VERNA DANKERS
University of Edinburgh

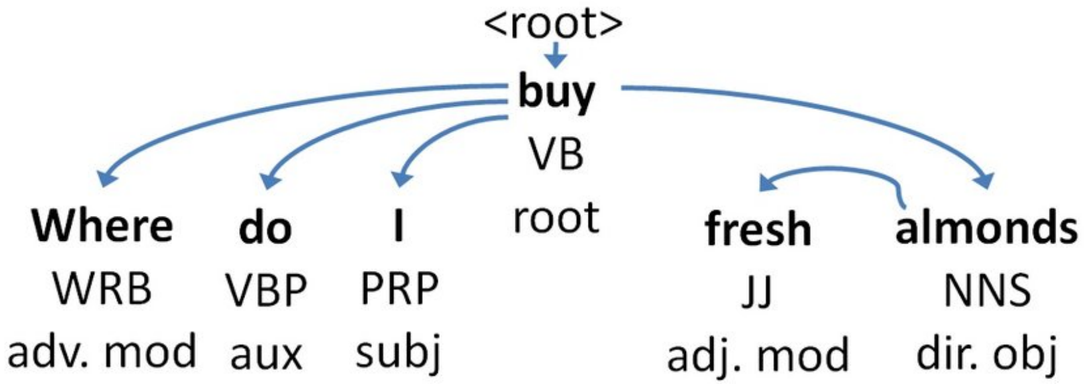
MULTITASK LEARNING

ATCS Lecture, April 16 2021

Contents

Introduction Motivation

- IMPROVE MAIN TASK THROUGH AUXILIARY TASKS



- MOVE TOWARDS A UNIFIED (NLP) ARCHITECTURE

Introduction Motivation

Moving towards a unified NLP architecture: GPT-3

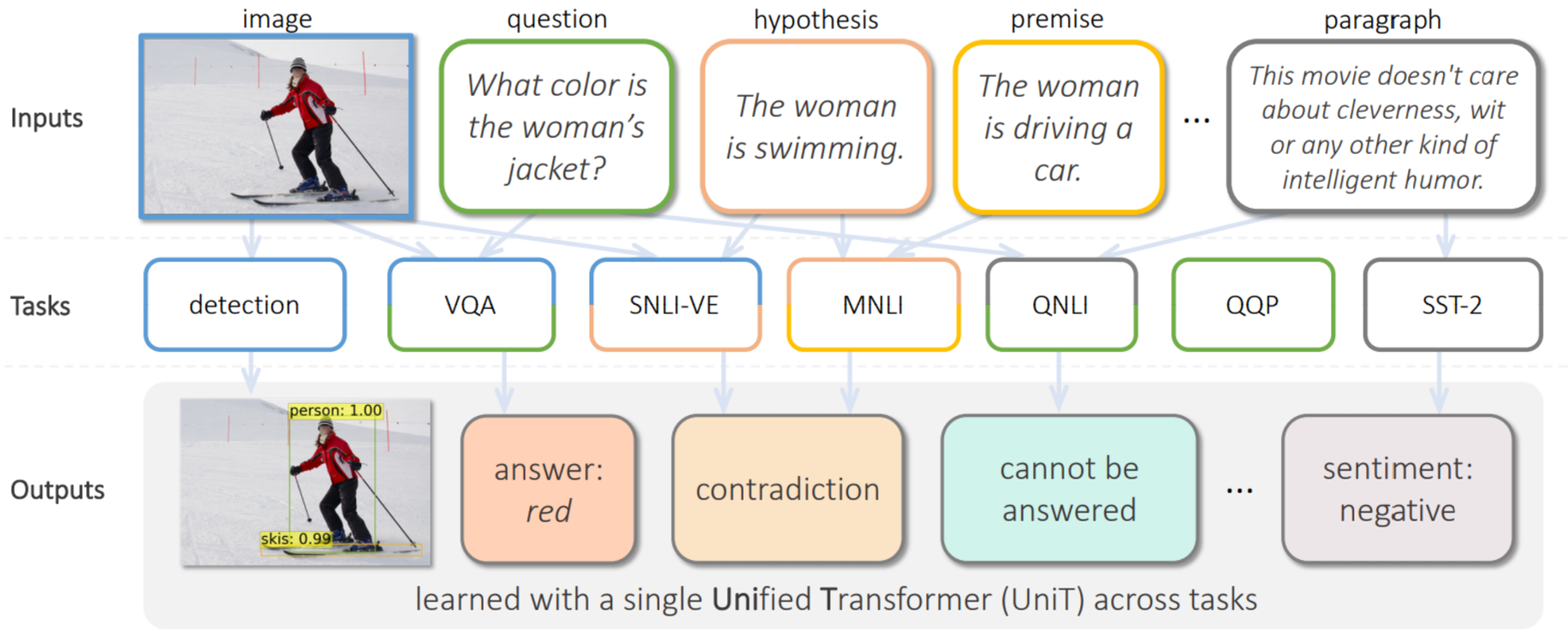
To "screeg" something is to swing a sword at it. An example of a sentence that uses the word screeg is:
We screeghed at each other for several minutes and then we went outside and ate ice cream.

Poor English input: I have tried to hit ball with bat, but my swing is has miss.
Good English output: I tried to hit the ball with the bat, but my swing missed.

Title: United Methodists Agree to Historic Split
Subtitle: Those who oppose gay marriage will form their own denomination
Article: **After two days of intense debate, the United Methodist Church has agreed to a historic split - one that is expected to end in the creation of a new denomination, one that will be "theologically and socially conservative," according to The Washington Post. The majority of delegates attending the church's annual General Conference in May voted to strengthen a ban on the ordination of LGBTQ clergy and to write new rules that will "discipline" clergy who officiate at same-sex weddings. But those who opposed these measures have a new plan: They say they will form a separate denomination by 2020, calling their church the Christian Methodist denomination.**

Introduction Motivation

UniT, multimodal transformer from Facebook AI, by Hu and Singh (2021). *preprint



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.

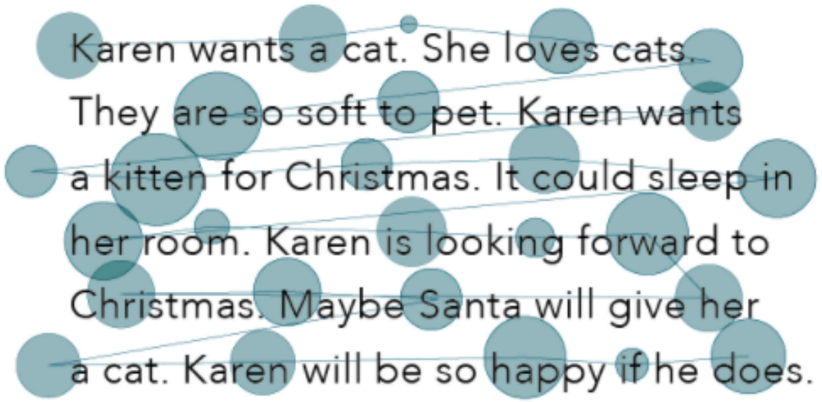
Introduction Inductive Biases

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

3 ATTRIBUTE SELECTION

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

E.g. use Barrett et al. (2018) use gaze prediction (auxiliary task) to allow other NLP tasks (i.a. sentiment analysis, abusive language detection) to focus on relevant input words.



Introduction Inductive Biases

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

4 EAVESDROPPING

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

Auxiliary tasks of sentiment and emotion prediction aid sarcasm detection in a multimodal model (Chauhan et al., 2020).

Approach

1 CHOOSE YOUR TASKS

2 DESIGN THE NETWORK ARCHITECTURE

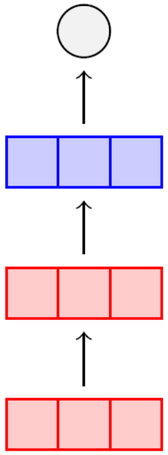
3 SELECT THE DATA

4 TASK PRIORITISATION DURING TRAINING

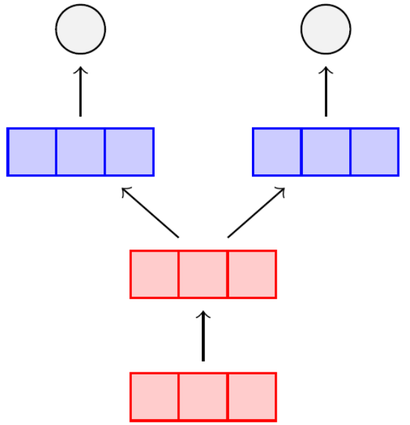
Task weighting in loss;

Regulation of parameter update frequencies (especially for separate datasets).

