TURL: experiments, insights and new ideas

Federico Belotti (federico.belotti1@unimib.it)

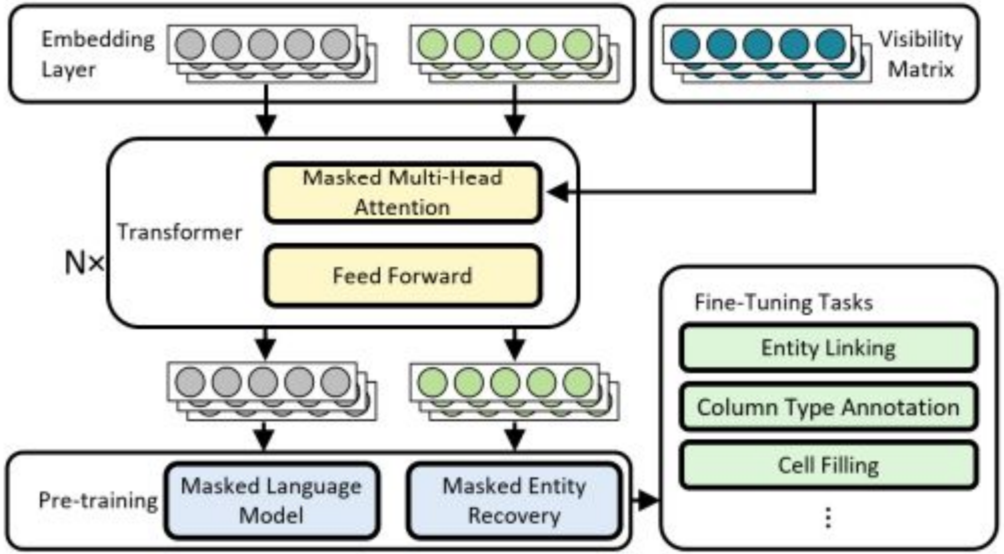
INSID&S Lab
Department of Informatics, Systems and Communication
Università degli Studi di Milano-Bicocca

TURL summary

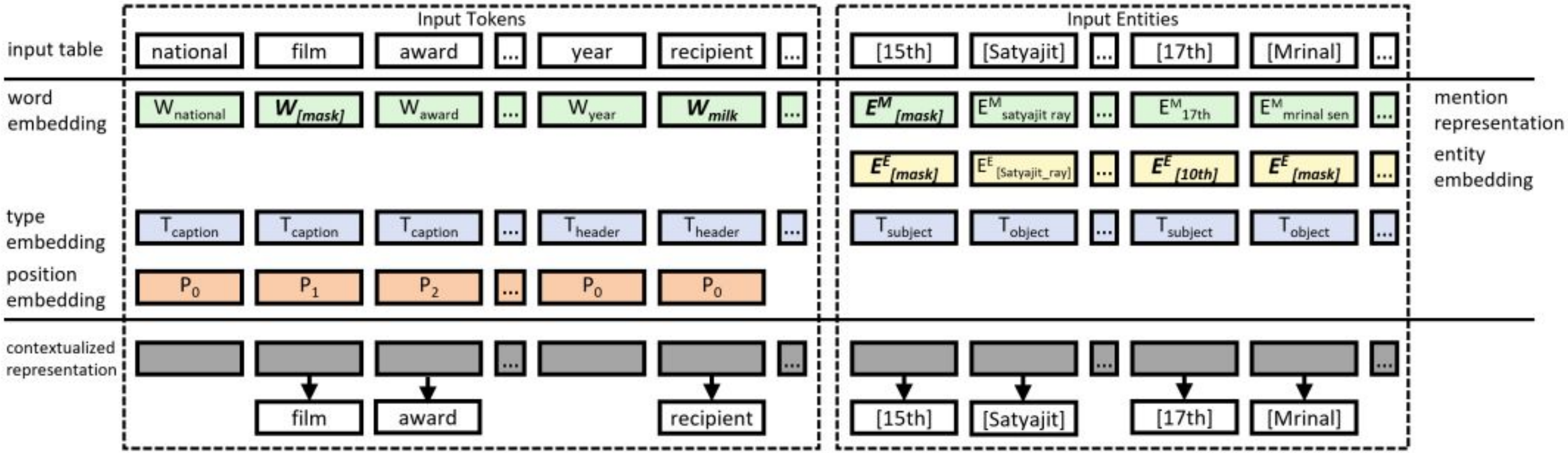
	Task					
	Annotation/Matching	Augmentation	QA	Fact Verification	Dialogue Generation	Data-to-Text
TURL	CEA CTA CPA	Row Population Cell Filling Schema Aug.	N.A.	N.A.	N.A.	N.A.

