The Thirty-Fifth AAAI Conference on Artificial Intelligence (AAAI-21)

KAN: Knowledge-aware Attention Network for Fake News Detection

Yaqian Dun[1,2], Kefei Tu[1,2], Chen Chen[1,2*], Chunyan Hou[3], Xiaojie Yuan[1,2]
[1] College of Computer Science, Nankai University, Tianjin, China
[2] Tianjin Key Laboratory of Network and Data Security Technology, Tianjin, China
[3] School of Computer Science and Engineering, Tianjin University of Technology, Tianjin, China
{dunyaqian, tukefei}@dbis.nankai.edu.cn, {nkchenchen, yuanxj}@nankai.edu.cn, chunyanhou@163.com

AAAI'21

211102 Chia-Chun Ho

Outline

Introduction

Related Work

Proposed Model

Experiments

Conclusion and Future work

Comments

Introduction

Social Media

- Social media has become a platform for people to obtain and exchange information.
- More and more people publish and read news online.
- Meanwhile, it also gradually becomes an ideal place for widespread of fake news.
- Since fake news distorts and fabricates facts maliciously, its extensive dissemination has extremely negative impacts on individuals and society.
- It's desirable and socially beneficial to detect fake news in social media.

Introduction

Fake news detection

- Early studies mainly focus on machine learning model based on feature engineering.
- After emergence of deep learning, various deep-learning-based approaches are proposed and greatly improve the detection performances.
- Although the existing deep learning methods have achieved great success to detect fake news based on the high-level feature representation of news contents.
    - They ignore the external knowledge by which people usually judge the authenticity of the news.

Introduction

Entity mentions in news content

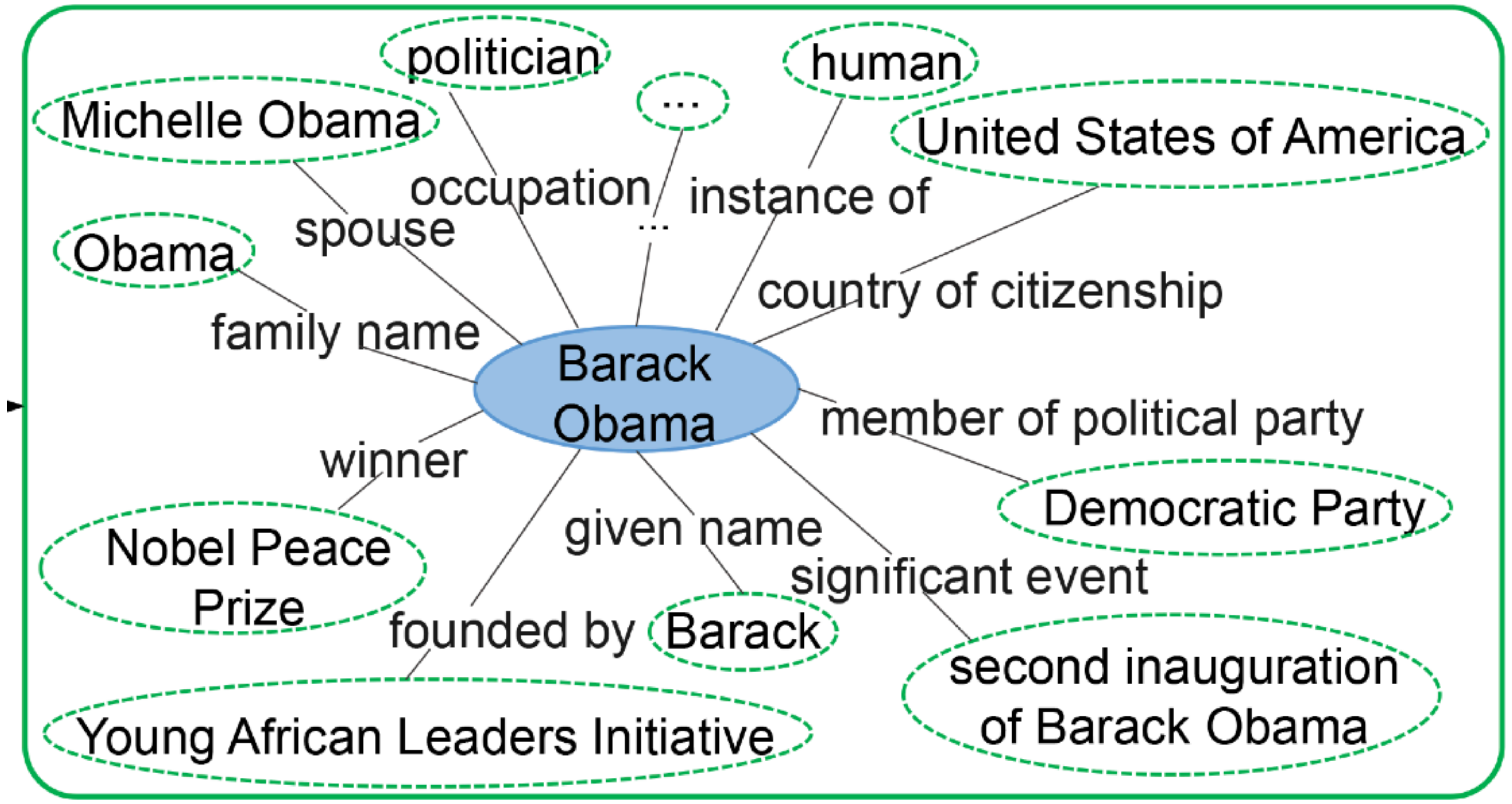
> MARCH 13, 2018 BY
>
> **Court Orders Obama To Pay $400 Million In Restitution**
>
> The West Texas Federal Appeals Court, operating out of the 33rd District, has ordered that Barack Obama repay $400 Million to the American people for funds he says were “lost” during an illegal transaction with Iranian hard-liners. Judge Gary Jones and Judge Amanda Perry stood together to overrule Judge Kris Weinshenker in a split decision.
>
> … …
>
> “Never before has a President taken a knee during his term and flaked on his duties like Obama did. This money is owed to the American perople and then some. Punitive damages weren't assessed. Had they been it is doubtful Mr. Obama would ever recover financially.”

