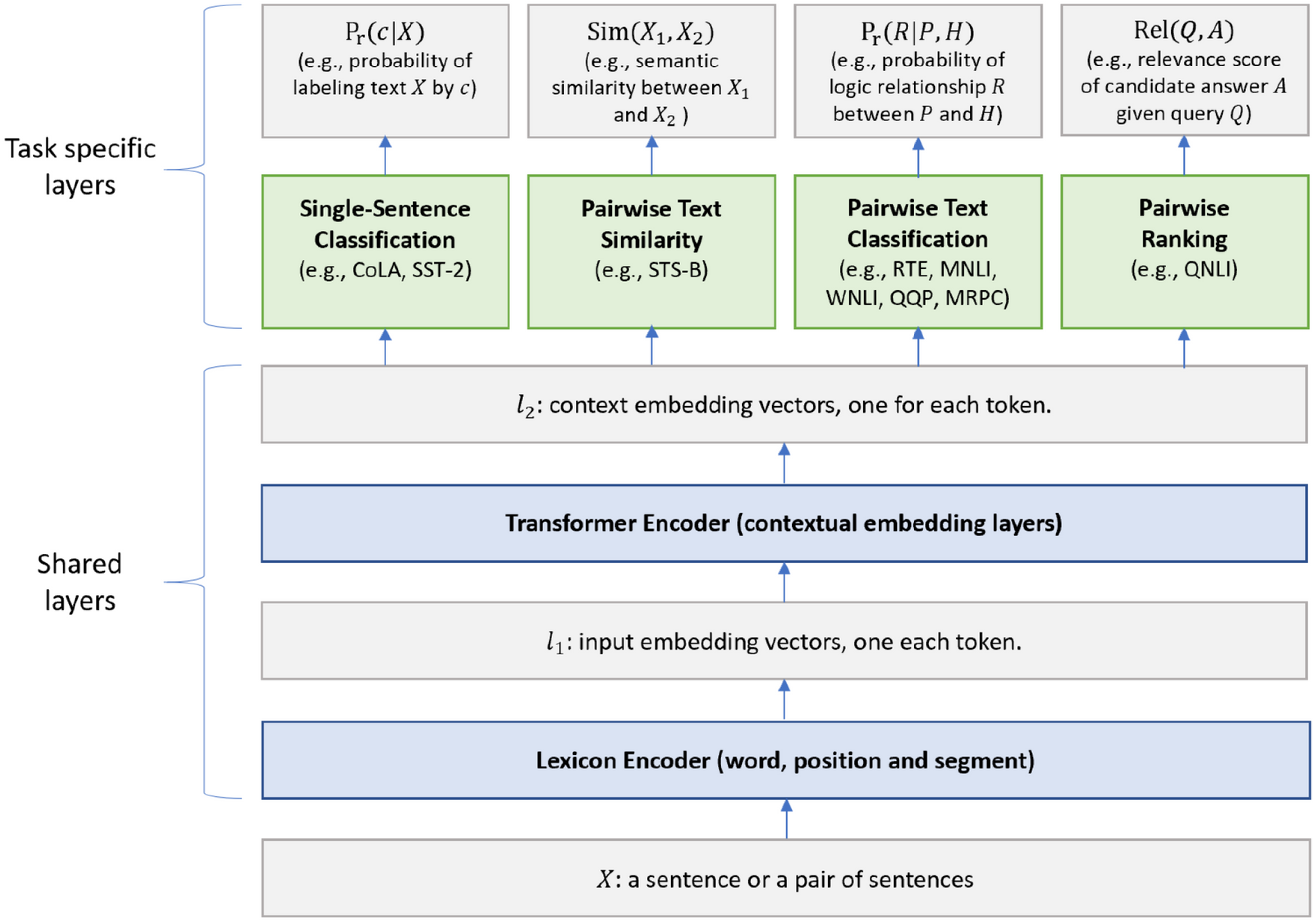
Approach Network Architecture



- HARD SHARING

DecaNLP (McCann et al., 2018)

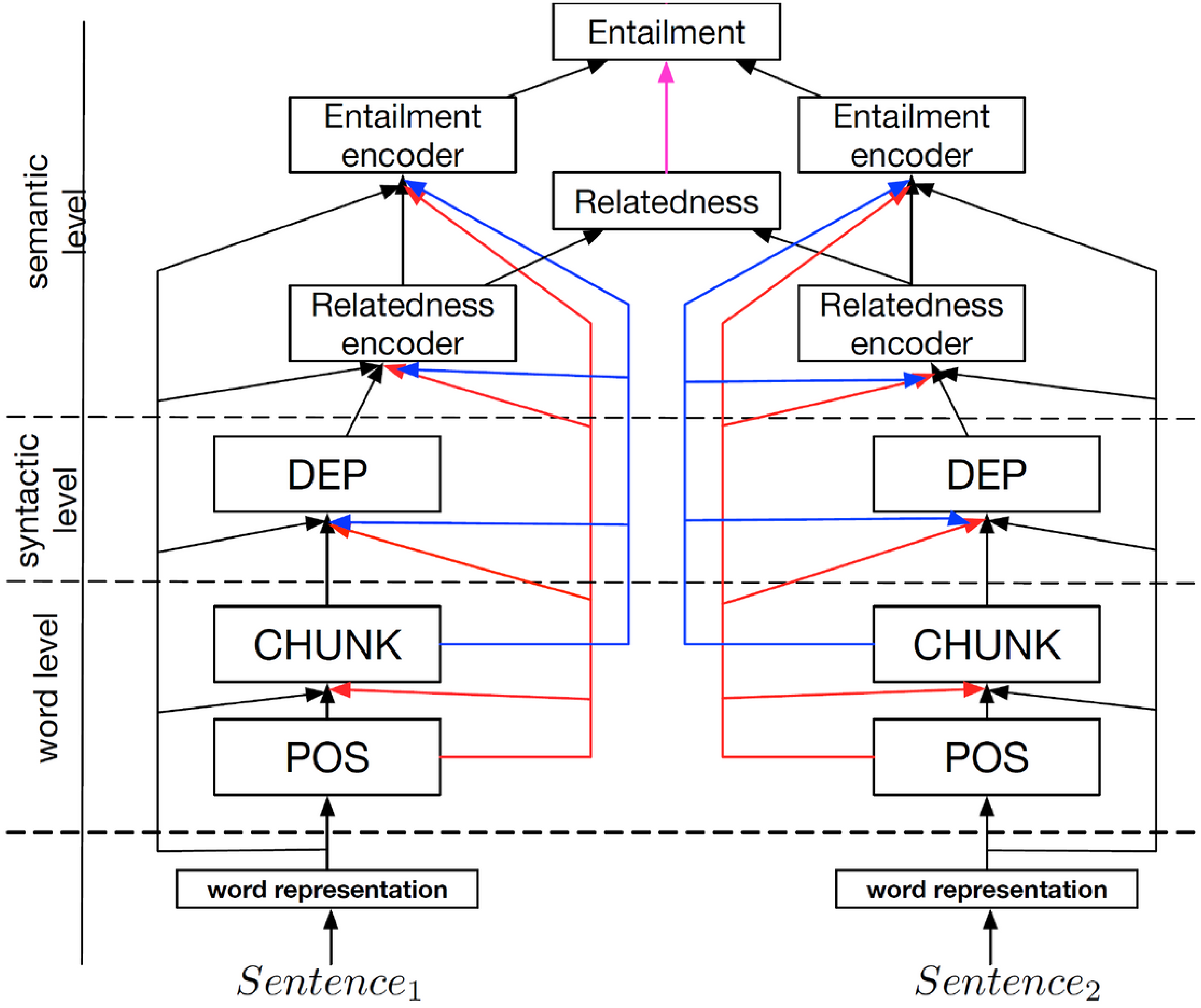
Approach Network Architecture



HARD SHARING

Liu et al. (2019) combine transfer learning with BERT and multitask learning to improve performance on GLUE.