	Input	Output	Transformer	Params
TURL	Flatten input table as: **[Table context, Table Header-1, …, Table Header-M, Row-1, …, Row-N]**	<u>Depends on the downstream task</u>: for CEA the probability distribution over the N candidates for a mention	Tiny-BERT (Encoder-only)	14.5M

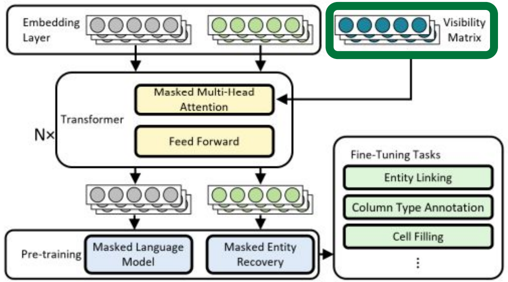
TURL: Architecture overview



TURL: Pre-training example



TURL: Structure-aware Transformer

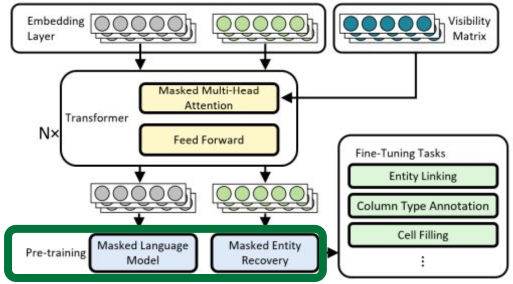


Standard BERT Transformer with modified attention matrix:

- Caption and topic entity are visible to all components of the table
- Entities and text content in the same row or the same column are visible to each other

Figure 5: Graphical illustration of masked self-attention by our visibility matrix. Each token/entity in a table can only attend to its directly connected neighbors (shown as edges here).

TURL: Pre-training objectives



- **MLM**: Given an input token sequence from [C;H], mask some percentage of the tokens at random and then predict masked tokens
- **MER**: Mask a certain percentage of input entity cells and then recover the linked entity based on surrounding entity cells and table metadata.
  - For some percentage of masked entities mask $\mathbf{e}^{e}$ only, such that the model receives additional entity mention information (from $\mathbf{e}^{m}$)

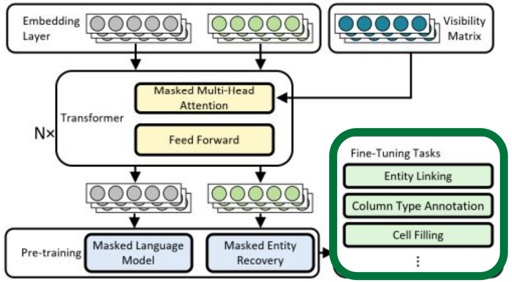
$$P(w) = \frac{\exp\left(\mathrm{LINEAR}(\mathbf{h}^{t}) \cdot \mathbf{w}\right)}{\sum_{w_k \in \mathcal{W}} \exp\left(\mathrm{LINEAR}(\mathbf{h}^{t}) \cdot \mathbf{w}_k\right)}$$

$$P(e) = \frac{\exp\left(\mathrm{LINEAR}(\mathbf{h}^{e}) \cdot \mathbf{e}^{e}\right)}{\sum_{e_k \in \mathcal{E}} \exp\left(\mathrm{LINEAR}(\mathbf{h}^{e}) \cdot \mathbf{e}^{e}_{k}\right)}$$

$$loss = \sum \log\left(P(w)\right) + \sum \log\left(P(e)\right)$$

Small candidate set

TURL: CEA fine-tuning



Given a fine-tuning EL dataset:

1. Each cell is a potential entity and input both entity mention and table metadata to TURL and obtain a contextualized representation $\mathbf{h}^{\mathbf{e}}$
2. Represent a Knowledge-Base Entity as $\mathbf{e}^{\mathbf{kb}} = [\mathrm{MEAN}_{w \in N}(\mathbf{w}), \mathrm{MEAN}_{w \in D}(\mathbf{w}), \mathrm{MEAN}_{t \in T}(\mathbf{t})]$
    - **w** initialized from the pre-trained word embedding
    - **t** is the entity-type embedding to be learned
3. Pre-trained entity embeddings are not used
    - goal is to link mentions to entities in a target KB, not necessarily those appear in the pre-training table corpus
4. The model is fine-tuned with a cross-entropy loss

Activities

1. Pre-training TURL in a distributed setting with SOTA methodologies
    - 4x80GB A100
    - Mixed-precision training (16bit)
    - Distributed Data Parallel (DDP)
    - Learning rate scaling (Accurate, Large Minibatch SGD: Training ImageNet in 1 Hour)
2. Pre-training TURL in a distributed setting with deduplicated data
3. Fine-tuning TURL on downstream tasks
    - Replicate authors results (CEA, CTA)
    - Test on SemTab challenge datasets in zero-shot (CEA)

Comparison with TURL

Task	Test dataset	Method	F1	Prec.	Rec.
CTA	TURL	TURL	0.9475	0.9495	0.9456
		$TURL_{our}$	0.9441	0.9410	0.9473
		$TURL_{our\text{-}dedup}$	0.9433	0.9383	0.9484
CEA	TURL	TURL	0.6800	0.7100	0.6600
		$TURL_{our}$	0.6691	0.6952	0.6450
		$TURL_{our\text{-}dedup}$	0.6686	0.6946	0.6445
CEA	TURL-T2D	TURL	0.8200	0.8800	0.7700
		$TURL_{our}$	0.8152	0.8728	0.7648
		$TURL_{our\text{-}dedup}$	0.8153	0.8728	0.7648

SemTab CEA datasets

Datasets from [Estimating Link Confidence for Human-in-the-Loop Table Annotation]

dataset	# tables	# columns	# rows	# entities (CEA)
Round1_T2D	64	323	9089	8078
Round3	2161	9736	152753	390456
Round4	22207	78750	475897	994920
2T-2020	180	802	194438	667243
HardTableR2	1750	5589	29280	47439
HardTableR3	7207	17902	58949	58948

SemTab CEA experimental setting

The tests are conducted as follows:

- For every mention in the ground-truth we collect
  - a list of at most 50 candidates with their relative description and types (from wikidata lookup)
- We then test in two settings:
  - we evaluate TURL only on those mentions that have at least one candidate and for which the list of candidates contains the ground-truth entity to be linked
  - we evaluate TURL in an Oracle-Retriever setting, i.e. a setting where we specifically inject the ground-truth entity in the first position of the candidate list, i.e. we suppose that the ground-truth entity can always be retrieved and is the one most relevant to the mention itself

SemTab CEA experimental setting

- TURL has been fine-tuned on the authors CEA dataset:
  - “The training set for fine-tuning TURL is based on our pre-training corpus, but with tables in the WikiGS removed. We also remove duplicate entity mentions and mentions where Wikidata Lookup fails to return the ground truth entity in candidates, and finally obtain 1,264,217 entity mentions in 192,728 tables to fine-tune our model for the entity linking task”

- In addition to the above, we have also tried to finetune TURL on both the TURL fine-tuning dataset and a SemTab challenge dataset

SemTab CEA experimental setting

We collect the F1, Precision and Recall as defined in the [SemTab 2023]()

$$Precision = \frac{correct\_annotations\#}{submitted\_annotations\#}$$

$$Recall = \frac{correct\_annotations\#}{ground\_truth\_annotations\#}$$

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

The entity to be linked is selected as in TURL, i.e.

max(0.8 * TURL-Top1-score, Wikidata-Top1-score)

where **{Method}-Top1-score** is the score computed by TURL for the best candidate retrieved by {Method}