- News content is highly condensed and comprised of a large number of entity mentions.
- A named entity could possibly denote different entity mentions because a named entity may have multiple textual forms, such as its aliases, abbreviations and alternate spellings.
- These knowledge-level judgments and connections are helpful to evaluate the credibility of the news.
- However, these mentions cannot be understood directly based on the text content of news.

Introduction
Knowledge graph

- To extract deep logical connections among entities, it's necessary to incorporate the knowledge information in knowledge graph.

- Knowledge graph is a multi-relational graph which is composed of entities as nodes and relationships as edges with different types.



Introduction

Knowledge graph

- This knowledge is beneficial to understand news because:
  - The ambiguous entity mentions usually occur in news contents.
  - The ambiguity of mentions can be avoided by associating each mention in news content with its corresponding entity in knowledge graph.
  - Knowledge graph also can provide more complementary information about relevant entities, which is helpful for learning knowledge-level relationships among entities in news and improving the performance of fake news detection.

Introduction

Knowledge-aware Attention Network (KAN)

- Propose the method that incorporates external knowledge from knowledge graph for fake news detection.
- First, identify entity mentions in the news contents, and the obtain corresponding entities through external knowledge graph such as YAGO, Freebase, Wikidata and Probase.
- Next, extract the entity context of each entity (i.e., it's directly connected neighbors in knowledge graph) as auxiliary information.
- Finally, these entities and their entity contexts serve as external knowledge so as to learn both semantic-level and knowledge-level representations of news.

Introduction

Knowledge-aware Attention Network (KAN)

- To fuse external knowledge into the model effectively, it's key to figure out the relative importance of each entity associated with news content.
- Thus, use News-Entities attention to calculate the semantic similarity between news contents and its corresponding entities, where each entity is assigned a weight to represent its importance.
- For the purpose of integrating entity contexts, design News-Entities and Entity contexts attention to assign a weight to the entity context by the vitality of the corresponding entity.
- Finally, the representations of these are concatenated and fed into a fully-connected network to predict the veracity of the news.