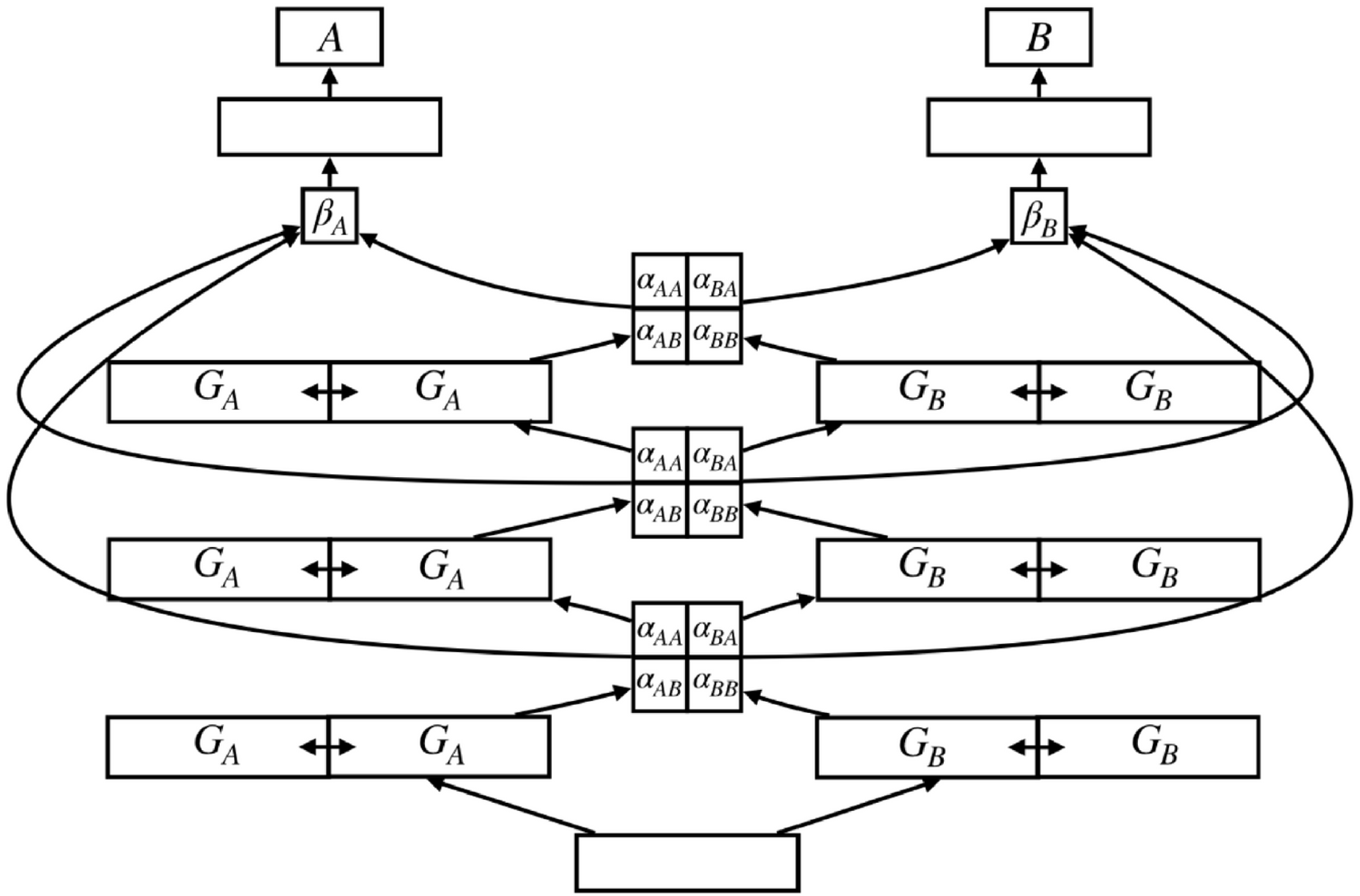
Approach Network Architecture



HIERARCHICAL SHARING

Joint-many model of Hashimoto et al. (2017).

Approach Network Architecture

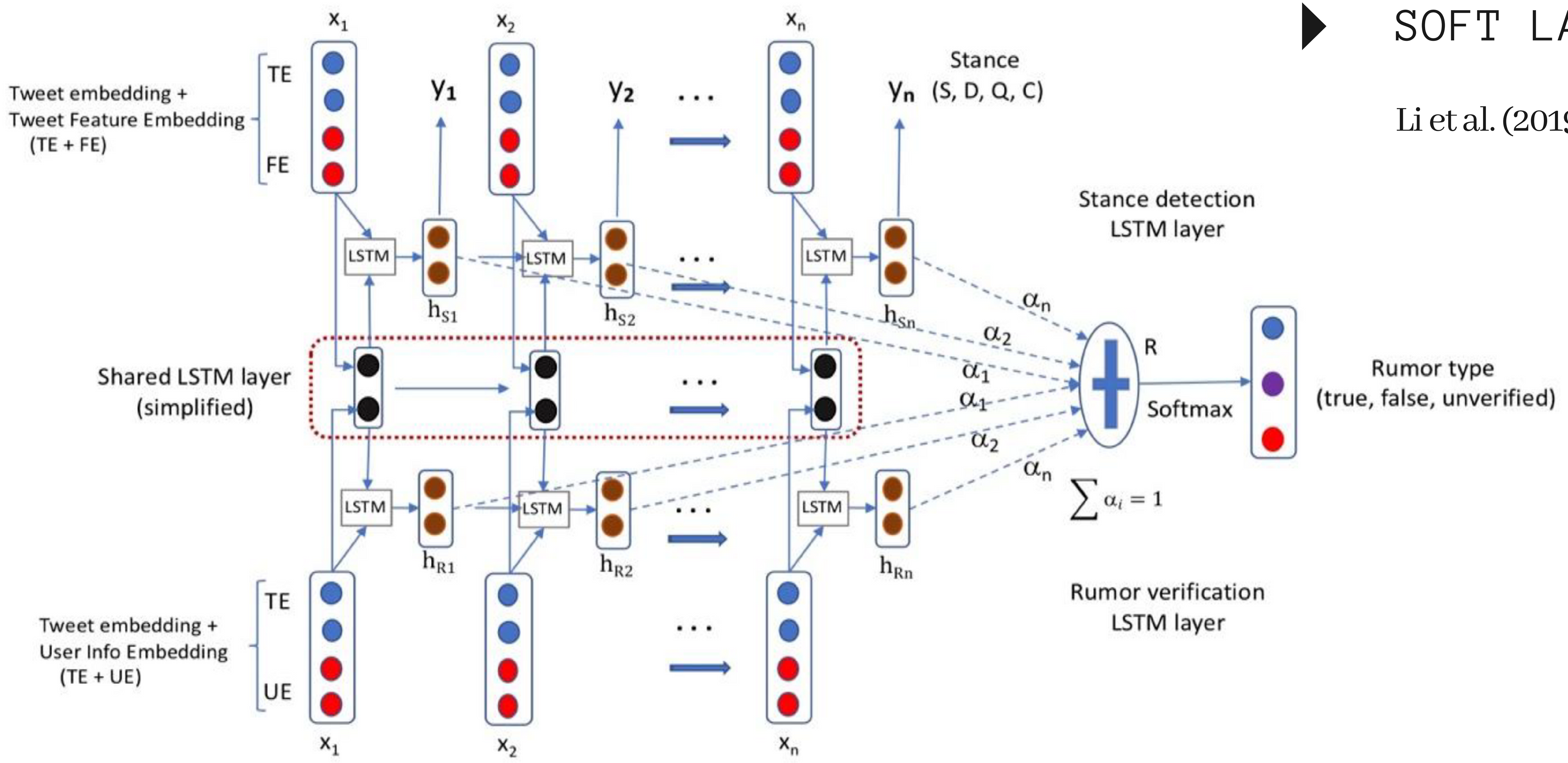


- SOFT SHARING

Sluice network of Ruder et al. (2019) uses cross-stitch units, skip connections and orthogonality constraints on subspaces of recurrent layers.

$$\begin{bmatrix} \tilde{h}_A \\ \tilde{h}_B \end{bmatrix} = \begin{bmatrix} \alpha_{AA} & \alpha_{AB} \\ \alpha_{BA} & \alpha_{BB} \end{bmatrix} \begin{bmatrix} h_A^\top & h_B^\top \end{bmatrix}$$

Approach Network Architecture



- SOFT LAYER SHARING

Li et al. (2019)

Approach Task Prioritisation

1 RANDOMISED TRAINING

(a) Uniform Task Selection (Søgaard and Goldberg, 2016).