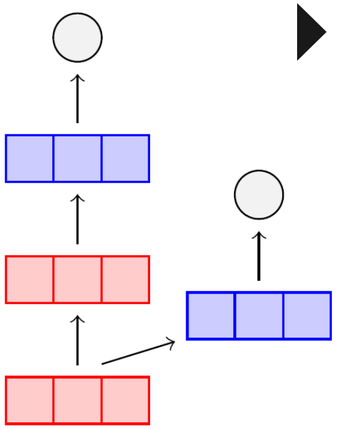
Approach Network Architecture



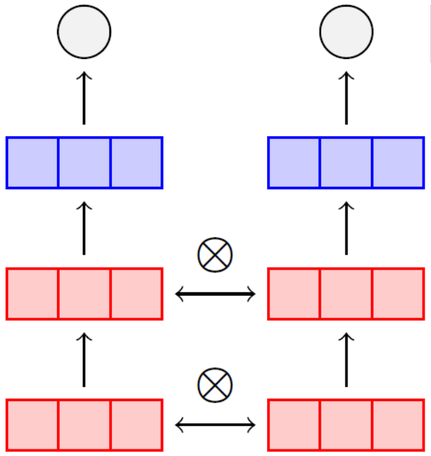
▶ FULL SHARING



▶ HARD SHARING

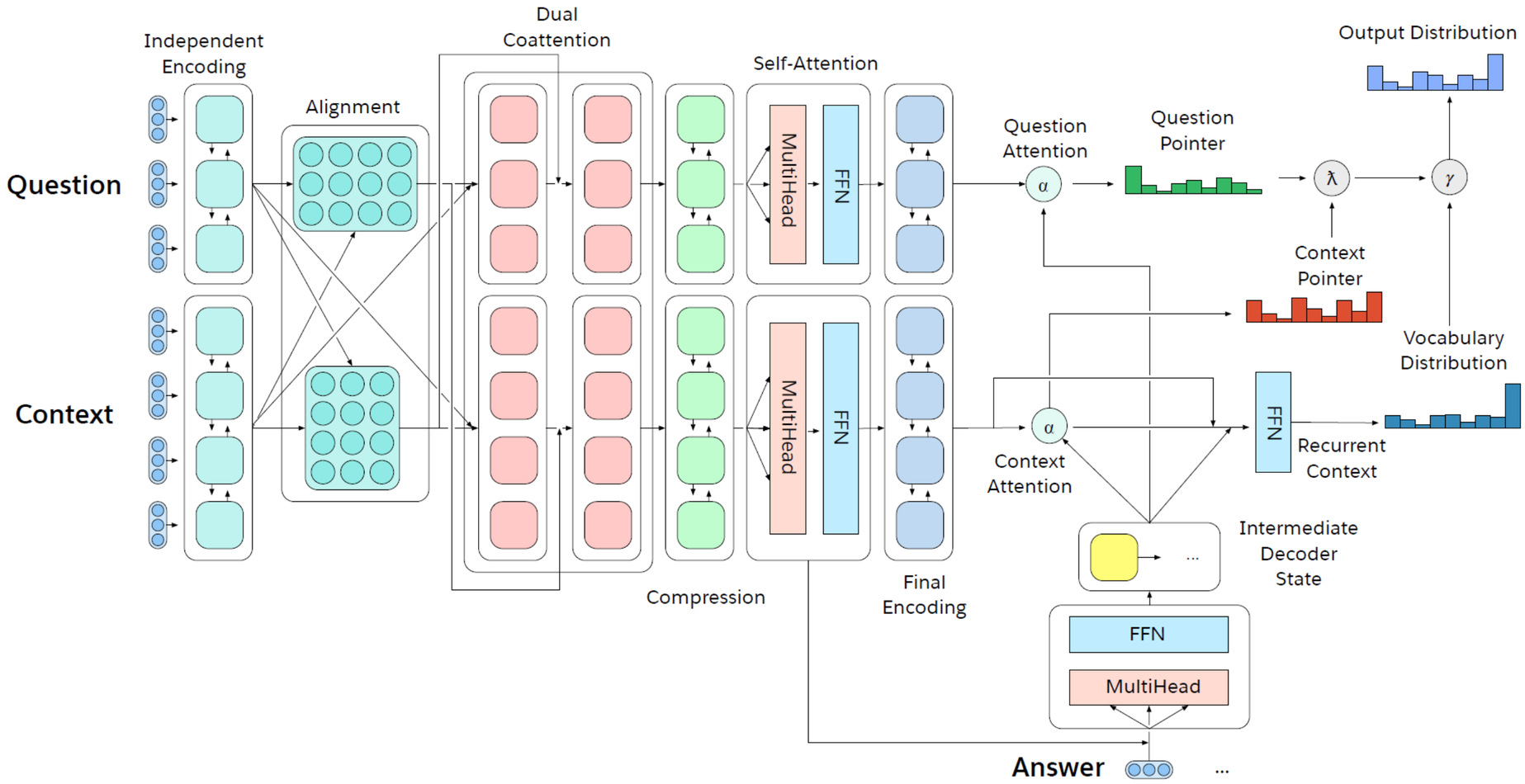


▶ HIERARCHICAL SHARING



▶ SOFT SHARING

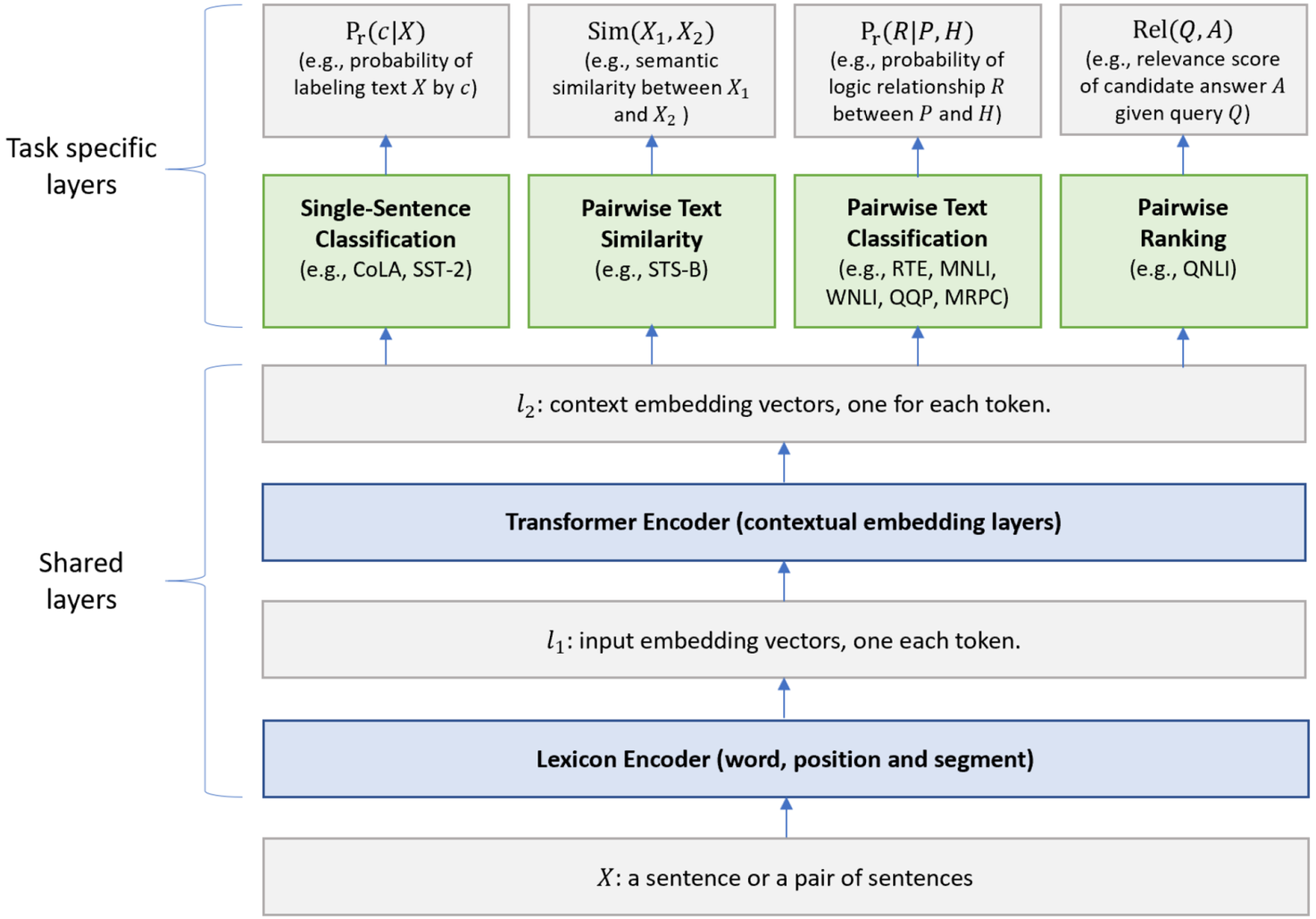
Approach Network Architecture



FULL SHARING

DecaNLP (McCann et al., 2018), define multiple tasks as question answering.

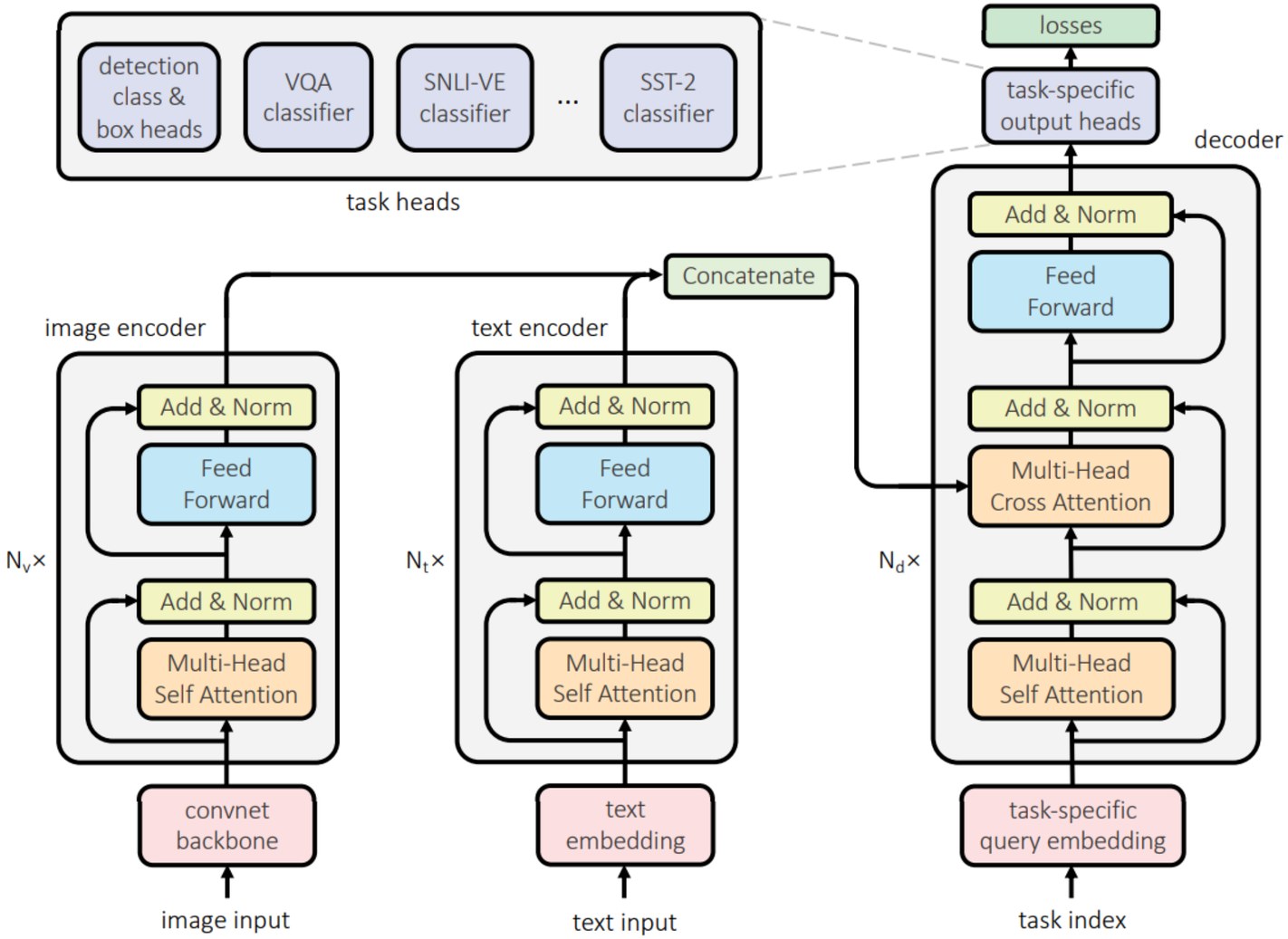
Approach Network Architecture



HARD SHARING

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

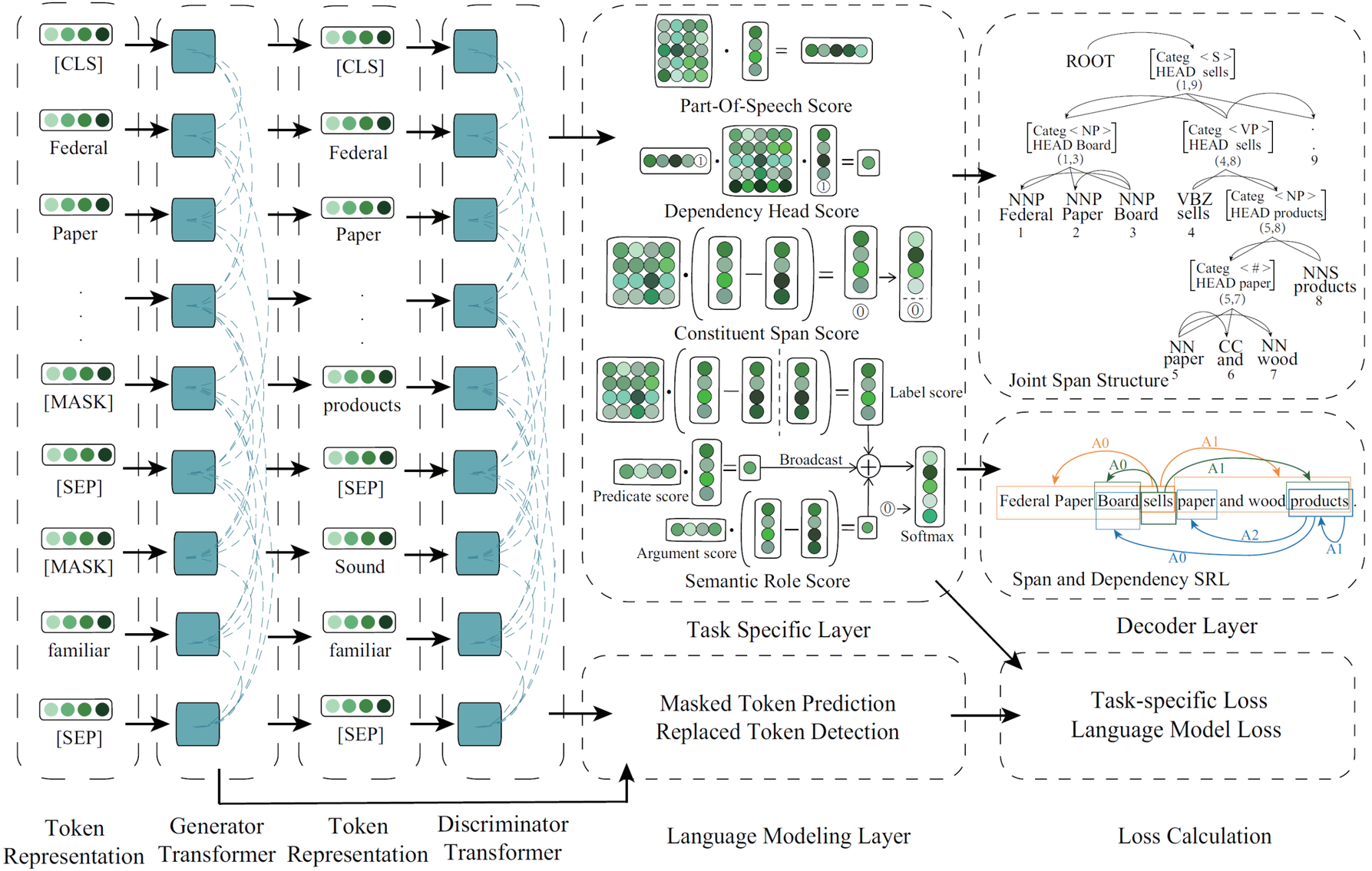
Approach Network Architecture



HARD SHARING

UniT by Hu and Singh (2021), with Transformer encoder and decoder shared by various tasks.

