Mind The Facts: Knowledge-Boosted Coherent Abstractive Text Summarization

Beliz Gunel[1], Chenguang Zhu[2], Michael Zeng[2], Xuedong Huang[2]
[1]Stanford University [2]Speech and Dialogue Research Group, Microsoft

Overview

Neural models have become successful at producing abstractive summaries that are human-readable and fluent. However, these models have two critical shortcomings:

1. They often don't respect the facts that are either included in the source article or are known to humans as commonsense knowledge.
2. They don't produce coherent summaries when the source article is long.

Contributions:

1. We incorporate entity-level knowledge from the Wikidata knowledge graph into the encoder-decoder architecture. This makes our model more fact-aware.
2. We utilize the ideas used in Transformer-XL language model in our architecture. This gives us coherent summaries even for long articles.

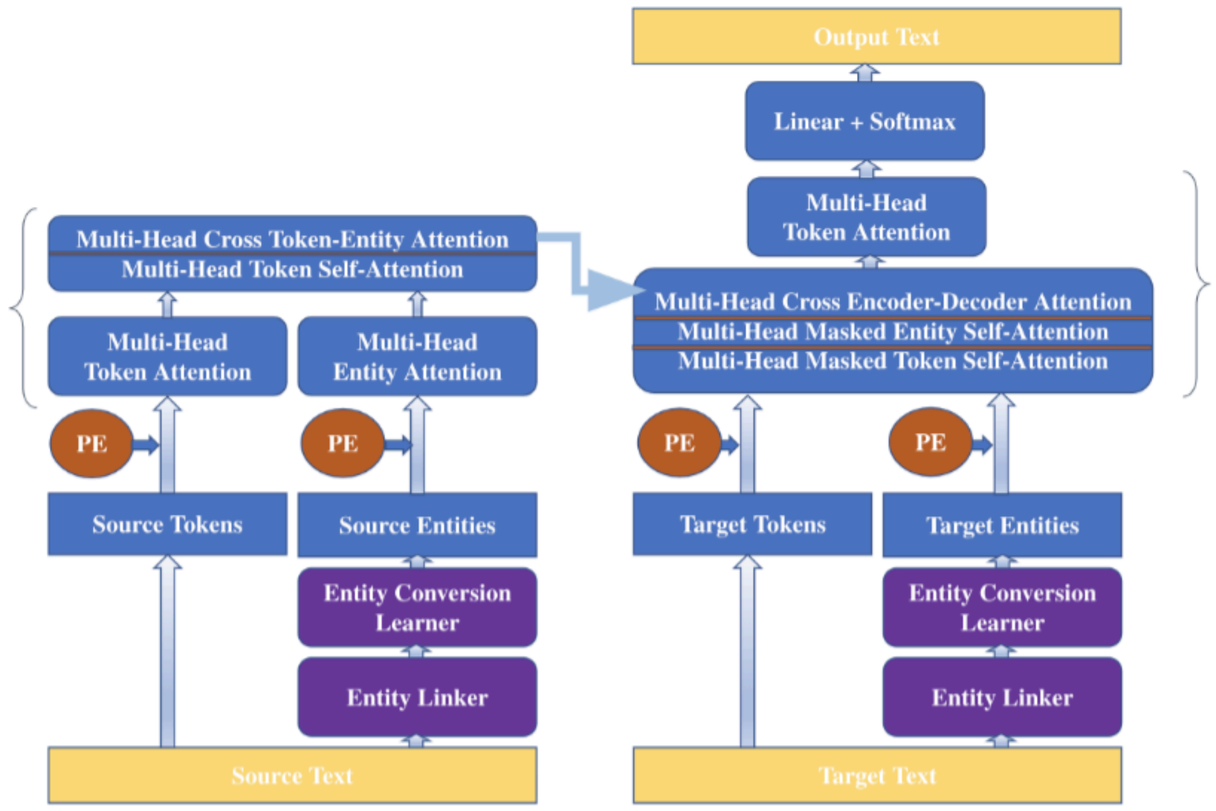


Figure 1. Our model architecture. PE stands for positional encoding. Single encoder and decoder layers are shown in parenthesis. In multi-layer architectures, layers in curly brackets are stacked.