(b) Proportional Task Selection (Sahn et al., 2018).

2 PERIODIC TASK ALTERNATIONS

Dong et al. (2015) use periodic task alternations with equal training ratios for every task.

Approach Task Prioritisation

1 RANDOMISED TRAINING

(a) Uniform Task Selection (Søgaard and Goldberg, 2016).
(b) Proportional Task Selection (Sahn et al., 2018).

2 PERIODIC TASK ALTERNATIONS

Dong et al. (2015) use periodic task alternations with equal training ratios for every task.

Approach Task Prioritisation

3 CONSECUTIVE TRAINING (HASHIMOTO ET AL., 2017)

In one epoch, iterate over the datasets in order of complexity;

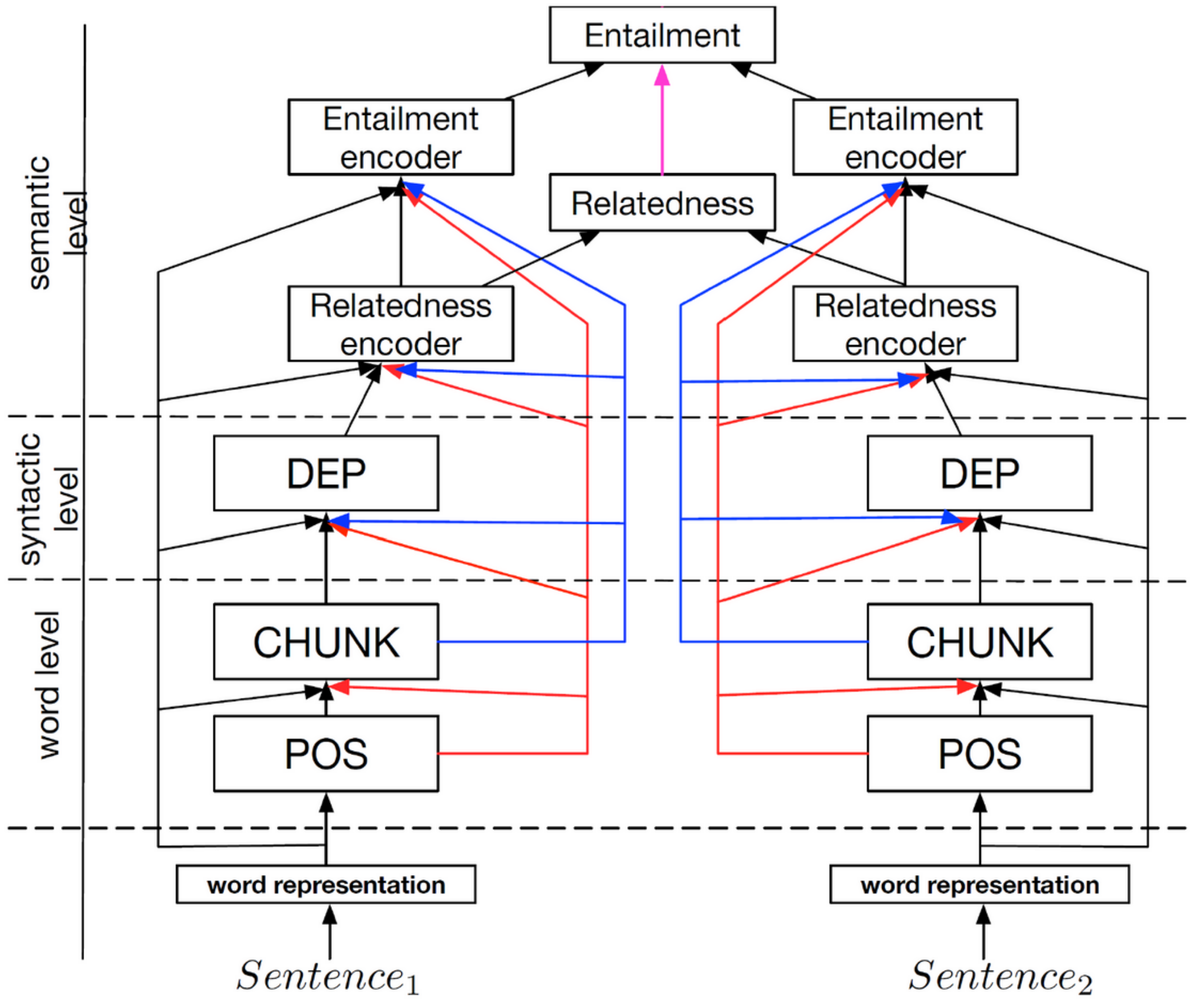
Introduce successive regularisation to avoid catastrophic forgetting.

$$J_5(\theta_{\text{ent}}) = -\sum_{(s,s')} \log p(y^{(5)}_{(s,s')} = \alpha | h^{(5)}_s, h^{(5)}_{s'}) + \lambda \|W_{\text{ent}}\|^2 + \delta \|\theta_{\text{rel}} - \theta'_{\text{rel}}\|^2,$$

task objective

task weight decay

successive regularisation



Approach Task Prioritisation

4 CURRICULUM LEARNING (BENGIO ET AL., 2009)

Start with easier subtasks and gradually increase the difficulty level.
Motivation from humans and animals who learn better when trained with a curriculum-like strategy.

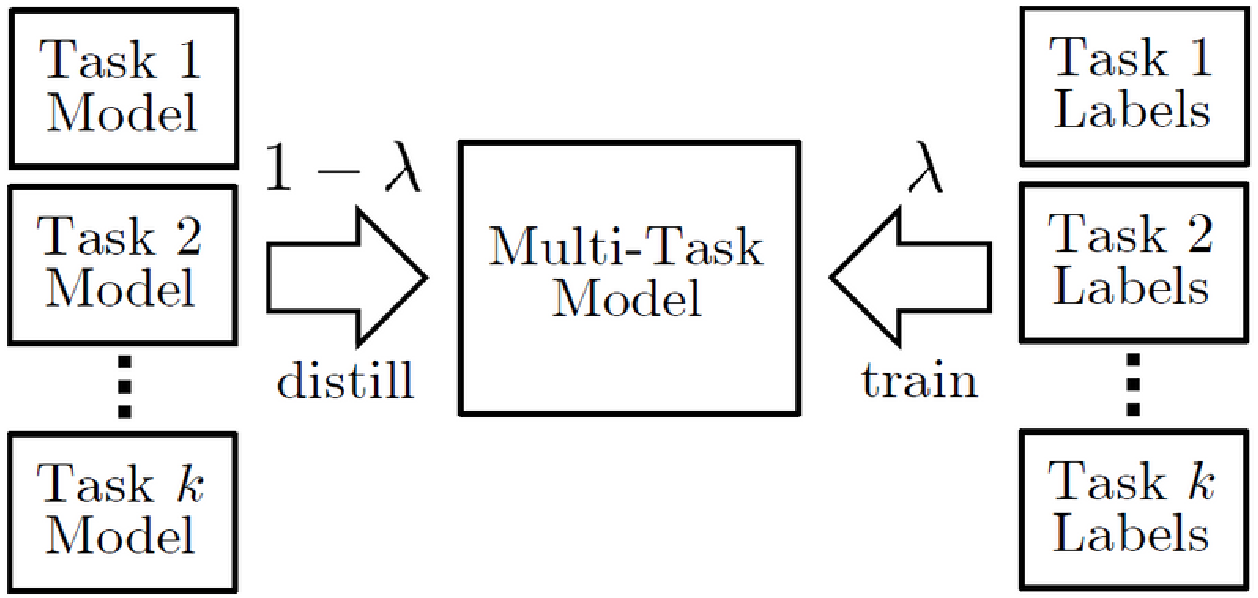
5 ANTI-CURRICULUM LEARNING

However, curriculum learning does not always work best: models converge faster on easier tasks.
McCann et al. (2018) of DecaNLP start with difficult tasks in phase 1 and add easy tasks in phase 2.