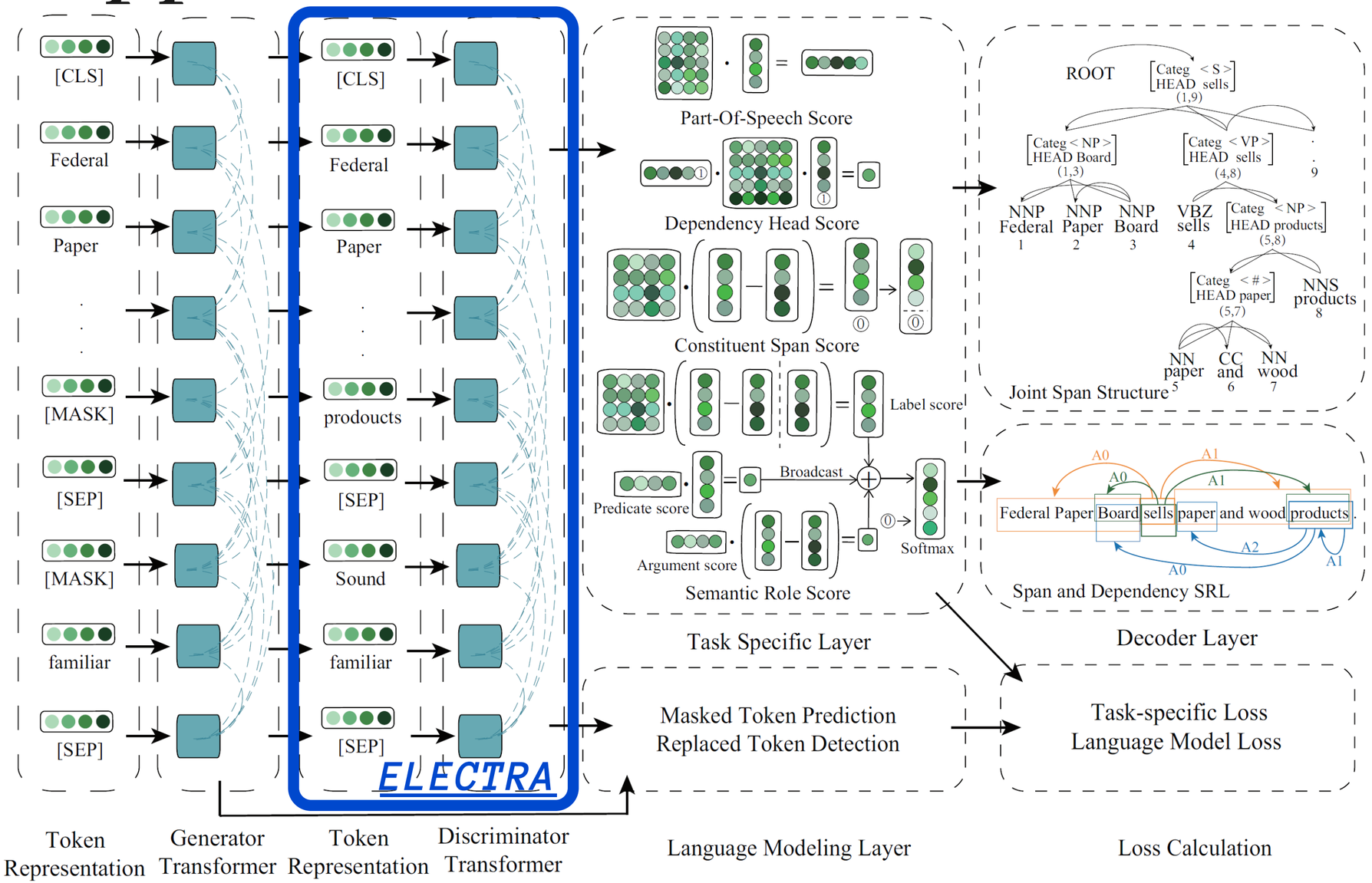
Approach Network Architecture



HARD SHARING

LIMIT-Bert by Zhou et al. (2020), that trains BERT with ELECTRA on 5 tasks and applies Syntactic/Semantic Phrase Masking.

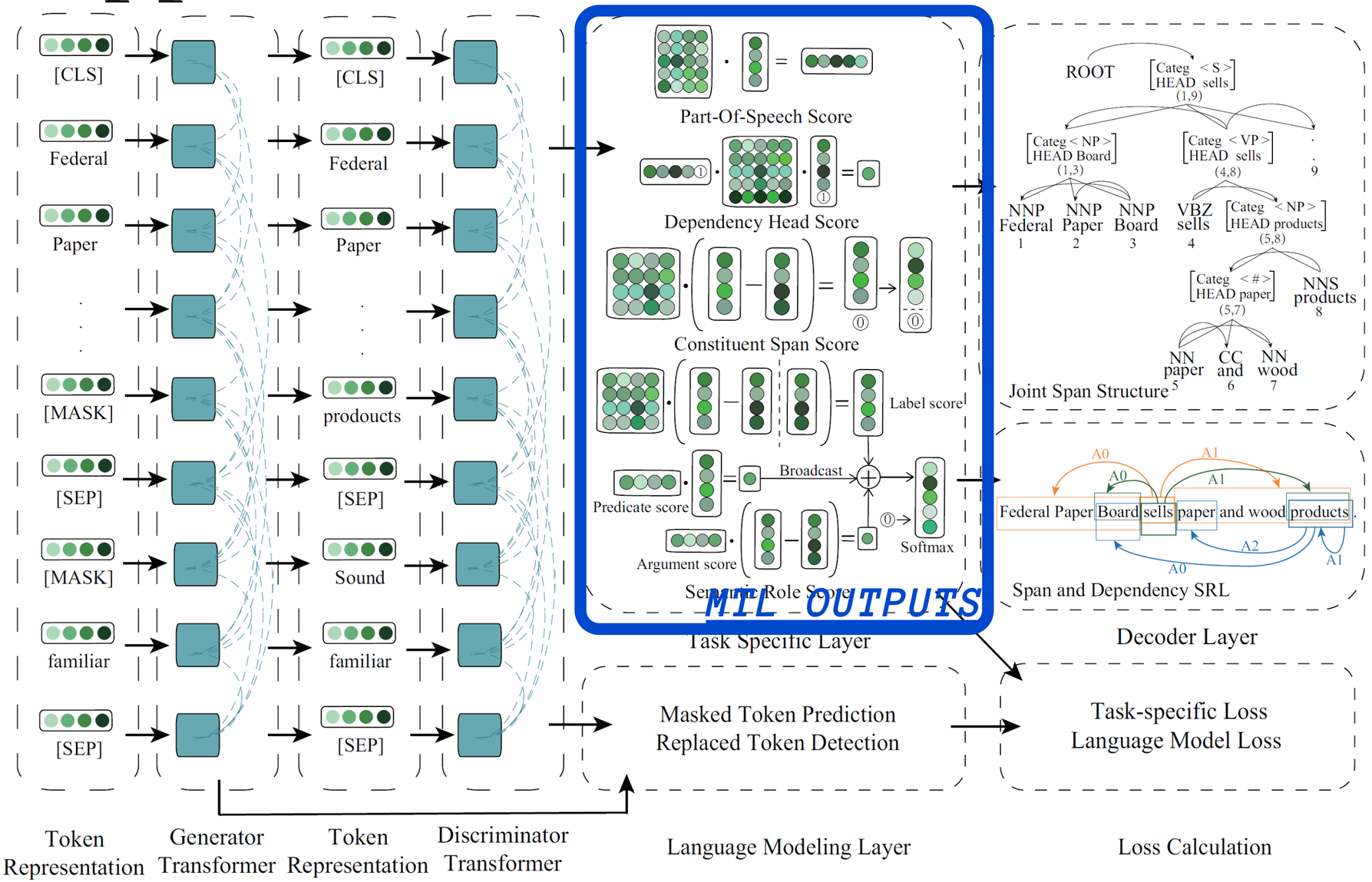
Approach Network Architecture



HARD SHARING

LIMIT-Bert by Zhou et al. (2020), that trains BERT with ELECTRA on 5 tasks and applies Syntactic/Semantic Phrase Masking.

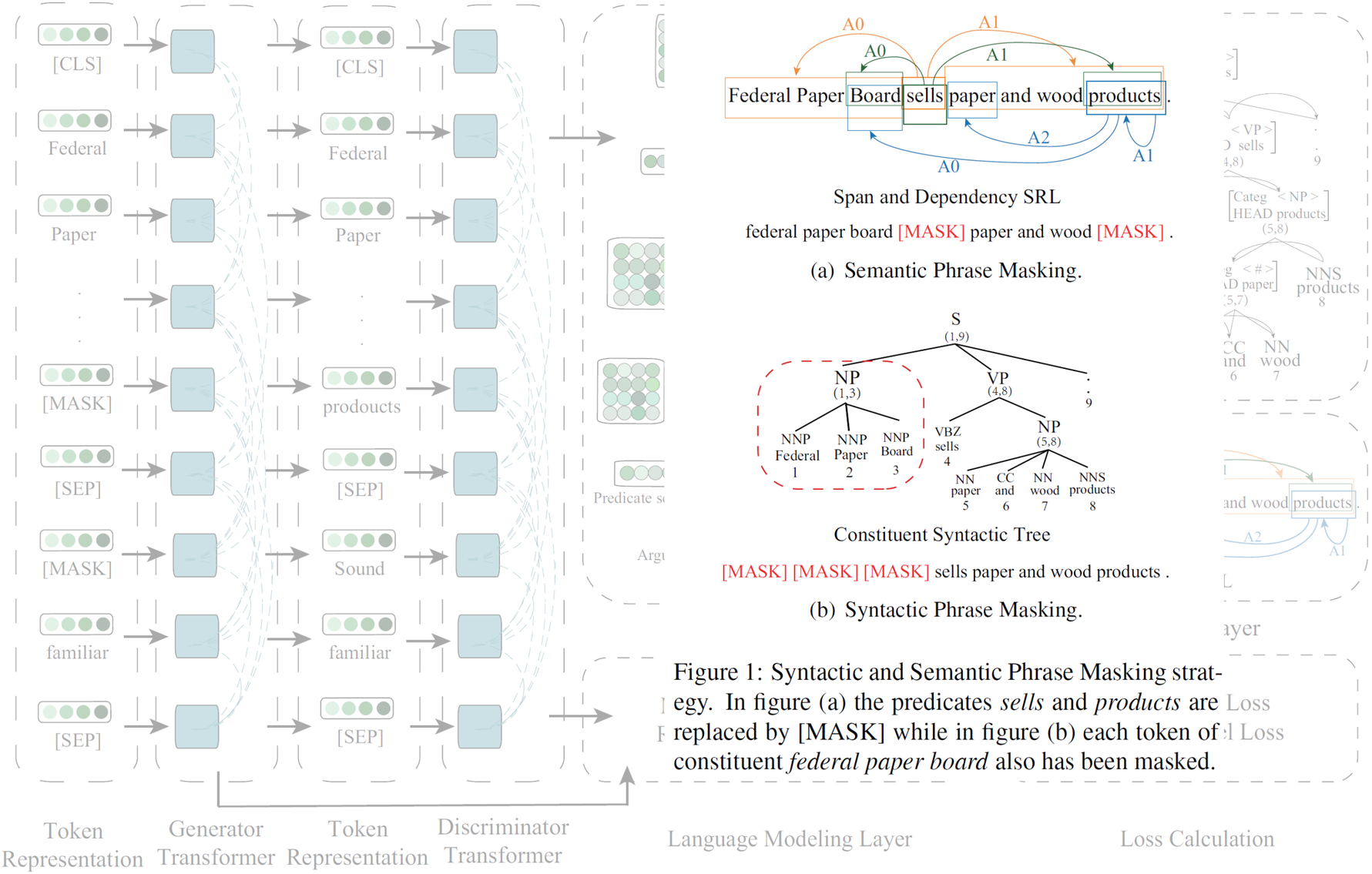
Approach Network Architecture



HARD SHARING

LIMIT-Bert by Zhou et al. (2020), that trains BERT with ELECTRA on 5 tasks and applies Syntactic/Semantic Phrase Masking.

Approach Network Architecture



Span and Dependency SRL

federal paper board [MASK] paper and wood [MASK] .

(a) Semantic Phrase Masking.

Constituent Syntactic Tree

[MASK] [MASK] [MASK] sells paper and wood products .