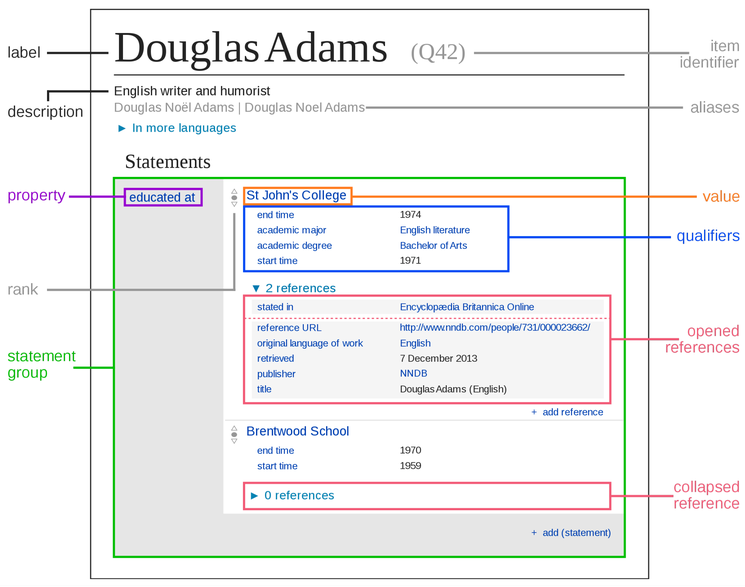
Wikidata Knowledge Graph Entity Embeddings

We sample part of **Wikidata** that has >5M entities, 25K relationship triples. We learn entity embeddings for these entities through **TransE** by Bordes et al. [NeurIPS'13], similar to ERNIE [ACL'19]. We minimize a margin-based ranking criterion over the entity and relationship set as in the following:

$$\mathcal{L} = \sum_{(h,\ell,t)\in S} \sum_{(h',\ell,t')\in S'_{(h,\ell,t)}} \left[\gamma + d(\boldsymbol{h}+\boldsymbol{\ell},\boldsymbol{t}) - d(\boldsymbol{h'}+\boldsymbol{\ell},\boldsymbol{t'})\right]_+$$

where $[x]_+$ denotes the positive part of x, $\gamma > 0$ is a margin hyperparameter, and

$$S'_{(h,\ell,t)} = \{(h',\ell,t)|h' \in E\} \cup \{(h,\ell,t')|t' \in E\}.$$



Quantitative Results

Table 1. Results on CNN/Daily Mail dataset. R used as an abbreviation for ROUGE.

Model	R-1	R-2	R-L
Transformer Baseline	33.351	12.473	30.663
Transformer-Entity w/ Random Entity Emb	33.047	11.536	30.487
Transformer-Entity w/ Wikidata KG Emb	33.741	12.171	31.076
Transformer-XL-Entity w/ Wikidata KG Emb (Our Model)	**33.804**	**12.509**	**31.225**

Table 2. Results on CNN/Daily Mail dataset with high density entities. R used as an abbreviation for ROUGE. >50 ent denotes the slice of test data that has more than 50 entities in the source article.

Model	R-1 (>50 ent)	R-2 (>50 ent)	R-L (>50 ent)
Transformer Baseline	33.423	12.46	30.97
Our model	**34.273**	**13.018**	**32.048**

Encoder-Decoder Architecture for Transformer-XL

Transformers have fixed-length context, which results in bad performance for long text. These fixed-length context segments do not respect sentence boundaries, resulting context fragmentation. Transformer-XL:

1. Introduces the notion of recurrence into a self-attention-based model by reusing hidden states from previous segments.
2. Introduces the idea of relative positional encoding.

Transformer-XL learns dependency that is about 80% longer than RNNs and 450% longer than vanilla Transformers!