Approach Task Prioritisation

6 *ALTERNATIVE* TEACHER DISTILLATION

Teaching distillation from teachter (STL architectures) to student (MTL architecture) (Clark et al., 2019).



7 *ALTERNATIVE* TRANSDUCTIVE AUXILIARY TASK SELF-LEARNING

Bjerva et al. (2019) use the auxiliary task to train a STL model, which generates labels on the main task dataset. Subsequently, they train a MTL model on both tasks.

Approach Task Weights

1 HUMAN SUPERVISION

Fixed curriculum through human supervision by introducing per-task weights in the loss function.

2 SELF-PACED LEARNING

Dynamical adjustment of task weights according to normalisation requirements
- e.g. GradNorm by Chen et al. (2018).

3 PROGRESS-SIGNAL BASED CURRICULUM

Reinforcement learning inspired - e.g. dynamic task prioritisation by Guo et al. (2018).

Approach Task Weights

1 HUMAN SUPERVISION

Fixed curriculum through human supervision by introducing per-task weights in the loss function.

2 SELF-PACED LEARNING

Dynamical adjustment of task weights according to normalisation requirements
- e.g. GradNorm by Chen et al. (2018).

3 PROGRESS-SIGNAL BASED CURRICULUM

Reinforcement learning inspired - e.g. dynamic task prioritisation by Guo et al. (2018).

Tasks to combine

- STUDY 1

  Bingel and Søgaard (2017) research sequence labelling tasks' beneficiality pairwise.

- STUDY 2

  Changpinyo et al. (2018) present similar research, but move beyond pairwise comparisons.

Task Relations Study (1)

Bingel and Søgaard (2017) research when and why MTL works for task pairs:

- 10 SEQUENCE LABELLING TASKS
- HARD SHARING MODEL

  GloVe embeddings, hard shared Bi-LSTM and task-specific output layers.
- RANDOM SELECTION TRAINING STRATEGY

Task Relations Study (1)

1. Logical type tagging (CCG)
2. Chunking (CHU)
3. Sentence compression (COM)
4. Semantic frames (FNT)
5. POS tagging (POS)
6. Hyperlink prediction (HYP)
7. Keyphrase detection (KEY)
8. MWE detection (MWE)
9. Super-sense tagging 1 (SEM)
10. Super-sense tagging 2 (STR)

	CCG	CHU	COM	FNT	POS	HYP	KEY	MWE	SEM	STR
CCG		1.4	0.45	0.58	1.8	0.24	0.3	0.45	1.4	0.84
CHU	-0.052		-0.15	-0.12	-0.45	-0.5	-0.22	-0.27	-0.099	-0.32
COM	-5	1.3		1.3	-1.4	-2.4	-4.8	0.82	-3	-0.63
FNT	-5.8	-1	-6.1		-9.4	-5.7	-3.6	-9.4	-3	-0.68
POS	4.9	2.9	1.9	0.9		-0.85	-0.26	1.3	3.4	2.9
HYP	12	4	-11	9.2	22		1.5	-7.7	23	8.1
KEY	5.7	3.2	-1	-0.43	-1.3	-2.6		-4.7	0.59	0.69
MWE	18	20	7.4	5.5	1.6	-3.8	-5.8		16	8.6
SEM	-5	-0.76	-1.2	-0.81	-0.85	-1.3	-0.83	-1.1		-1.7
STR	-1.7	1.5	-0.26	-0.72	0.037	-1.5	-1.4	-1.6	1.7	

Gains and losses (%) in F1 for including auxiliary tasks (columns) with main tasks (rows).

Task Relations Study (1)

Most beneficial auxiliary task:

1. Logical type tagging (CCG)
2. Chunking (CHU)
3. Sentence compression (COM)
4. Semantic frames (FNT)
5. POS tagging (POS)
6. Hyperlink prediction (HYP)
7. Keyphrase detection (KEY)
8. MWE detection (MWE)
9. Super-sense tagging 1 (SEM)
10. Super-sense tagging 2 (STR)

	CCG	CHU	COM	FNT	POS	HYP	KEY	MWE	SEM	STR
CCG		1.4	0.45	0.58	1.8	0.24	0.3	0.45	1.4	0.84
CHU	-0.052		-0.15	-0.12	-0.45	-0.5	-0.22	-0.27	-0.099	-0.32
COM	-5	1.3		1.3	-1.4	-2.4	-4.8	0.82	-3	-0.63
FNT	-5.8	-1	-6.1		-9.4	-5.7	-3.6	-9.4	-3	-0.68
POS	4.9	2.9	1.9	0.9		-0.85	-0.26	1.3	3.4	2.9
HYP	12	4	-11	9.2	22		1.5	-7.7	23	8.1
KEY	5.7	3.2	-1	-0.43	-1.3	-2.6		-4.7	0.59	0.69
MWE	18	20	7.4	5.5	1.6	-3.8	-5.8		16	8.6
SEM	-5	-0.76	-1.2	-0.81	-0.85	-1.3	-0.83	-1.1		-1.7
STR	-1.7	1.5	-0.26	-0.72	0.037	-1.5	-1.4	-1.6	1.7	

Gains and losses (%) in F1 for including auxiliary tasks (columns) with main tasks (rows).

Task Relations Study (1)

Tasks that benefit most:

1. Logical type tagging (CCG)
2. Chunking (CHU)
3. Sentence compression (COM)
4. Semantic frames (FNT)
5. POS tagging (POS)
6. Hyperlink prediction (HYP)
7. Keyphrase detection (KEY)
8. MWE detection (MWE)
9. Super-sense tagging 1 (SEM)
10. Super-sense tagging 2 (STR)

	CCG	CHU	COM	FNT	POS	HYP	KEY	MWE	SEM	STR
CCG		1.4	0.45	0.58	1.8	0.24	0.3	0.45	1.4	0.84
CHU	-0.052		-0.15	-0.12	-0.45	-0.5	-0.22	-0.27	-0.099	-0.32
COM	-5	1.3		1.3	-1.4	-2.4	-4.8	0.82	-3	-0.63
FNT	-5.8	-1	-6.1		-9.4	-5.7	-3.6	-9.4	-3	-0.68
POS	4.9	2.9	1.9	0.9		-0.85	-0.26	1.3	3.4	2.9
HYP	12	4	-11	9.2	22		1.5	-7.7	23	8.1
KEY	5.7	3.2	-1	-0.43	-1.3	-2.6		-4.7	0.59	0.69
MWE	18	20	7.4	5.5	1.6	-3.8	-5.8		16	8.6
SEM	-5	-0.76	-1.2	-0.81	-0.85	-1.3	-0.83	-1.1		-1.7
STR	-1.7	1.5	-0.26	-0.72	0.037	-1.5	-1.4	-1.6	1.7	

Gains and losses (%) in F1 for including auxiliary tasks (columns) with main tasks (rows).