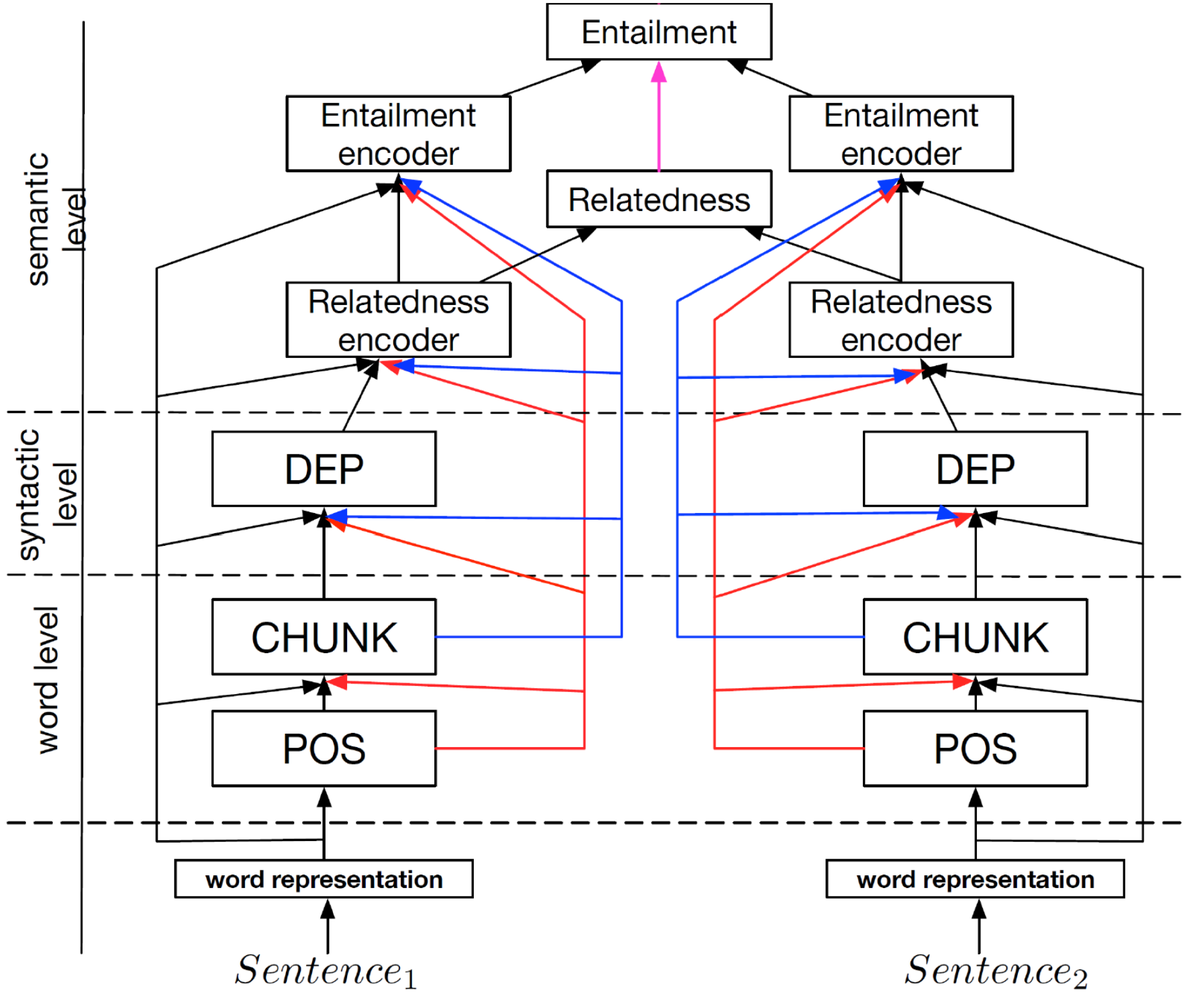
(b) Syntactic Phrase Masking.

Figure 1: Syntactic and Semantic Phrase Masking strategy. In figure (a) the predicates *sells* and *products* are replaced by [MASK] while in figure (b) each token of constituent *federal paper board* also has been masked.

HARD SHARING

LIMIT-Bert by Zhou et al. (2020), that trains BERT with ELECTRA on 5 tasks and applies Syntactic/Semantic Phrase Masking.

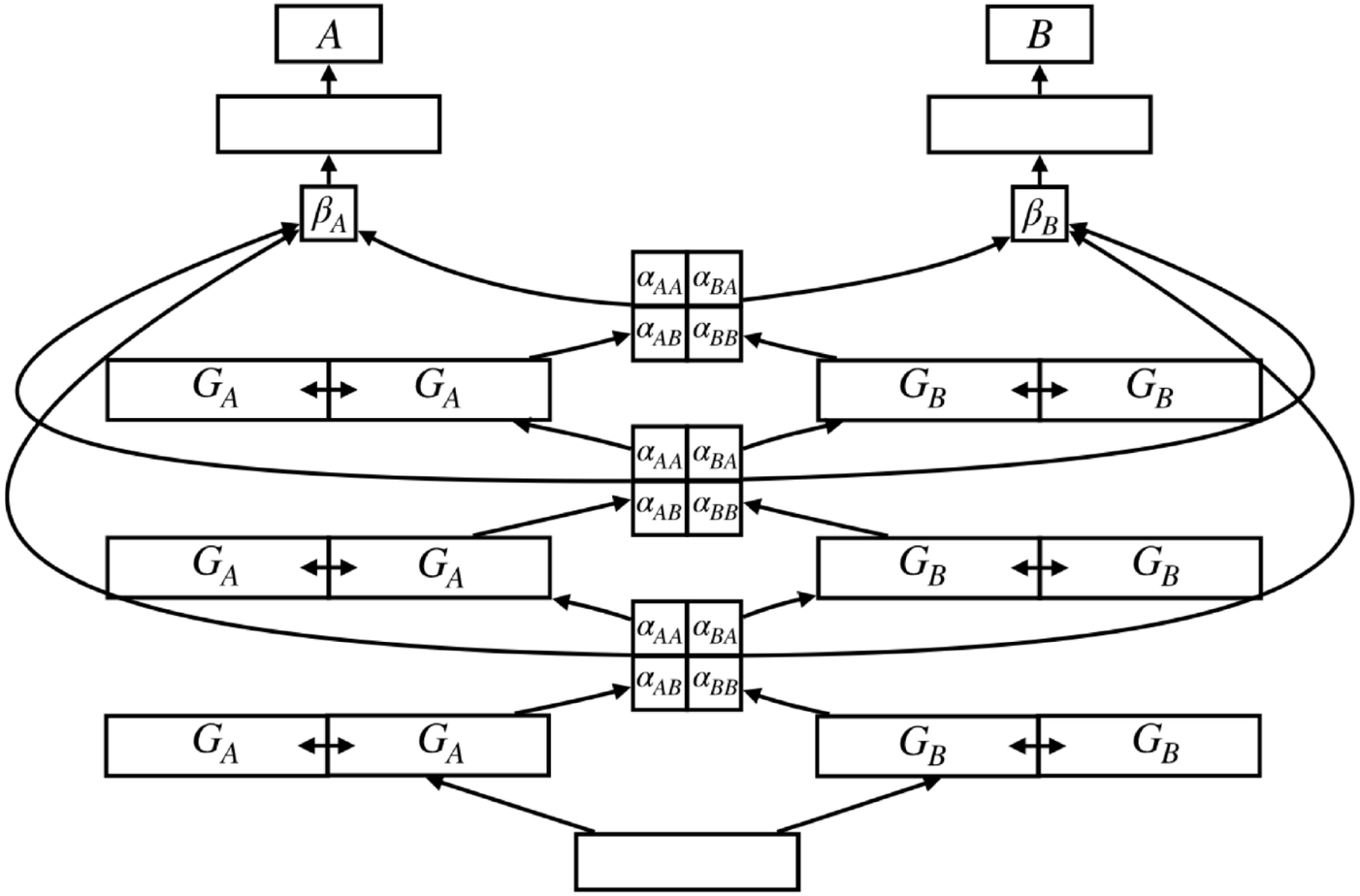
Approach Network Architecture



HIERARCHICAL SHARING

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

Approach Network Architecture

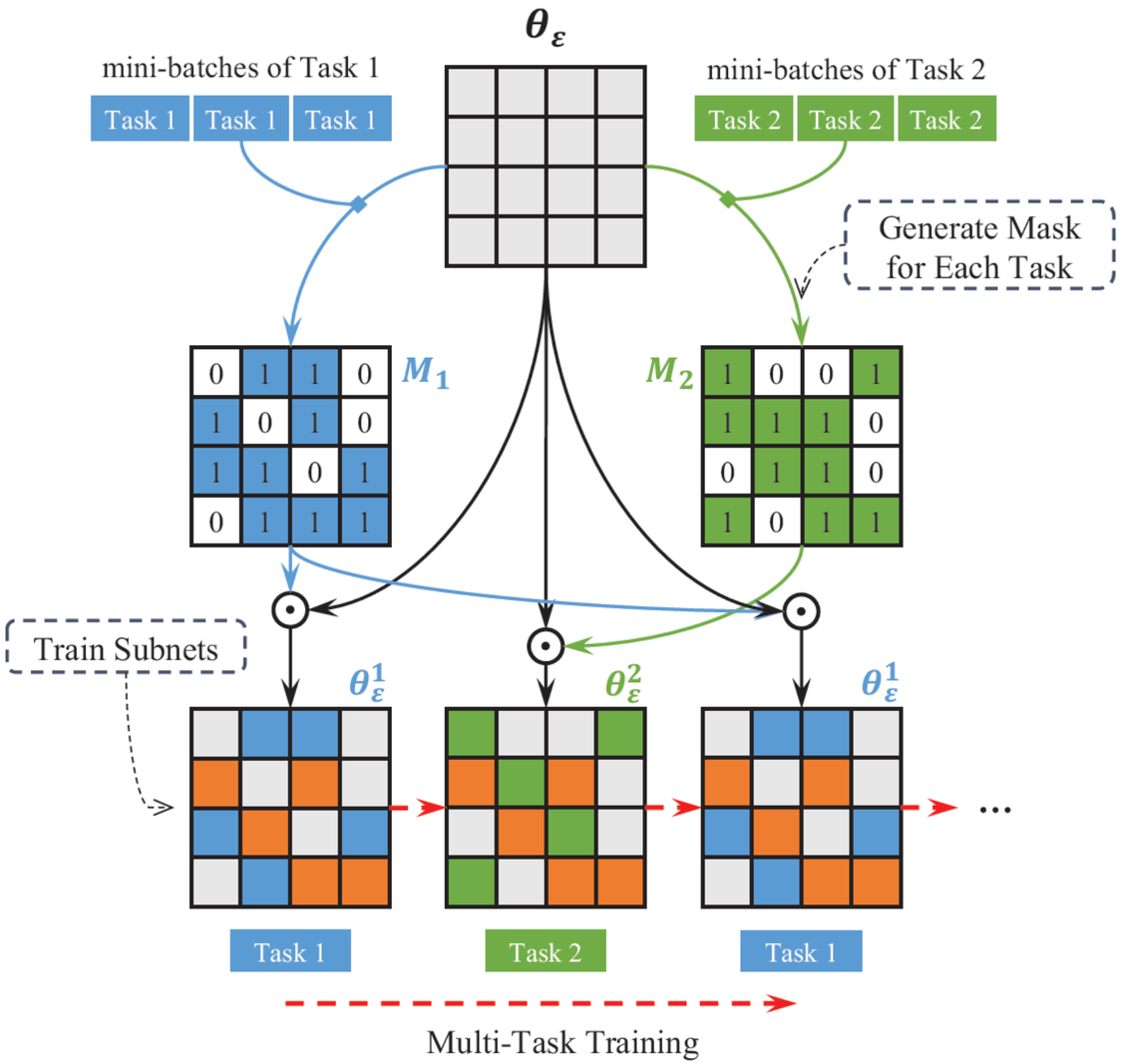


SOFT SHARING

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

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

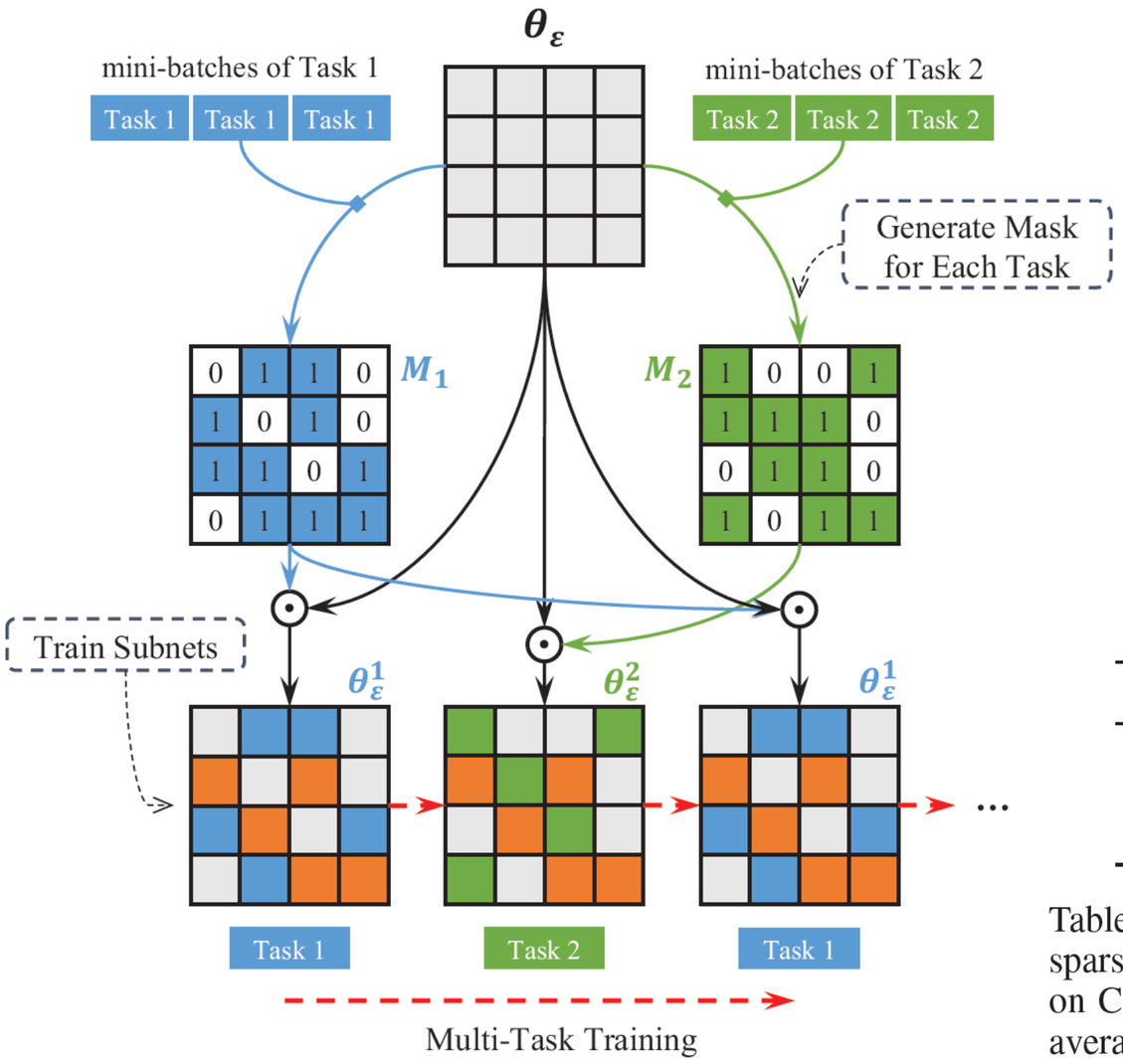
Approach Network Architecture



SOFT SHARING (SPARSE SHARING)

Sun et al. (2020) train an over parametrized network with masks per subtask, on POS tagging, NER, chunking with a CNN-LSTM.
Can also model hierarchical sharing and hard sharing.