SemTab CEA - T2D Round1

Method	F1	Prec.	Rec.
$TURL_{our}$	0.7968	0.8697	0.7352
$TURL_{our\text{-}dedup}$	0.8126	0.8869	0.7498
$TURL_{our\text{-}oracle\text{-}retriever}$	0.8815	/	/
$TURL_{our\text{-}dedup\text{-}oracle\text{-}retriever}$	0.8970	/	/
Alligator	0.8600	/	/
SemTab Top Scorer	0.9000	/	/

SemTab CEA - T2D Round1

Method	F1 / F1 (TURL + R3)	Prec. / Prec. (TURL + R3)	Rec. / Rec. (TURL + R3)
TURL_{our}	0.7968 / 0.8197	0.8697 / 0.8947	0.7352 / 0.7563
$\text{TURL}_{\text{our-dedup}}$	0.8126	0.8869	0.7498
$\text{TURL}_{\text{our-oracle-retriever}}$	0.8815 / 0.9071	/	/
$\text{TURL}_{\text{our-dedup-oracle-retriever}}$	0.8970	/	/
Alligator	0.8600	/	/
SemTab Top Scorer	0.9000	/	/

SemTab CEA - HardTables Round2

Method	F1	Prec.	Rec.
TURL_{our}	0.7048	0.7792	0.6433
$\text{TURL}_{\text{our-dedup}}$	0.6897	0.7625	0.6296
$\text{TURL}_{\text{our-oracle-retriever}}$	0.8124	/	/
$\text{TURL}_{\text{our-dedup-oracle-retriever}}$	0.8020	/	/
Alligator	0.9300	/	/
SemTab Top Scorer	0.9800	/	/

SemTab CEA - HardTables Round2

Method	F1 / F1 (TURL + R3)	Prec.	Rec.
$TURL_{our}$	0.7048	0.7792	0.6433
$TURL_{our\text{-}dedup}$	0.6897	0.7625	0.6296
$TURL_{our\text{-}oracle\text{-}retriever}$	0.8124 / 0.8740	/	/
$TURL_{our\text{-}dedup\text{-}oracle\text{-}retriever}$	0.8020	/	/
Alligator	0.9300	/	/
SemTab Top Scorer	0.9800	/	/

SemTab CEA - HardTables Round3

Method	F1	Prec.	Rec.
$TURL_{our}$	0.3419	0.4086	0.2940
$TURL_{our\text{-}dedup}$	0.3000	0.3578	0.2574
$TURL_{our\text{-}oracle\text{-}retriever}$	0.5550	/	/
$TURL_{our\text{-}dedup\text{-}oracle\text{-}retriever}$	0.5177	/	/
Alligator	0.6200	/	/
SemTab Top Scorer	0.9700	/	/

SemTab CEA - 2T Round4 (50 mentions per table)

Method	F1	Prec.	Rec.
TURL_{our}	0.3071	0.3860	0.2550
$\text{TURL}_{\text{our-dedup}}$	0.1898	0.2386	0.1576
$\text{TURL}_{\text{our-oracle-retriever}}$	0.6256	/	/
$\text{TURL}_{\text{our-dedup-oracle-retriever}}$	0.5109	/	/
Alligator	0.8900	/	/
SemTab Top Scorer	0.9000	/	/

SemTab CEA - Round3 2019 (50 mentions per table)

Method	F1	Prec.	Rec.
TURL_{our}	0.6192	0.7907	0.5088
$\text{TURL}_{\text{our-dedup}}$	0.6032	0.7702	0.4957
$\text{TURL}_{\text{our-oracle-retriever}}$	0.8348	/	/
$\text{TURL}_{\text{our-dedup-oracle-retriever}}$	0.8178	/	/
Alligator	0.7600	/	/
SemTab Top Scorer	0.9700	/	/

SemTab CEA - Round4 2020 (50 mentions per table)