Task Relations Study (1)

Symbiotic relations:

1 Logical type tagging (CCG)
2 Chunking (CHU)
3 Sentence compression (COM)
4 Semantic frames (FNT)
5 POS tagging (POS)
6 Hyperlink prediction (HYP)
7 Keyphrase detection (KEY)
8 MWE detection (MWE)
9 Super-sense tagging 1 (SEM)
10 Super-sense tagging 2 (STR)

	CCG	CHU	COM	FNT	POS	HYP	KEY	MWE	SEM	STR
CCG		1.4	0.45	0.58	1.8	0.24	0.3	0.45	1.4	0.84
CHU	-0.052		-0.15	-0.12	-0.45	-0.5	-0.22	-0.27	-0.099	-0.32
COM	-5	1.3		1.3	-1.4	-2.4	-4.8	0.82	-3	-0.63
FNT	-5.8	-1	-6.1		-9.4	-5.7	-3.6	-9.4	-3	-0.68
POS	4.9	2.9	1.9	0.9		-0.85	-0.26	1.3	3.4	2.9
HYP	12	4	-11	9.2	22		1.5	-7.7	23	8.1
KEY	5.7	3.2	-1	-0.43	-1.3	-2.6		-4.7	0.59	0.69
MWE	18	20	7.4	5.5	1.6	-3.8	-5.8		16	8.6
SEM	-5	-0.76	-1.2	-0.81	-0.85	-1.3	-0.83	-1.1		-1.7
STR	-1.7	1.5	-0.26	-0.72	0.037	-1.5	-1.4	-1.6	1.7	

Gains and losses (%) in F1 for including auxiliary tasks (columns) with main tasks (rows).

Task Relations Study (1)

Using logistic regression, they predict MTL gains and losses from dataset statistics (e.g. size or label distribution entropy) and STL model characteristics (e.g. loss curve values).

- GOOD PREDICTOR: LOSS PLATEAU

  MTL gains are more likely for main tasks that quickly plateau with non-plateauing auxiliary tasks.

- GOOD PREDICTOR: LABEL ENTROPY AUXILIARY TASK

- BAD PREDICTOR: DATASET SIZES

Task Relations Study (2)

Changpinyo et al. (2018) move beyond pairwise comparisons:

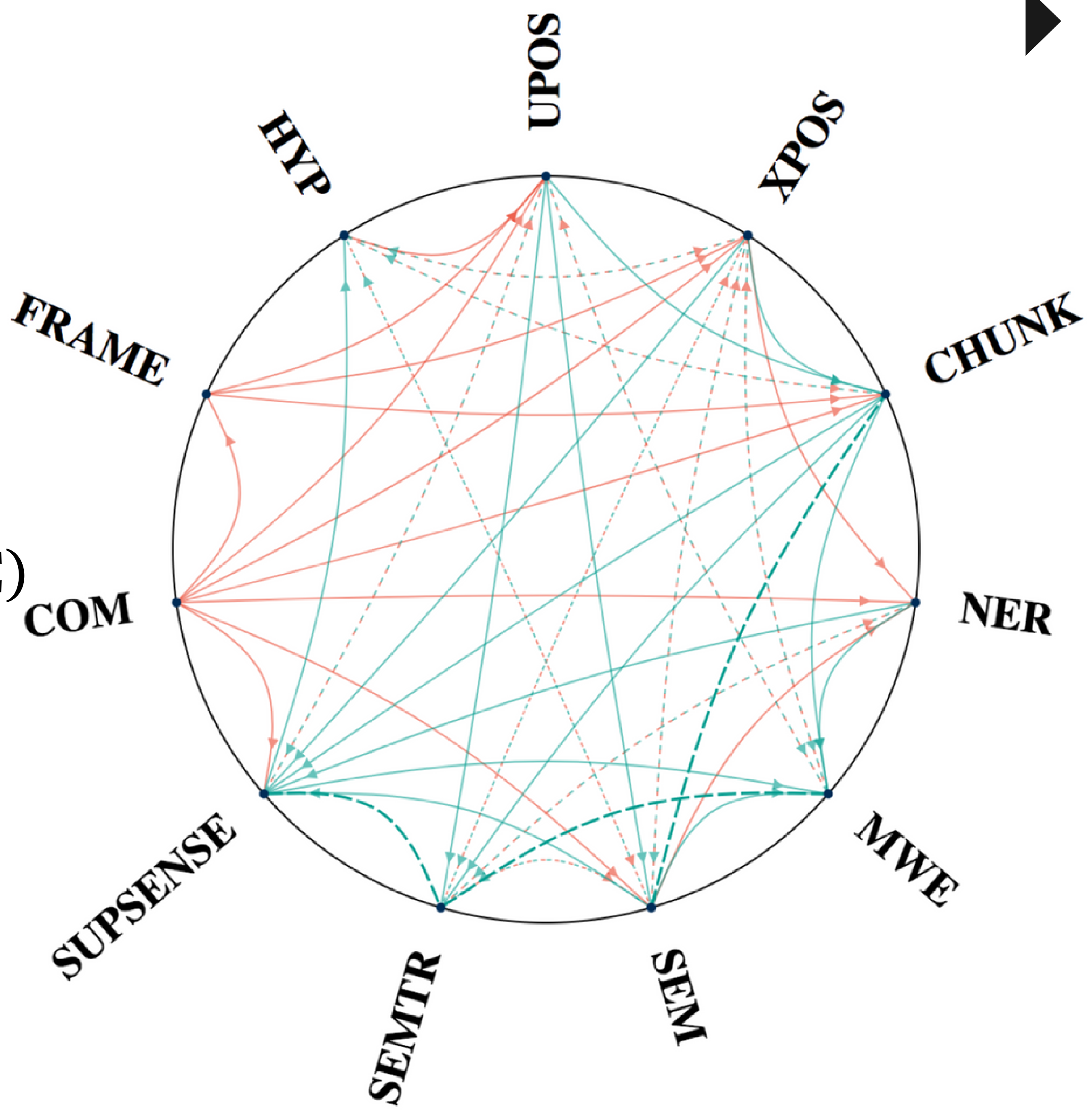
- 11 SEQUENCE LABELLING TASKS
- HARD SHARING MODELS

  (1) Hard sharing with task-specific output layers.

  (2) Hard sharing of all layers , but with task embeddings.
- UNIFORM TRAINING STRATEGY

Task Relations Study (2)

1 POS tagging (UPOS, XPOS)
2 Chunking (CHUNK)
3 Named Entity Recognition (NER)
4 MWE identification (MWE)
5 Super-sense tagging (SEM, SUPSENSE)
6 Semantic trait tagging (SEMTR)
7 Sentence compression (COM)
8 Semantic frame prediction (FRAME)
9 Hyperlink detection (HYP)