Approach Network Architecture



SOFT SHARING (SPARSE SHARING)

Sun et al. (2020) train an over parametrized network with masks per subtask, on POS tagging, NER, chunking with a CNN-LSTM.
Can also model hierarchical sharing and hard sharing.

Task Pairs	Mask OR	$\Delta(S^2 - HS)$
POS & NER	0.18	0.4
NER & Chunking	0.20	0.34
POS & Chunking	0.50	0.05

Table 7: Mask Overlap Ratio (OR) and the improvement for sparse sharing (S^2) compared to hard sharing (HS) of tasks on CoNLL-2003. The improvement is calculated using the average performance on the test set.

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 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.

4 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

5 CONSECUTIVE TRAINING (HASHIMOTO ET AL., 2017)

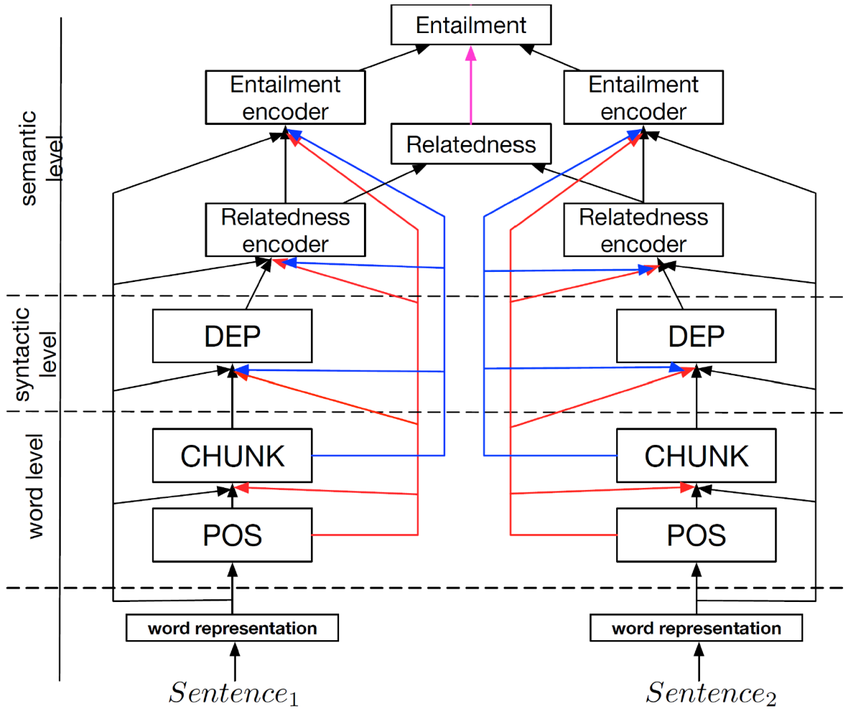
In one epoch, iterate over the datasets in order of complexity;

Introduce successive regularisation to avoid catastrophic forgetting.

task objective

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

task weight decay successive regularisation



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 to force tasks to learn at a similar pace -
e.g. GradNorm by Chen et al. (2018), or Dynamic Weight Average by Liu et al. (2019).

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 to force tasks to learn at a similar pace -
e.g. GradNorm by Chen et al. (2018), or Dynamic Weight Average by Liu et al. (2019).

3 PROGRESS-SIGNAL BASED CURRICULUM

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

Tasks to combine

- STUDY 1

  "Identifying beneficial task relations for multi-task learning in deep neural networks" by Bingel and Søgaard (2017)

- STUDY 2

  "Multi-task learning of pairwise sequence classification tasks over disparate label spaces" by Augenstein et al. (2018)

- STUDY 3

  "Estimating the influence of auxiliary tasks for multi-task learning of sequence tagging tasks." by Schröder and Biemann (2020)

Task Relations Examples

- "Multitask Learning for Complaint Identification and Sentiment Analysis" (Singh et al., 2021)
- "Multitask Learning of Negation and Speculation using Transformers" (Khandelwal, 2020)
- "The Pragmatics behind Politics: Modelling Metaphor, Framing and Emotion in Political Discourse" (Cabot et al., 2020)
- "Multi-Task Learning for Metaphor Detection with Graph Convolutional Neural Networks and Word Sense Disambiguation" (Le et al., 2020)
- "Happy Are Those Who Grade without Seeing: A Multi-Task Learning Approach to Grade Essays Using Gaze Behaviour" (Mathias et al., 2020)

Task Relations Study (1)

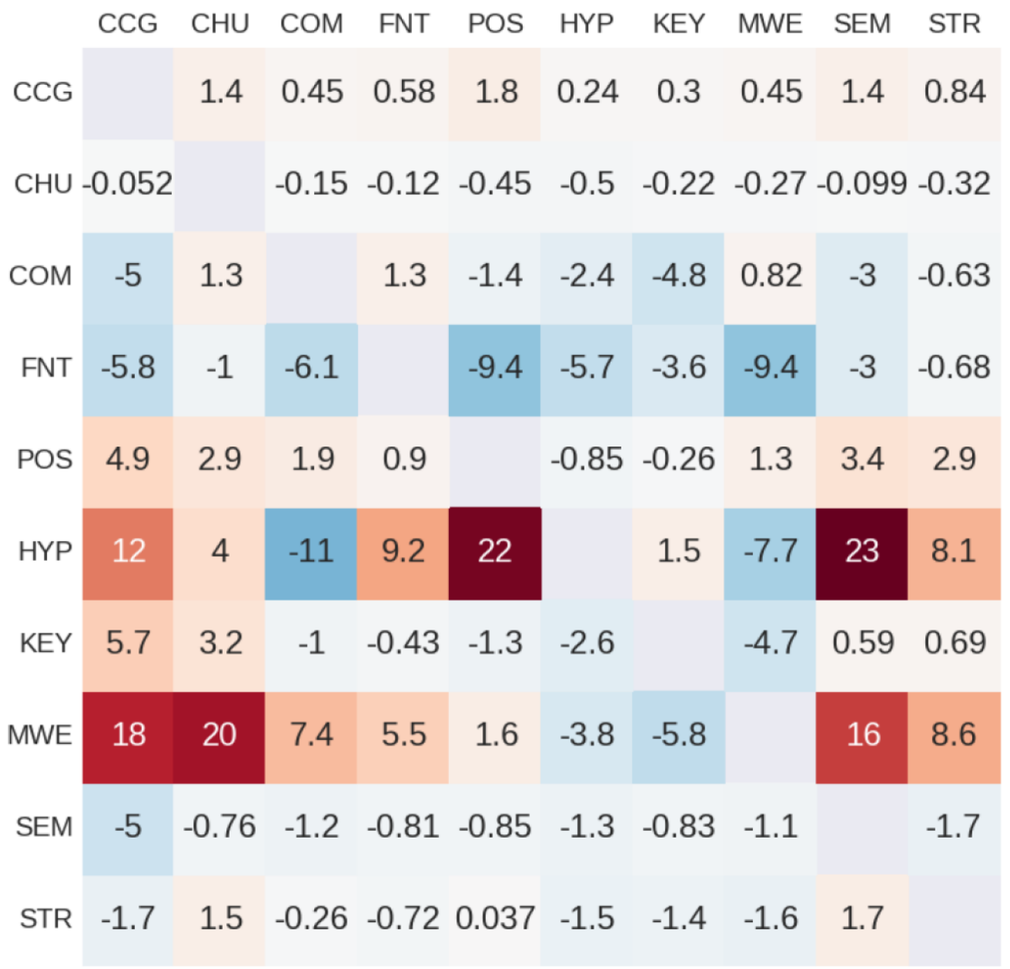
Bingel and Søgaard (2017) research when and why MTL works for task pairs:

- 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)



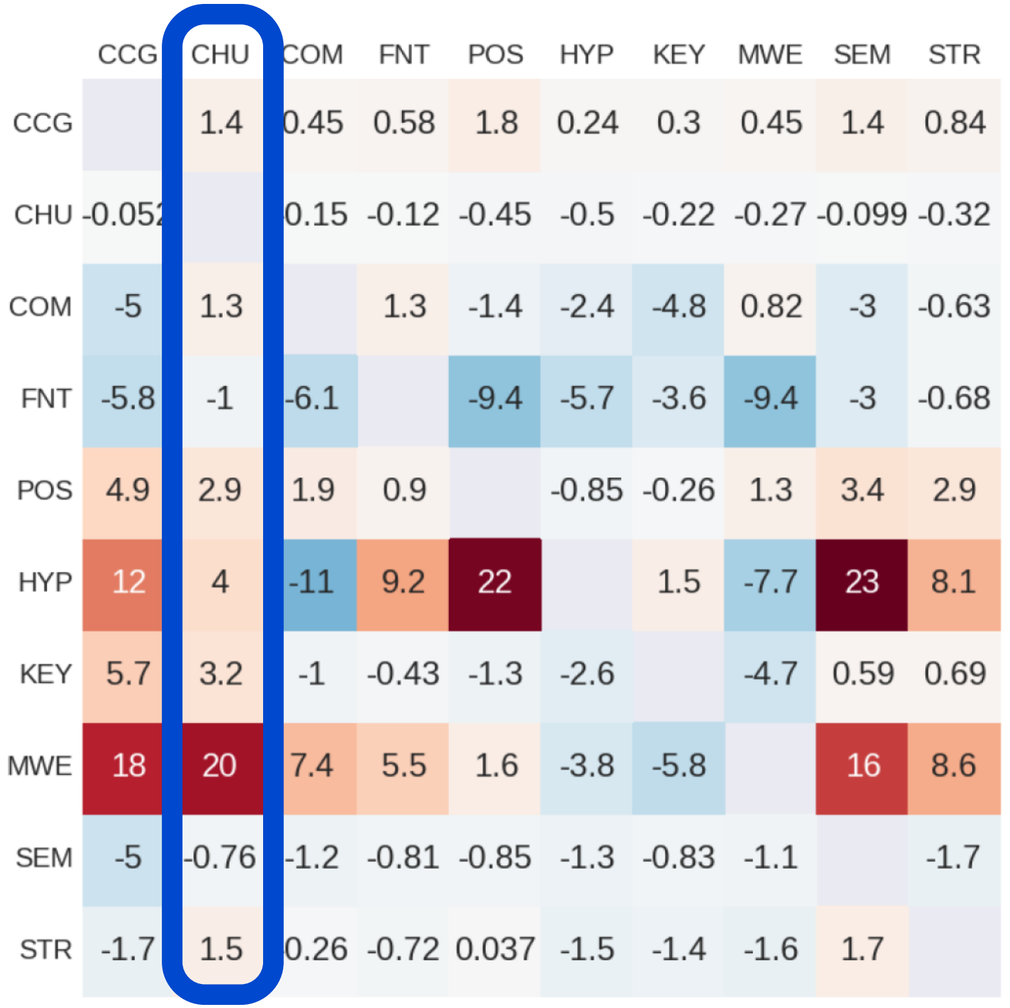
	CCG	CHU	COM	FNT	POS	HYP	KEY	MWE	SEM	STR
CCG		1.4	0.45	0.58	1.8	0.24	0.3	0.45	1.4	0.84
CHU	-0.052		-0.15	-0.12	-0.45	-0.5	-0.22	-0.27	-0.099	-0.32
COM	-5	1.3		1.3	-1.4	-2.4	-4.8	0.82	-3	-0.63
FNT	-5.8	-1	-6.1		-9.4	-5.7	-3.6	-9.4	-3	-0.68
POS	4.9	2.9	1.9	0.9		-0.85	-0.26	1.3	3.4	2.9
HYP	12	4	-11	9.2	22		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)