Transformer

$$\mathbf{A}^{\text{abs}}_{i,j} = \underbrace{\mathbf{E}_{x_i}^\top \mathbf{W}_q^\top \mathbf{W}_k \mathbf{E}_{x_j}}_{(a)} + \underbrace{\mathbf{E}_{x_i}^\top \mathbf{W}_q^\top \mathbf{W}_k \mathbf{U}_j}_{(b)} + \underbrace{\mathbf{U}_i^\top \mathbf{W}_q^\top \mathbf{W}_k \mathbf{E}_{x_j}}_{(c)} + \underbrace{\mathbf{U}_i^\top \mathbf{W}_q^\top \mathbf{W}_k \mathbf{U}_j}_{(d)}.$$

Transformer-XL (Relative Pos. Enc.)

$$\mathbf{A}^{\text{rel}}_{i,j} = \underbrace{\mathbf{E}_{x_i}^\top \mathbf{W}_q^\top \mathbf{W}_{k,E} \mathbf{E}_{x_j}}_{(a)} + \underbrace{\mathbf{E}_{x_i}^\top \mathbf{W}_q^\top \mathbf{W}_{k,R} \mathbf{R}_{i-j}}_{(b)} + \underbrace{u^\top \mathbf{W}_{k,E} \mathbf{E}_{x_j}}_{(c)} + \underbrace{v^\top \mathbf{W}_{k,R} \mathbf{R}_{i-j}}_{(d)}.$$

Transformer-XL Full

$$\begin{aligned}
\widetilde{\mathbf{h}}_\tau^{n-1} &= \left[\text{SG}(\mathbf{m}_\tau^{n-1}) \circ \mathbf{h}_\tau^{n-1}\right] \quad \text{(Recurrence)} \\
\mathbf{q}_\tau^n, \mathbf{k}_\tau^n, \mathbf{v}_\tau^n &= \mathbf{h}_\tau^{n-1}\mathbf{W}_q^{n\top}, \widetilde{\mathbf{h}}_\tau^{n-1}\mathbf{W}_{k,E}^{n\top}, \widetilde{\mathbf{h}}_\tau^{n-1}\mathbf{W}_v^{n\top} \\
\mathbf{A}_{\tau,i,j}^n &= \mathbf{q}_{\tau,i}^{n\top}\mathbf{k}_{\tau,j}^n + \mathbf{q}_{\tau,i}^{n\top}\mathbf{W}_{k,R}^n\mathbf{R}_{i-j} \\
&\quad + u^\top \mathbf{k}_{\tau,j} + v^\top \mathbf{W}_{k,R}^n \mathbf{R}_{i-j} \\
\mathbf{a}_\tau^n &= \text{Masked-Softmax}(\mathbf{A}_\tau^n)\mathbf{v}_\tau^n \\
\mathbf{o}_\tau^n &= \text{LayerNorm}(\text{Linear}(\mathbf{a}_\tau^n) + \mathbf{h}_\tau^{n-1}) \\
\mathbf{h}_\tau^n &= \text{Positionwise-Feed-Forward}(\mathbf{o}_\tau^n)
\end{aligned}$$

Quantitative Example

Ground Truth Summary
Steve McClaren is expected to take Newcastle job if Derby don't go up. Rams are currently battling for Championship promotion via the play-offs. Paul Clement is a leading candidate for job. Derby will make formal contact with Real Madrid if McClaren leaves.

Transformer Baseline Output
Steve McClaren is a leading candidate to replace Steve McClaren. The 42-year-old has established a reputation as one of European football's leading coaches in recent years, working on mainly under Carlo Ancelotti. The former Manchester United boss is keen to secure promotion into the Premier League next season.

Output of Our Model
Paul Clement is a leading candidate to replace Steve McClaren at Derby County. The former England boss has established a reputation as one of Europe's leading football coaches in recent years. Clement is currently a Real Madrid coach.

Figure 2. Comparison of the transformer baseline output and the output of our proposed model. Ground truth summary is sampled from the CNN/Daily Mail summarization corpus. Baseline model makes factual errors, while our model respects the facts through incorporating entity-level knowledge from Wikidata knowledge graph.

Baseline model makes several factual errors based on our manual fact-checking:

1. Neither McClaren nor Paul Clement was 42 years old at the time.
2. Neither McClaren nor Paul Clement worked as a Manchester United boss.

On the other hand, our model respects the facts through incorporating world knowledge from Wikidata. Again, based on our fact checking, we find:

1. Clement was working in Real Madrid before he was appointed the manager of Derby County.
2. Although "England boss" is too broad, he did work at Chelsea 2009-2011.