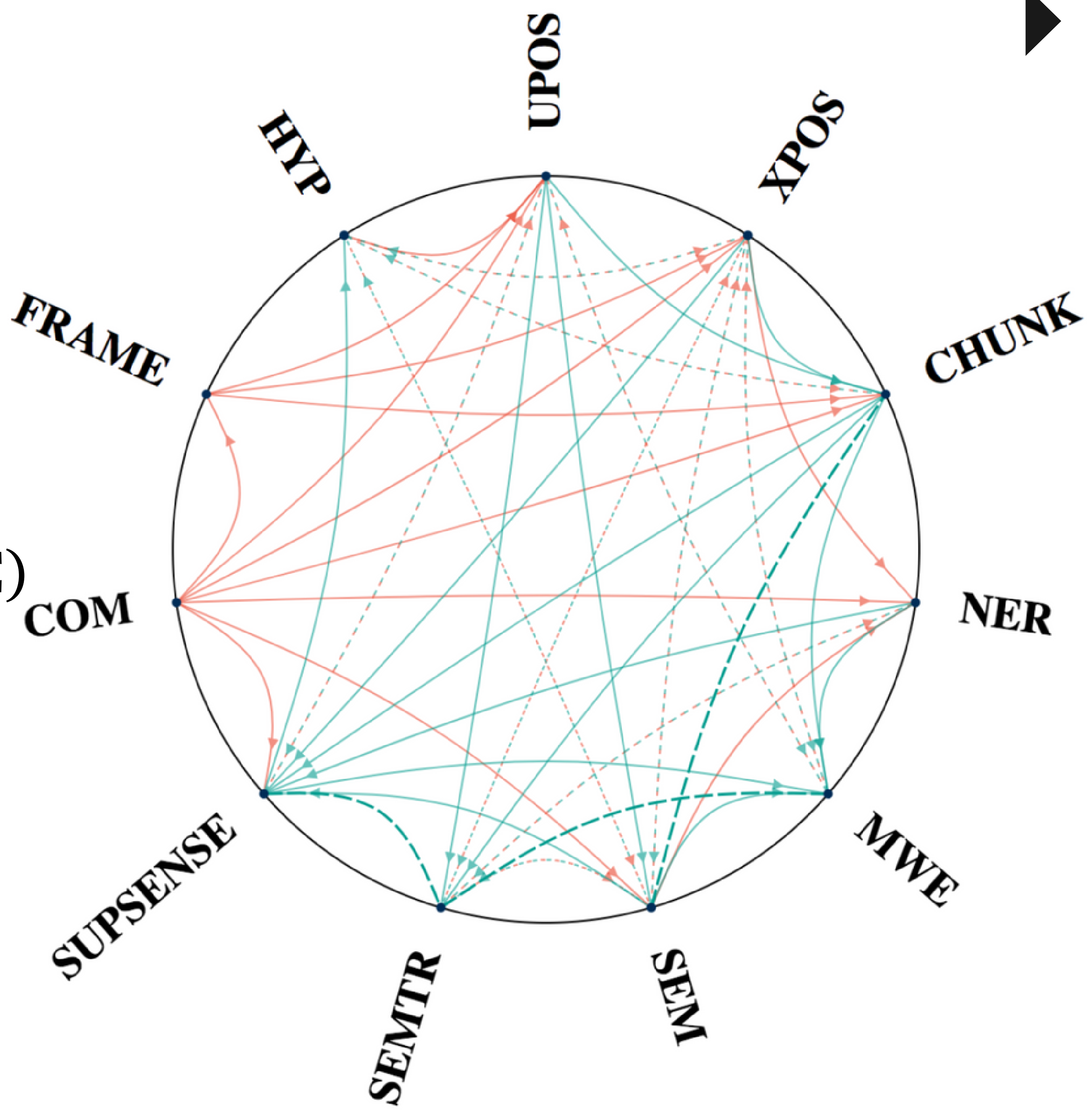


- Pairwise MTL relations, green is beneficial, red is harming, dotted is asymmetric.

Task Relations Study (2)

Main tasks that benefit:

1. POS tagging (UPOS, XPOS)
2. Chunking (CHUNK)
3. Named Entity Recognition (NER)
4. MWE identification (MWE)
5. Super-sense tagging (SEM, SUPSENSE)
6. Semantic trait tagging (SEMTR)
7. Sentence compression (COM)
8. Semantic frame prediction (FRAME)
9. Hyperlink detection (HYP)

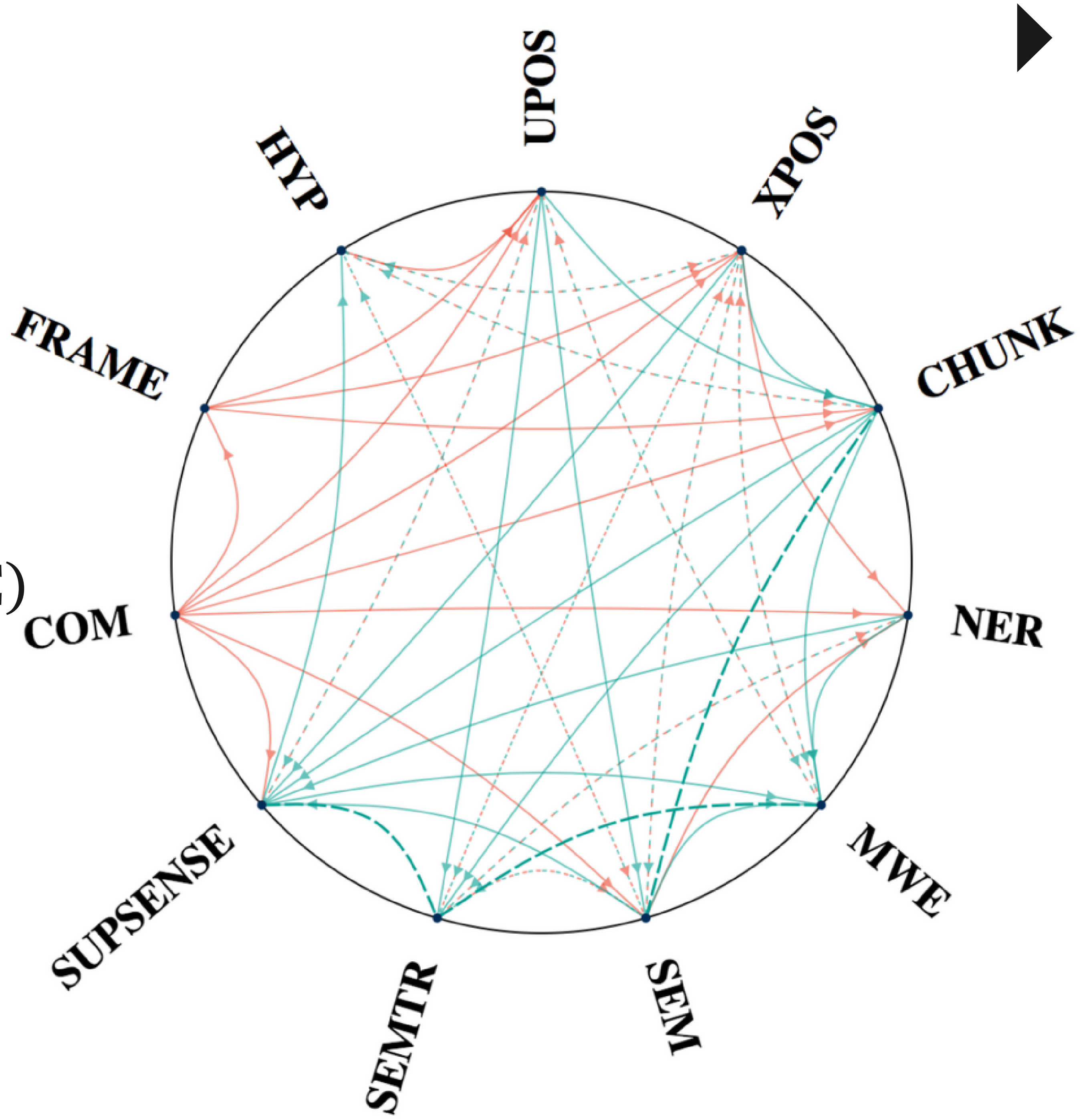


- Pairwise MTL relations, green is beneficial, red is harming, dotted is asymmetric.