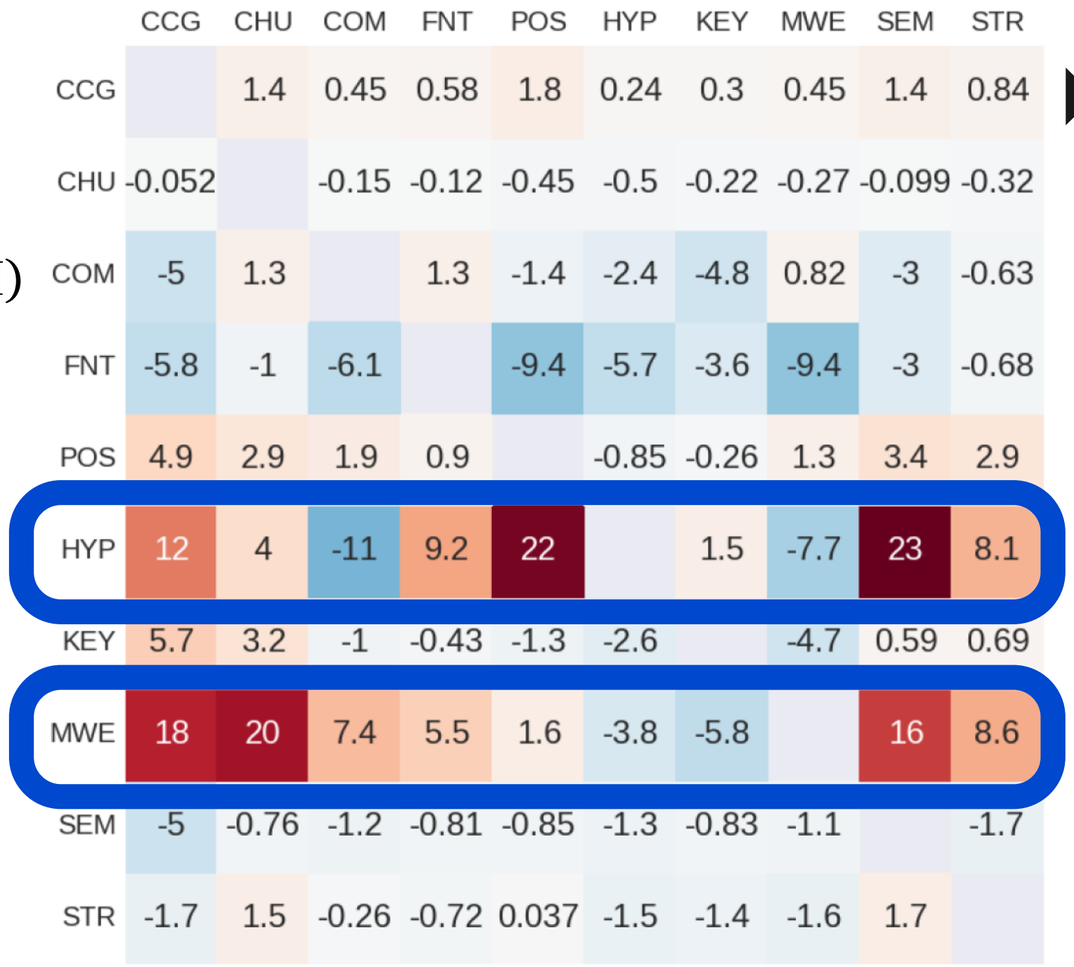
	CCG	CHU	COM	FNT	POS	HYP	KEY	MWE	SEM	STR
CCG		1.4	0.45	0.58	1.8	0.24	0.3	0.45	1.4	0.84
CHU	-0.052		0.15	-0.12	-0.45	-0.5	-0.22	-0.27	-0.099	-0.32
COM	-5	1.3		1.3	-1.4	-2.4	-4.8	0.82	-3	-0.63
FNT	-5.8	-1	-6.1		-9.4	-5.7	-3.6	-9.4	-3	-0.68
POS	4.9	2.9	1.9	0.9		-0.85	-0.26	1.3	3.4	2.9
HYP	12	4	-11	9.2	22		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)



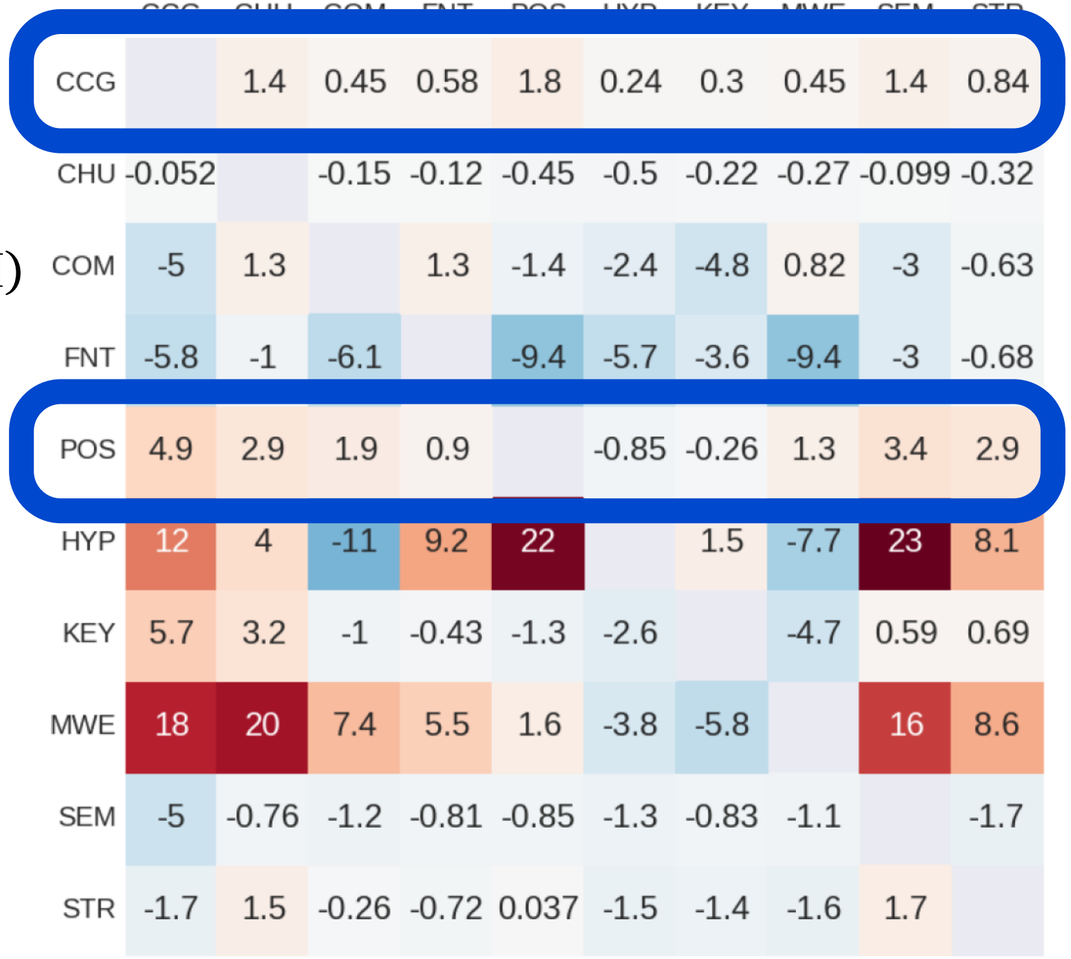
	CCG	CHU	COM	FNT	POS	HYP	KEY	MWE	SEM	STR
CCG		1.4	0.45	0.58	1.8	0.24	0.3	0.45	1.4	0.84
CHU	-0.052		-0.15	-0.12	-0.45	-0.5	-0.22	-0.27	-0.099	-0.32
COM	-5	1.3		1.3	-1.4	-2.4	-4.8	0.82	-3	-0.63
FNT	-5.8	-1	-6.1		-9.4	-5.7	-3.6	-9.4	-3	-0.68
POS	4.9	2.9	1.9	0.9		-0.85	-0.26	1.3	3.4	2.9
HYP	12	4	-11	9.2	22		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 relation:

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)

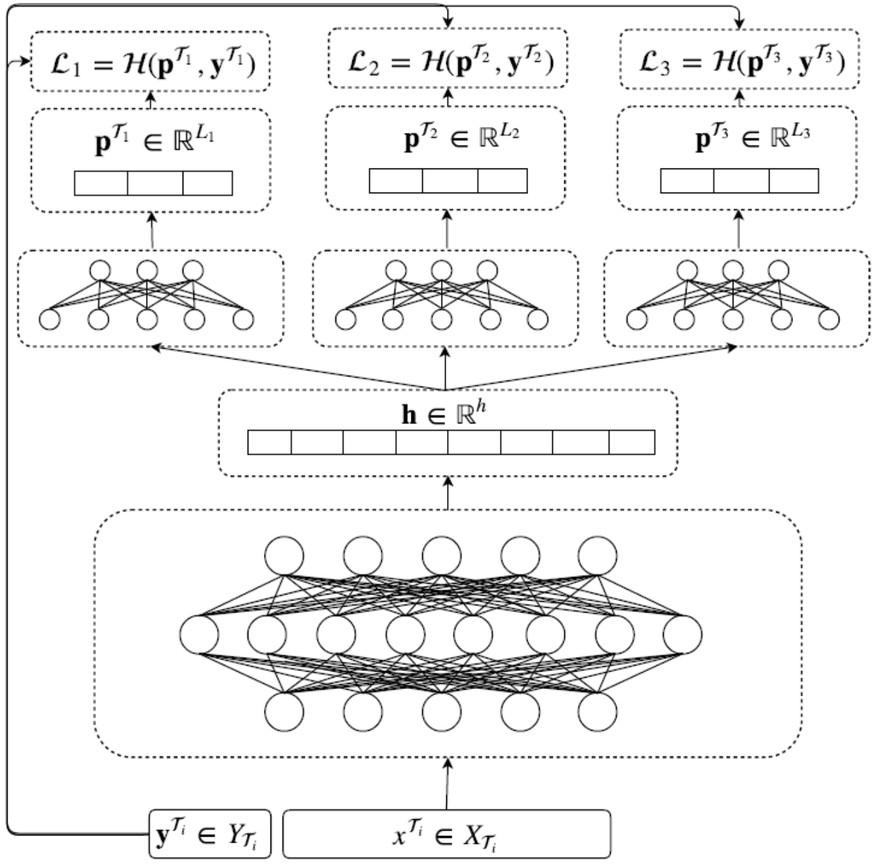
Using logistic regression, they predict MTL gains and losses from dataset statistics (e.g. size or label distribution entropy) and STL model (e.g. loss curve values).

- GOOD PREDICTORS
  - loss plateau
  - label entropy auxiliary task
  - out of vocabulary rate

- BAD PREDICTORS
  - Jensen Shannon Divergence of train/test bag-of-words
  - Dataset size differences

Task Relations Study (2)