Method	F1	Prec.	Rec.
$TURL_{our}$	0.6682	0.7567	0.5983
$TURL_{our\text{-}dedup}$	0.6570	0.7440	0.5883
$TURL_{our\text{-}oracle\text{-}retriever}$	0.8067	/	/
$TURL_{our\text{-}dedup\text{-}oracle\text{-}retriever}$	0.7934	/	/
Alligator	0.9100	/	/
SemTab Top Scorer	0.9900	/	/

SemTab CEA - Round4 2020 (50 mentions per table)

Method	F1 / F1 (TURL + R3)	Prec. / Prec. (TURL + R3)	Rec. / Rec. (TURL + R3)
TURL_{our}	0.6682 / 0.7588	0.7567 / 0.8592	0.5983 / 0.6794
$\text{TURL}_{\text{our-dedup}}$	0.6570	0.7440	0.5883
$\text{TURL}_{\text{our-oracle-retriever}}$	0.8067 / 0.8882	/	/
$\text{TURL}_{\text{our-dedup-oracle-retriever}}$	0.7934	/	/
Alligator	0.9100	/	/
SemTab Top Scorer	0.9900	/	/

TURL: insights

- Table-oriented: TURL processes the entire table, so the batching is done over tables! This could hinder the scalability of the model. Can we change the perspective and transform TURL to be cell-oriented?
- Results on CEA are not fixed: given TURL table-orientation, the candidates set considered for the scoring (i.e. the set of candidates to be scored against a particular mention in a table) is the concatenation of all the candidates of every mention in a table. This has multiple problems:
    - Candidates replication: if two different mentions have their candidates set intersection not empty. What happens to the score?
    - For scalability constraints, specially for large table with a lot of mentions to be linked, it is useful to split a table considering N-mention-per-table, so to have a bounded set of candidates (i.e. if we keep at most 50 candidates per mention and 50-mentions-per-table, then we would have at most 2500 candidates). In this case the score changes for every N, since the softmax has different support for different N values!

TURL insights: No context for CEA

Missing context

```
['name', 'portrait', 'birth', 'marriages', 'death'],
[[[0, 0], 'Sancho I Garcés'],
 [[0, 2], 'García Jiménez'],
 [[0, 3], 'Toda of Navarre'],
 [[1, 0], 'Jimeno Garcés'],
 [[1, 2], 'García Jiménez'],
 [[2, 0], 'García Sánchez I'],
 [[2, 2], 'Sancho I Garcés'],
 [[2, 3], 'Andregota Galíndez'],
 [[3, 0], 'Sancho II Garcés Abarca'],
 [[3, 2], 'García Sánchez I'],
 [[3, 3], 'Urraca Fernández'],
 [[4, 0], 'García Sánchez II'],
 [[4, 2], 'Sancho II Garcés Abarca'],
 [[5, 2], 'García Sánchez II'],
 [[6, 0], 'García Sánchez III'],
 [[6, 4], 'Atapuerca'],
 [[7, 0], 'Sancho IV Garcés'],
 [[7, 2], 'García Sánchez III'],
 [[7, 4], 'Peñalén']]
```

Name	Portrait	Birth	Marriage(s)	Death
Sancho I Garcés 905–925		son of García Jiménez and Dadildis de Pallars	Toda of Navarre 6 children	11 December 925 Resa
Jimeno Garcés 925–931		son of García Jiménez and Dadildis de Pallars	Sancha of Navarre 3 children	29 May 931
García Sánchez I 931–970		919 son of Sancho I Garcés and Toda of Navarre	Andregota Galíndez of Aragón 2 children Teresa Ramírez of León 3 children	22 February 970 aged 51
Sancho II Garcés Abarca 970–994		after 935 son of García Sánchez I and Andregota	Urraca Fernández 4 children	December 994
García Sánchez II 994–1000/04		son of Sancho II Garcés Abarca and Urraca Fernández	Jimena Fernández of Cea 981 4 children	1000/04
Sancho III the Great 1004–1035		985 son of García Sánchez II and Jimena Fernández of Cea	Muniadona of Castile 1010 4 children	18 October 1035
García Sánchez III 1035–1054		1016 son of Sancho III the Great and Muniadona of Castile	Estefanía of Barcelona 1038 9 children	15 September 1054 Atapuerca
Sancho IV Garcés 1054–1076		1039 son of García Sánchez III and Estefanía of Barcelona	Placencia 1068 3 children	4 June 1076 Peñalén