Task Relations Study (2)

Auxiliary tasks that are beneficial:

1. POS tagging (UPOS, XPOS)
2. Chunking (CHUNK)
3. Named Entity Recognition (NER)
4. MWE identification (MWE)
5. Super-sense tagging (SEM, SUPSENSE)
6. Semantic trait tagging (SEMTR)
7. Sentence compression (COM)
8. Semantic frame prediction (FRAME)
9. Hyperlink detection (HYP)

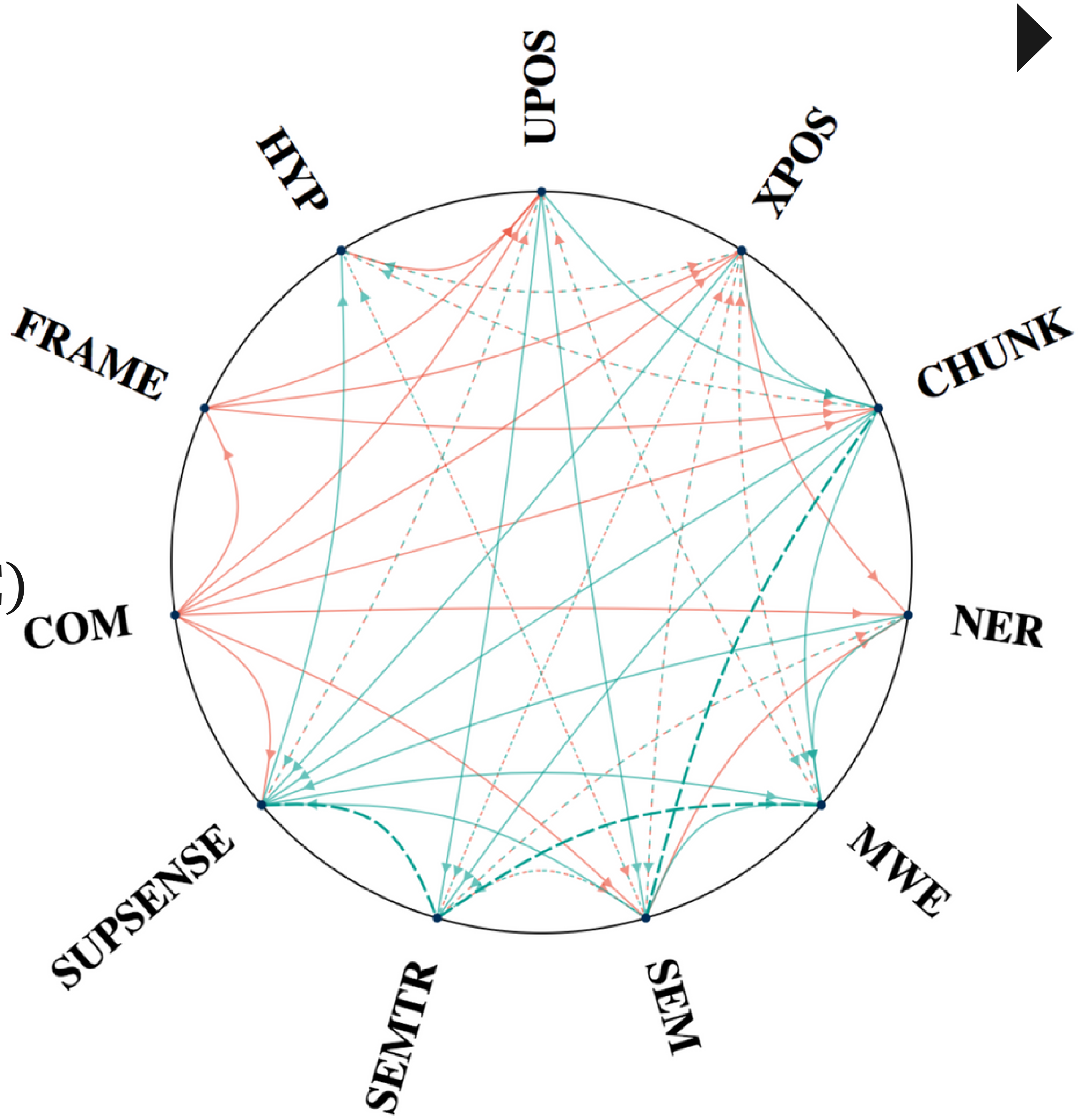


- Pairwise MTL relations, green is beneficial, red is harming, dotted is asymmetric.

Task Relations Study (2)

Harmful task:

1 POS tagging (UPOS, XPOS)
2 Chunking (CHUNK)
3 Named Entity Recognition (NER)
4 MWE identification (MWE)
5 Super-sense tagging (SEM, SUPSENSE)
6 Semantic trait tagging (SEMTR)
7 Sentence compression (COM)
8 Semantic frame prediction (FRAME)
9 Hyperlink detection (HYP)

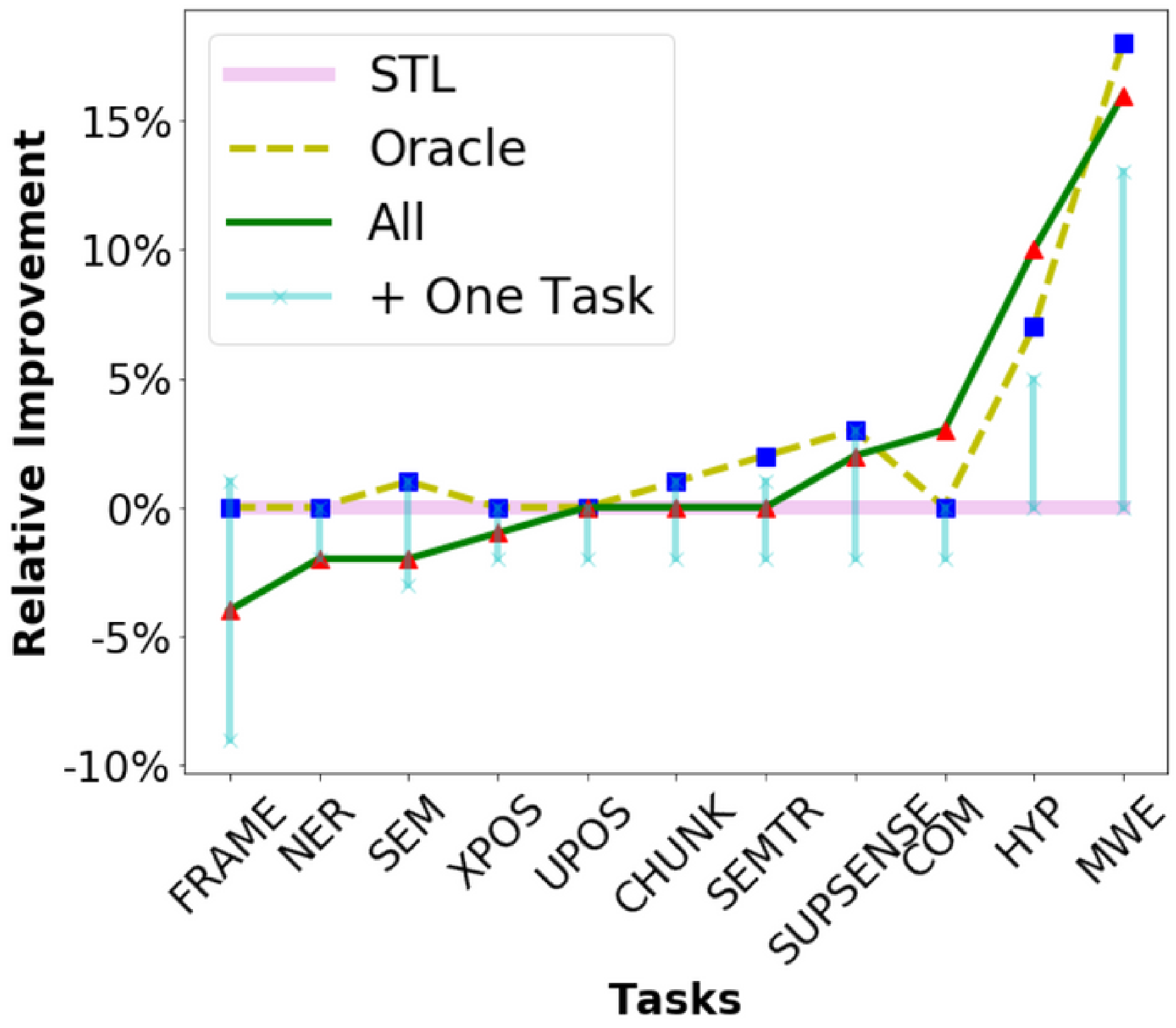


- Pairwise MTL relations, green is beneficial, red is harming, dotted is asymmetric.