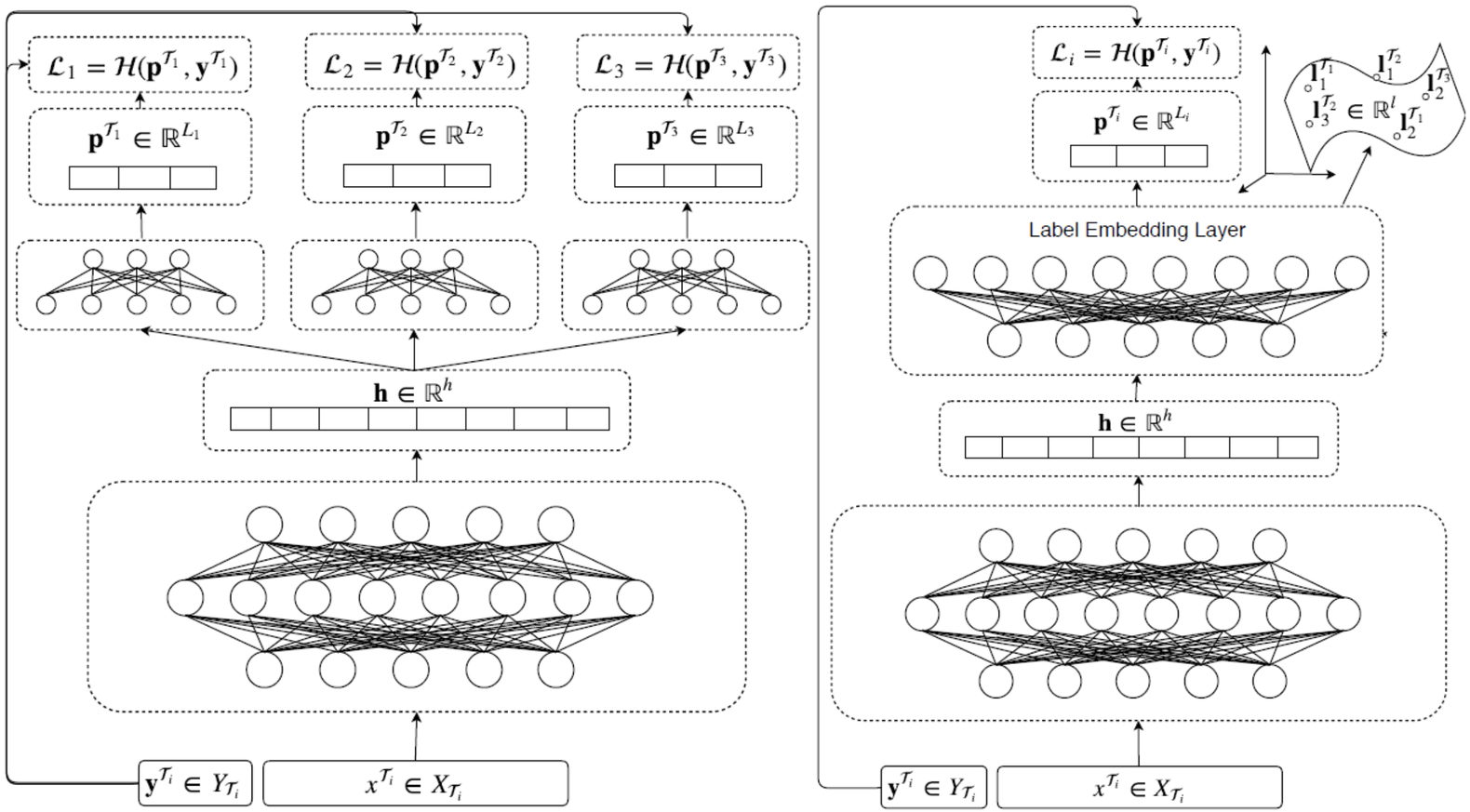
Augenstein et al. (2018) introduce a Label Embedding Space and Label Transfer Network.



(a) Multi-task learning

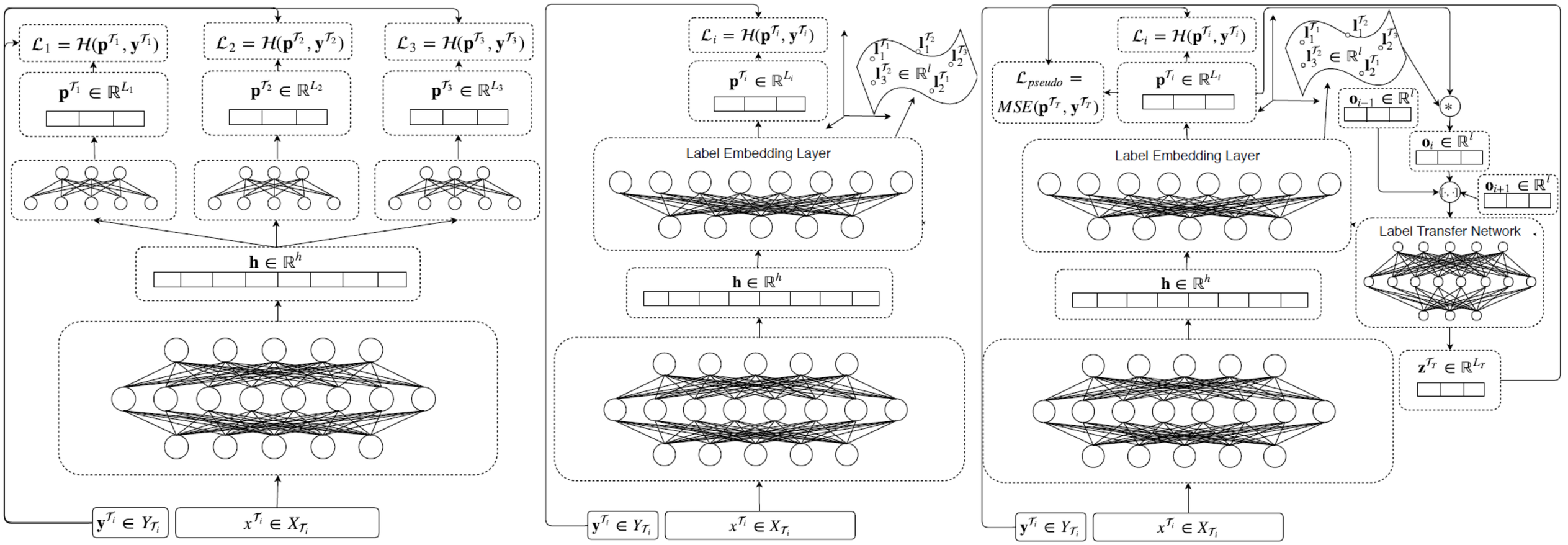
Task Relations Study (2)

Augenstein et al. (2018) introduce a Label Embedding Space and Label Transfer Network.



Task Relations Study (2)

Augenstein et al. (2018) introduce a Label Embedding Space and Label Transfer Network.



(a) Multi-task learning (b) MTL with LEL (c) Semi-supervised MTL with LTN

Task Relations Study (2)

Augenstein et al. (2018) introduce a Label Embedding Space and Label Transfer Network.

- Positive sentiment similar across tasks
  (e.g. sentiment analysis, stance detection, fake news detection);
- 2 clusters for negative and neutral;

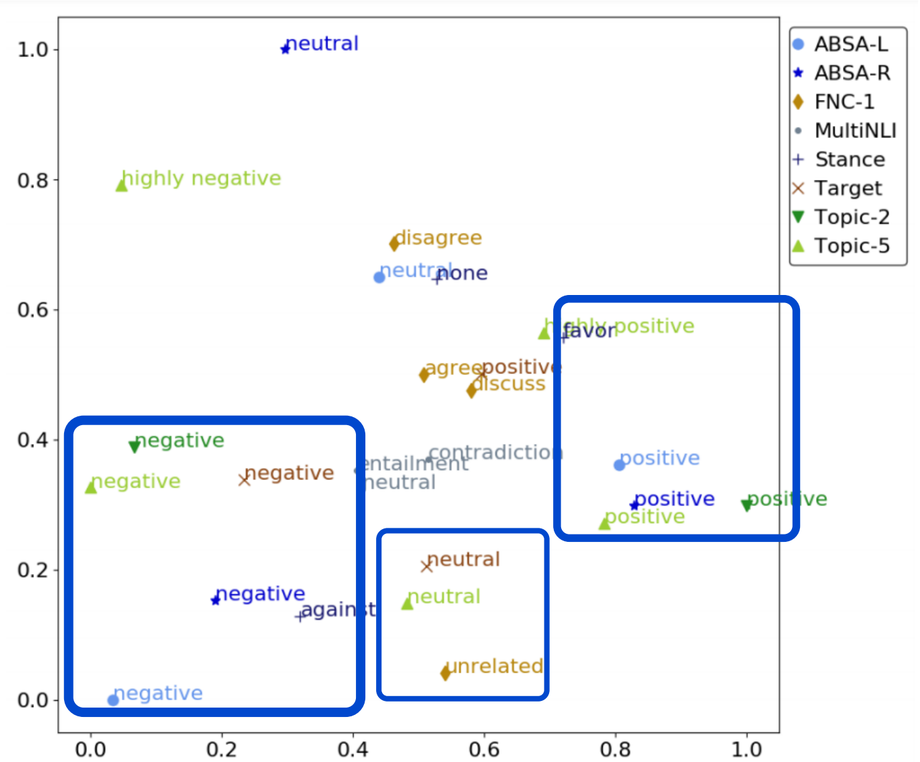


Figure 2: Label embeddings of all tasks. Positive, negative, and neutral labels are clustered together.

Task Relations Study (2)

Augenstein et al. (2018) introduce a Label Embedding Space and Label Transfer Network.

- Positive sentiment similar across tasks
  (e.g. sentiment analysis, stance detection, fake news detection);
- 2 clusters for negative and neutral;
- MNLI labels do not fit in.

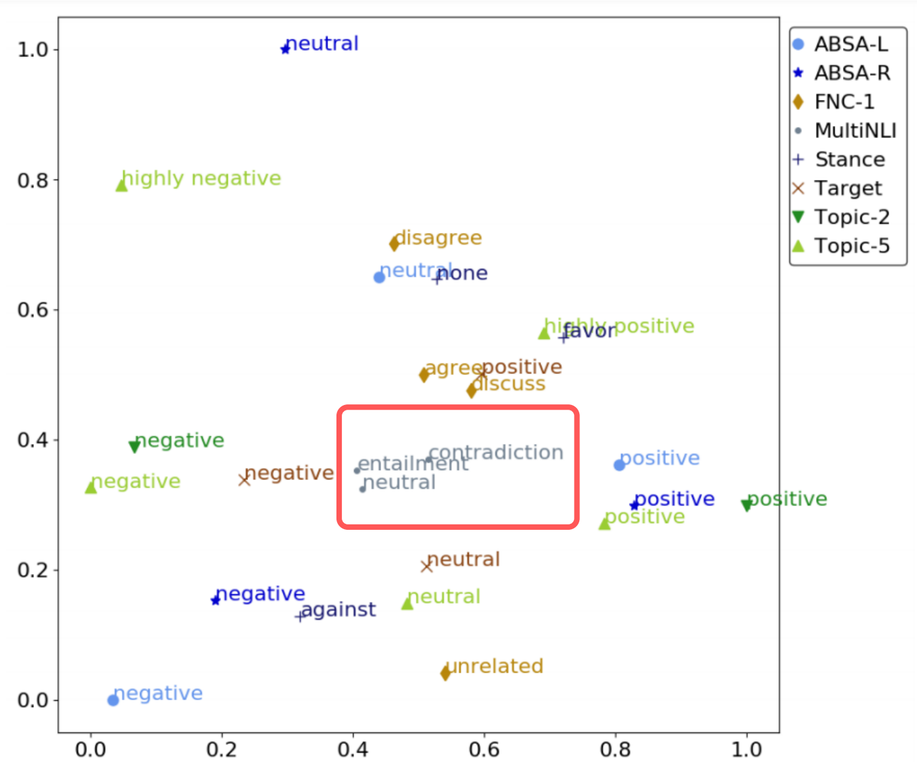


Figure 2: Label embeddings of all tasks. Positive, negative, and neutral labels are clustered together.

Task Relations Study (3)

Schröder & Biemann (2020) measure impact of the similarity of main and auxiliary data.

- Text overlap as normalised mutual information of word counts, corrected for shared vocabulary size;

	l'_1	l'_2	...	l'_M	Σ
l_1	c_{11}	c_{12}	...	c_{1M}	$c_{1.}$
l_2	c_{21}	c_{22}	...	c_{2M}	$c_{2.}$
...	...	...		...	...
l_N	c_{N1}	c_{N2}	...	c_{NM}	$c_{N.}$
Σ	$c_{.1}$	$c_{.2}$	...	$c_{.M}$	c

Table 1: Contingency table for a comparison of label sets L and L' with N and M unique labels