Missing context

```
['name', 'portrait', 'birth', 'marriages', 'death'],
[[[0, 0], 'Sancho I Garcés'],
 [[0, 2], 'García Jiménez'],
 [[0, 3], 'Toda of Navarre'],
 [[1, 0], 'Jimeno Garcés'],
 [[1, 2], 'García Jiménez'],
 [[2, 0], 'García Sánchez I'],
 [[2, 2], 'Sancho I Garcés'],
 [[2, 3], 'Andregota Galíndez'],
 [[3, 0], 'Sancho II Garcés Abarca'],
 [[3, 2], 'García Sánchez I'],
 [[3, 3], 'Urraca Fernández'],
 [[4, 0], 'García Sánchez II'],
 [[4, 2], 'Sancho II Garcés Abarca'],
 [[5, 2], 'García Sánchez II'],
 [[6, 0], 'García Sánchez III'],
 [[6, 4], 'Atapuerca'],
 [[7, 0], 'Sancho IV Garcés'],
 [[7, 2], 'García Sánchez III'],
 [[7, 4], 'Peñalén']]
```

Name	Portrait	Birth	Marriage(s)	Death
Sancho I Garcés 905–925		son of García Jiménez and Dadildis de Pallars	Toda of Navarre 6 children	11 December 925 Resa
Jimeno Garcés 925–931		son of García Jiménez and Dadildis de Pallars	Sancha of Navarre 3 children	29 May 931
García Sánchez I 931–970		919 son of Sancho I Garcés and Toda of Navarre	Andregota Galíndez of Aragón 2 children Teresa Ramírez of León 3 children	22 February 970 aged 51
Sancho II Garcés Abarca 970–994		after 935 son of García Sánchez I and Andregota	Urraca Fernández 4 children	December 994
García Sánchez II 994–1000/04		son of Sancho II Garcés Abarca and Urraca Fernández	Jimena Fernández of Cea 981 4 children	1000/04
Sancho III the Great 1004–1035		985 son of García Sánchez II and Jimena Fernández of Cea	Muniadona of Castile 1010 4 children	18 October 1035
García Sánchez III 1035–1054		1016 son of Sancho III the Great and Muniadona of Castile	Estefanía of Barcelona 1038 9 children	15 September 1054 Atapuerca
Sancho IV Garcés 1054–1076		1039 son of García Sánchez III and Estefanía of Barcelona	Placencia 1068 3 children	4 June 1076 Peñalén

TURL insights: static dataset

- TURL dataset is static: once a table has been pre-processed there's no way to change it, by, for example, exchanging rows or columns, where this can be viewed as augmentation at table level.
    - For CEA, by default there's no split if the entity number is greater than a predefined integer (during the pre-training phase this number is set to 100)
- This prevent us to apply different table contexts for duplicated tables at runtime
- This prevent us also to apply different context for a given mention at runtime
- This prevent us also to grab different candidates at runtime (pre-training phase)

New ideas

- Cell-oriented: let TURL be cell-oriented, i.e. instead of processing a whole table let it process single cells
  - [Table^{m} context (page title, table caption, …), $e^{m}_{i,j}$, $\text{context}^{m}_{i,\ j-k:j+k}$, $\text{context}^{m}_{i-k:i+k,\ j}$]
  - K is the neighbourhood dimension
    - K can be changed from batch to batch to help the model generalize to different contexts dimension
- Same Column Prediction: additional pre-training signal (inspired by BERT's Next Sentence Prediction)
  - Given two cells, predict whether they belong to the same column/row
- Pre-training with N candidates (generated once, out of M>>N)
  - N can be changed from batch to batch to help the model generalize to different candidate list

New ideas

- NIL prediction
  - Add the ability to answer NIL (add a new token in the vocabulary)
- Consider context during CEA fine-tuning
- Fix candidates set during CEA fine-tuning to be at most N
- Make TURL use and understand wikidata statements?

Thank you for your

$$\text{Attention}(Q,K,V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V$$