Task Relations Study (2)

Compare Oracle (only beneficial tasks) to pairwise, STL and all:

- ORACLE >= STL
- ORACLE > PAIRWISE
- ORACLE > ALL

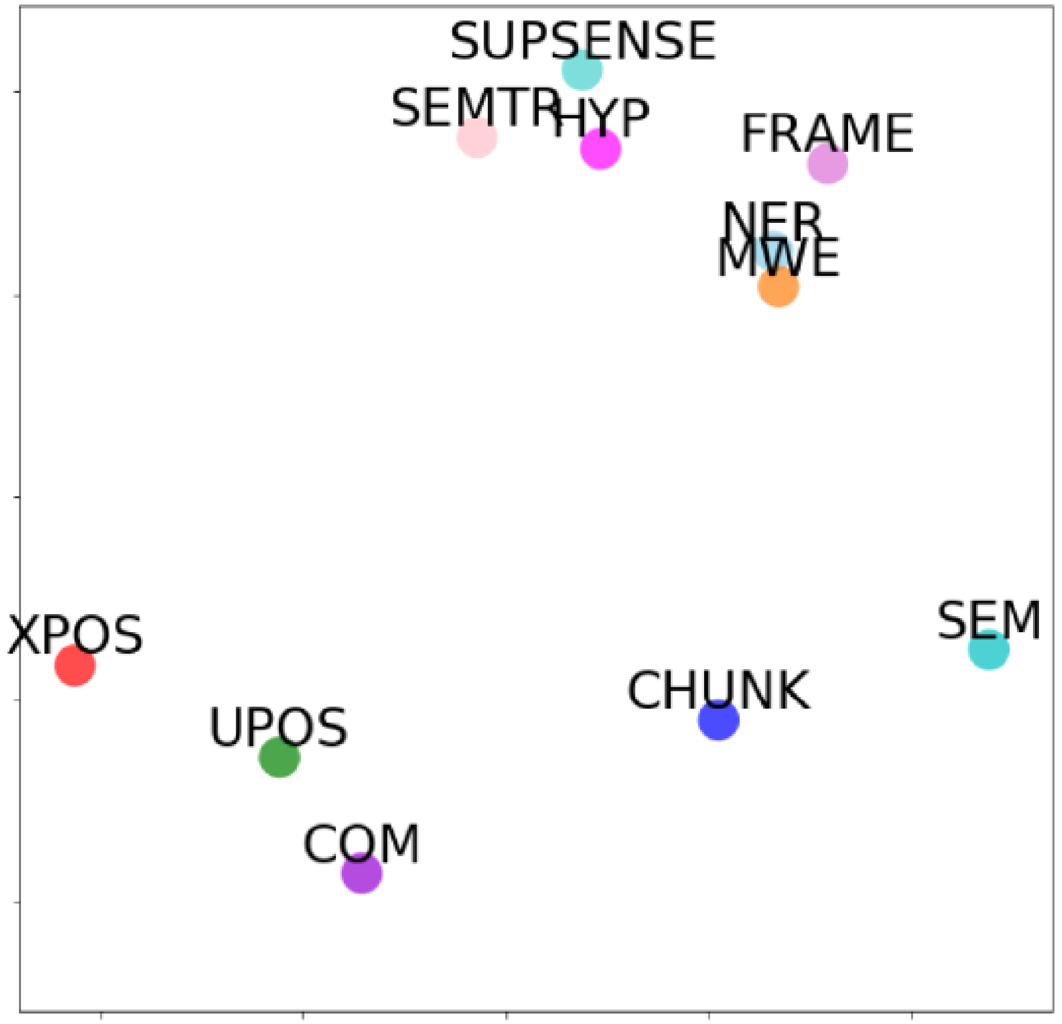


- Relative gains and losses for all experimental setups.

Task Relations Study (2)

The authors visualise task embeddings learnt in hard-shared setup with task embeddings:

- SYNTACTIC VS. SEMANTIC
- DATASET NOT INDICATIVE
- LABEL ENTROPY NOT INDICATIVE



- t-SNE visualisation of task embeddings.

It's Q&A time: raise your digital Zoom hand!

🙋🏾‍♂️

References

Barrett, M., Bingel, J., Hollenstein, N., Rei, M., and Søgaard, A. (2018) Sequence classification with human attention. In Proceedings of the 22nd Conference on Computational Natural Language Learning (pages 302-312)

Bengio, Y., Louradour, J., Collobert, R., and Weston, J. (2009). Curriculum learning. In Proceedings of the 26th annual international conference on machine learning (pages 41-48)

Bingel, J., and Søgaard, A. (2017). Identifying beneficial task relations for multi-task learning in deep neural networks. In Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers (pages 164-169)

Bjerva, J., Kann, K., and Augenstein, I. (2019). Transductive Auxiliary Task Self-Training for Neural Multi-Task Models. EMNLP-IJCNLP 2019, 253.

Caruana, R. (1993) Multitask learning: A knowledge-based source of inductive bias. In ICML.

Changpinyo, S., Hu, H., and Sha, F. (2018) Multi-Task Learning for Sequence Tagging: An Empirical Study. In Proceedings of the 27th International Conference on Computational Linguistics (pages 2965-2977)

Chen, Z., Badrinarayanan, V., Lee, c., and Rabinovich, A. (2018) Gradnorm: Gradient normalization for adaptive loss balancing in deep multitask networks. In International Conference on Machine Learning (pages 793-802)

Cheng, H., Fang, H., and Ostendorf, M. (2015) Open-domain name error detection using a multi-task RNN. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (pages 737-746)

References

Clark, K., Luong, M. T., Khandelwal, U., Manning, C. D., and Le, Q. (2019). BAM! Born-Again Multi-Task Networks for Natural Language Understanding. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (pages 5931-5937)

Dankers, V., Rei, M., Lewis, M., and Shutova, E. (2019). Modelling the interplay of metaphor and emotion through multitask learning. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP) (pages 2218-2229)

Dong, D., Wu, H., He, W., Yu, D., and Wang, H. (2015). Multi-Task Learning for Multiple Language Translation. In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers) (pages 1723-1732)

Guo, M., Haque, A., Huang, D., Yeung, S., and Fei-Fei, L. (2018) Dynamic task prioritization for multitask learning. In European Conference on Computer Vision (pages 282-299)

Hashimoto, K., Xiong, C., Tsuruoka, Y., and Socher, R. (2017). A Joint Many-Task Model: Growing a Neural Network for Multiple NLP Tasks. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (pages 1923-1933)

Li, Q., Zhang, Q., and Si, L. (2019). Rumor Detection by Exploiting User Credibility Information, Attention and Multi-task Learning. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (pages 1173-1179)

Søgaard, A., and Goldberg, Y. (2016) Deep multi-task learning with low level tasks supervised at lower layers. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers) (pages 231-235)

References

Liu, X., He, P., Chen, W., and Gao, J. (2019). Multi-Task Deep Neural Networks for Natural Language Understanding. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (pages 4487-4496)

McCann, B., Keskar, N. S., Xiong, C., and Socher, R. (2018). The natural language decathlon: Multitask learning as question answering. arXiv preprint arXiv:1806.08730.

Rei, M. (2017). Semi-supervised Multitask Learning for Sequence Labeling. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pages 2121-2130)

Ruder, S., Bingel, J., Augenstein, I., and Søgaard, A. (2019). Latent multi-task architecture learning. In Proceedings of the AAAI Conference on Artificial Intelligence, (pages 4822-4829)

Sanh, V., Wolf, T., and Ruder, s. (2018) A hierarchical multi-task approach for learning embeddings from semantic tasks. In Thirty-Second AAAI Conference on Artificial Intelligence.

Søgaard, A., and Goldberg, Y. (2016) Deep multi-task learning with low level tasks supervised at lower layers. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers) (pages 231-235)