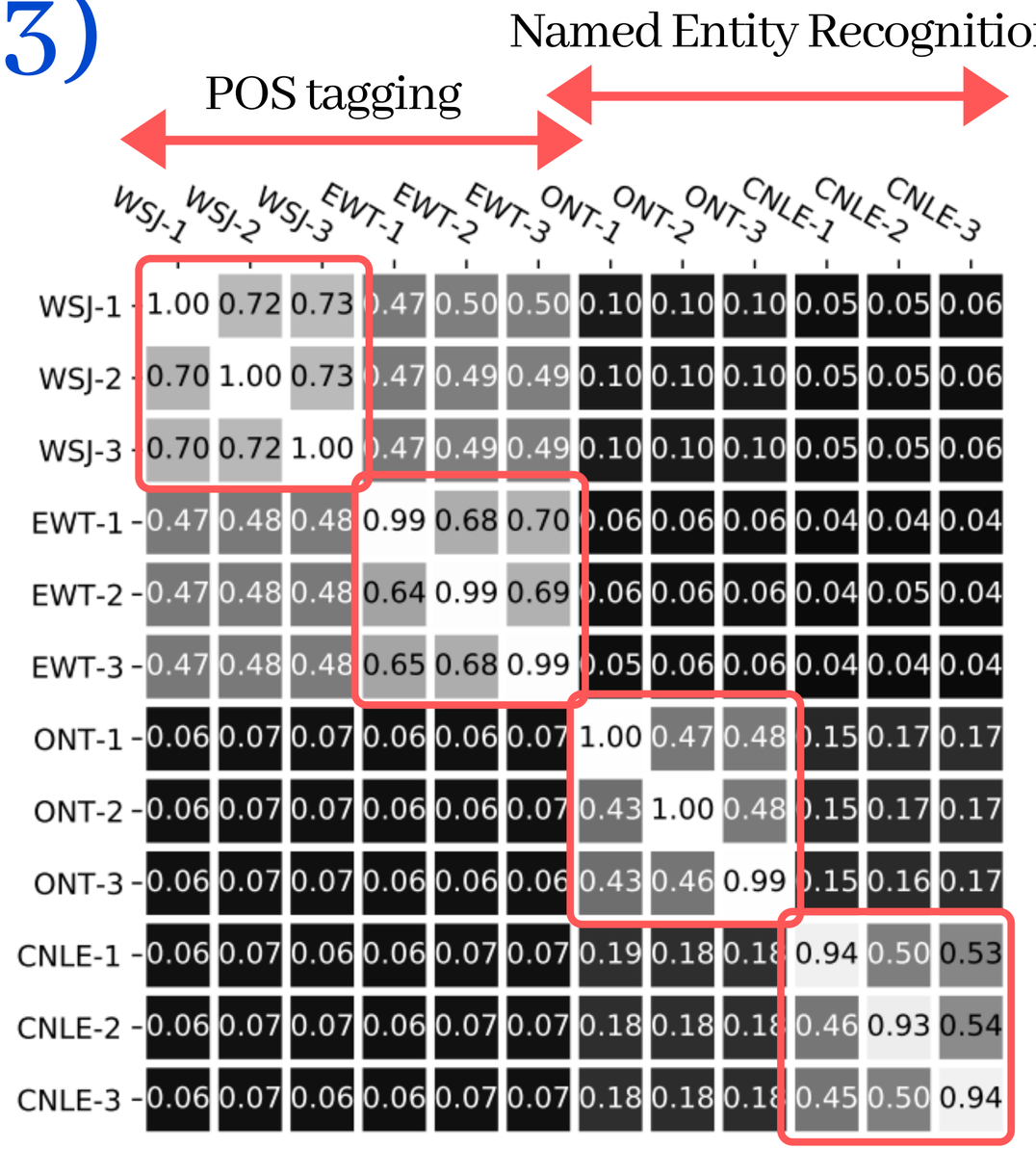
$$NMI(L, L')_{joint} = \frac{I(L; L')}{H(L, L')} = \frac{\sum_{i=1}^{N} \sum_{j=1}^{M} \frac{c_{ij}}{c} \log_2 \left(\frac{c_{ij} c}{c_{i.} c_{.j}} \right)}{-\sum_{i=1}^{N} \sum_{j=1}^{M} \frac{c_{ij}}{c} \log_2 \left(\frac{c_{ij}}{c} \right)}.$$

$$TO = \frac{2 \cdot NMI \cdot SV}{NMI + SV}$$

Task Relations Study (3)

Schröder & Biemann (2020) measure impact of the similarity of main and auxiliary data.

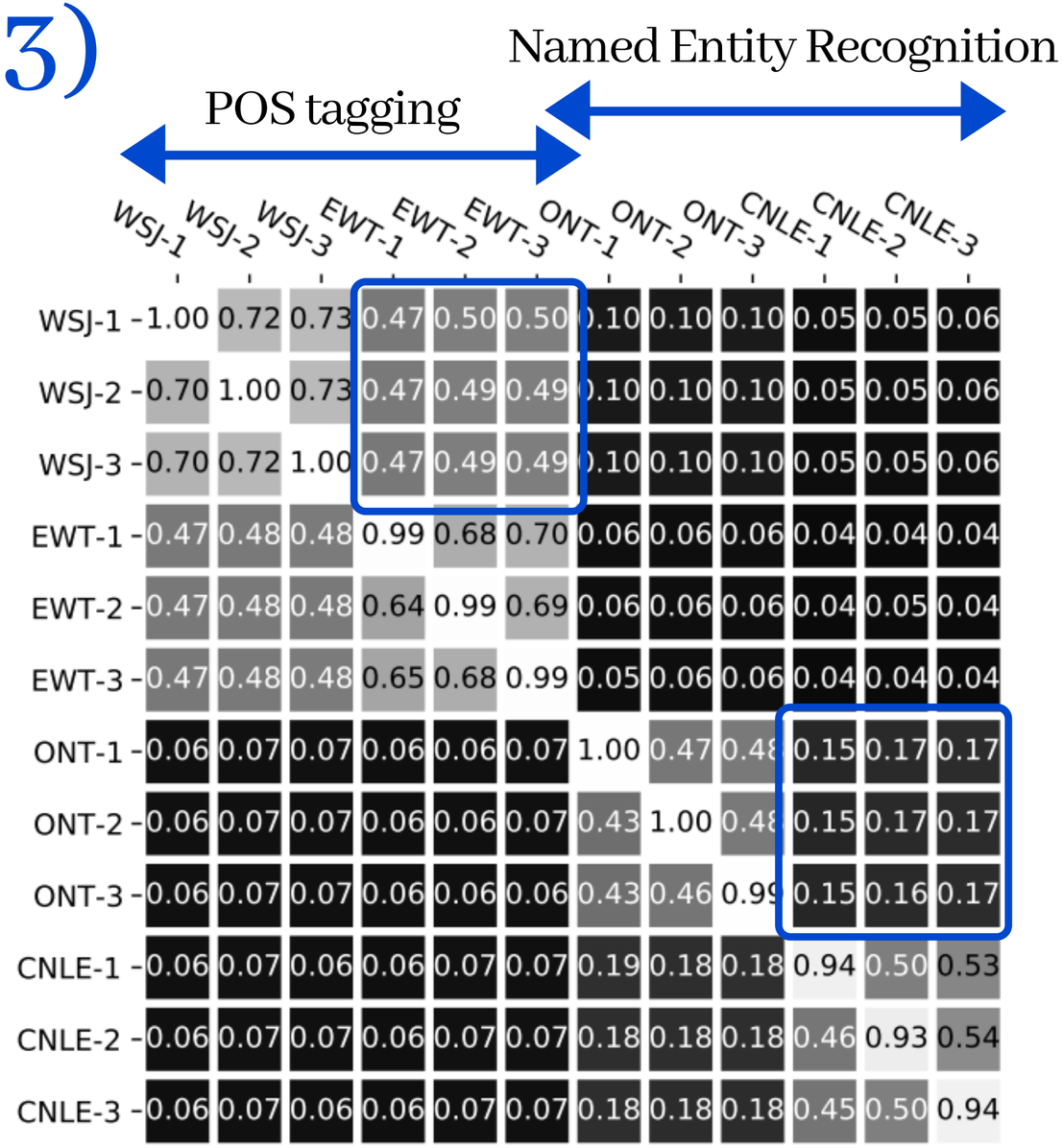
- Compare data subsets for POS tagging and NER;
- NMI is highest within the datasets;



Task Relations Study (3)

Schröder & Biemann (2020) measure impact of the similarity of main and auxiliary data.

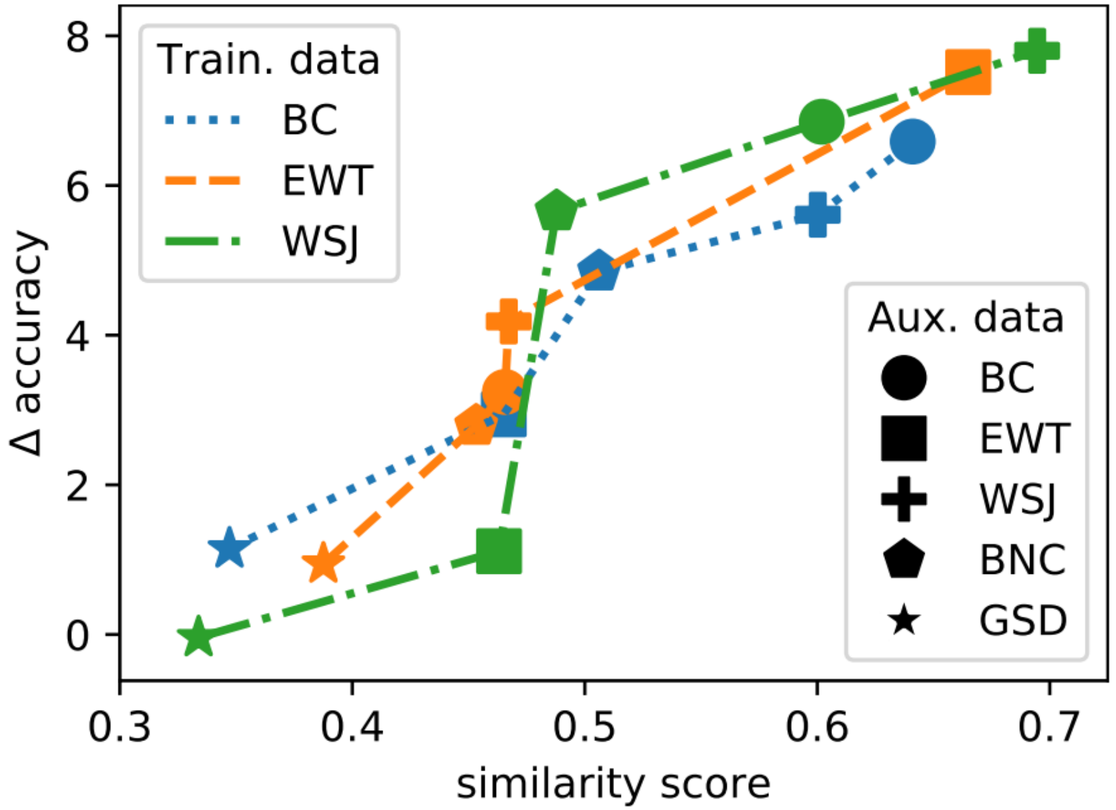
- Compare data subsets for POS tagging and NER;
- NMI is highest within the datasets;
- Find datasets per task more similar to each other than across tasks.



Task Relations Study (3)

Schröder & Biemann (2020) measure impact of the similarity of main and auxiliary data.

- Compare performance gain to similarity;
- Main task of POS tagging vs. auxiliary task of POS tagging;
- More similarity, more gain.



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

References

Augenstein, I., Ruder, S., & Søgaard, A. (2018, January). Multi-Task Learning of Pairwise Sequence Classification Tasks over Disparate Label Spaces. In NAACL-HLT.

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)

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. arXiv preprint arXiv:2005.14165.

Caruana, R. A. (1993). Multitask Learning: A Knowledge-Based Source of Inductive Bias. In Machine Learning Proceedings 1993: Proceedings of the Tenth International Conference on Machine Learning, University of Massachusetts, Amherst, June 27-29, 1993 (p. 41). Morgan Kaufmann.

Cabot, P. L. H., Dankers, V., Abadi, D., Fischer, A., & Shutova, E. (2020, November). The Pragmatics behind Politics: Modelling Metaphor, Framing and Emotion in Political Discourse. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings (pp. 4479-4488).

Chen, Z., Badrinarayanan, V., Lee, C. Y., & Rabinovich, A. (2018, July). Gradnorm: Gradient normalization for adaptive loss balancing in deep multitask networks. In International Conference on Machine Learning (pp. 794-803). PMLR.

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)

References

Hashimoto, K., Xiong, C., Tsuruoka, Y., & Socher, R. (2017, September). 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 (pp. 1923-1933).

Hu, R., & Singh, A. (2021). Transformer is all you need: Multimodal multitask learning with a unified transformer. arXiv preprint arXiv:2102.10772.

Khandelwal, A., & Britto, B. K. (2020, November). Multitask Learning of Negation and Speculation using Transformers. In Proceedings of the 11th International Workshop on Health Text Mining and Information Analysis (pp. 79-87).

Le, D., Thai, M., & Nguyen, T. (2020, April). Multi-Task Learning for Metaphor Detection with Graph Convolutional Neural Networks and Word Sense Disambiguation. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 34, No. 05, pp. 8139-8146).

Liu, S., Johns, E., & Davison, A. J. (2019). End-to-end multi-task learning with attention. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 1871-1880).

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

Mathias, S., Murthy, R., Kanojia, D., Mishra, A., & Bhattacharyya, P. (2020, December). Happy Are Those Who Grade without Seeing: A Multi-Task Learning Approach to Grade Essays Using Gaze Behaviour. In Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing (pp. 858-872).

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

Muangkammuen, P., & Fukumoto, F. (2020, December). Multi-task Learning for Automated Essay Scoring with Sentiment Analysis. In Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing: Student Research Workshop (pp. 116-123).

Ruder, S., Bingel, J., Augenstein, I., & Søgaard, A. (2019, July). Latent multi-task architecture learning. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 33, No. 01, pp. 4822-4829).

References

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.

Schröder, F., & Biemann, C. (2020, July). Estimating the influence of auxiliary tasks for multi-task learning of sequence tagging tasks. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (pp. 2971-2985).

Singh, A., Saha, S., Hasanuzzaman, M., & Dey, K. (2021). Multitask Learning for Complaint Identification and Sentiment Analysis. Cognitive Computation, 1-16.

Sun, T., Shao, Y., Li, X., Liu, P., Yan, H., Qiu, X., & Huang, X. (2020, April). Learning sparse sharing architectures for multiple tasks. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 34, No. 05, pp. 8936-8943).

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)

Zhou, J., Zhang, Z., Zhao, H., & Zhang, S. (2020, November). LIMIT-BERT: Linguistics Informed Multi-Task BERT. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings (pp. 4450-4461).