Introduction

Contribution

- Propose to incorporate entities and their entity contexts which are distilled from knowledge graph for fake news detection.
- Propose a Knowledge-aware Attention Network for fake news detection.
- To integrate knowledge into news more reasonable and effective, introduce two attention mechanisms to obtain the relative importance of entities and entity contexts.
- Conduct extensive experiments on three standard datasets for fake news detection.

Related Work

of fake news detection

- Early works ('11/'12) mainly focus on designing a complementary set of hand-crafted features based on linguistic features.
- To expand beyond the specificity of hand-crafted features, someone propose a deep neural network (EANN ('18)) to capture multi-modal data features for fake news detection.
- Someone ('18) utilize CNN and GRU to capture useful patterns from user profiles.
- The authors use knowledge graph to capture latent knowledge-level connections among news entities for better exploration.

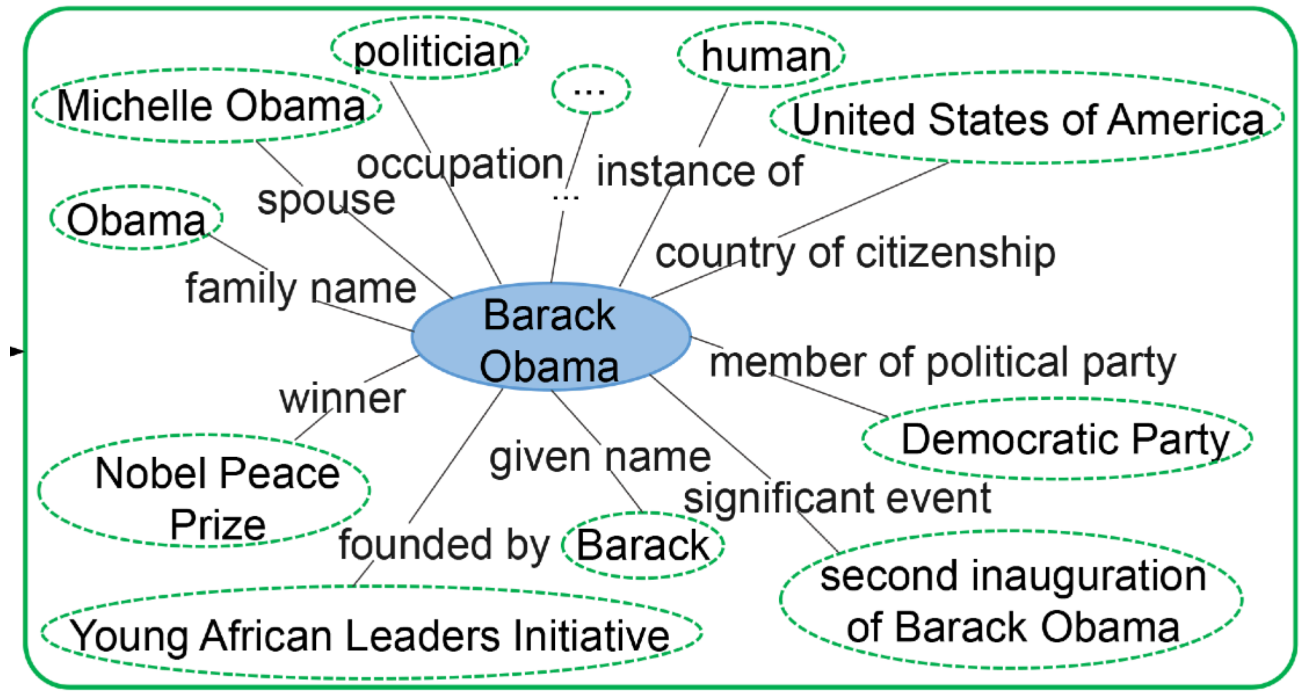
Problem Statement
Notations

- Defined fake news detection task can be defined as a binary classification problem.
- Each piece of news S is composed of a sequence of words. i.e., $\boldsymbol{S} = \{w_1, w_2, \cdots, w_n\}$
- One or several words may be associated with an entity e_i in the knowledge graph.
- Each e_i have many immediate neighbors in the knowledge graph. The neighbor entities of entity e_i is defined as "entity context" $ec(e_i)$ of the entity e_i.

Problem Statement

Problem formulation

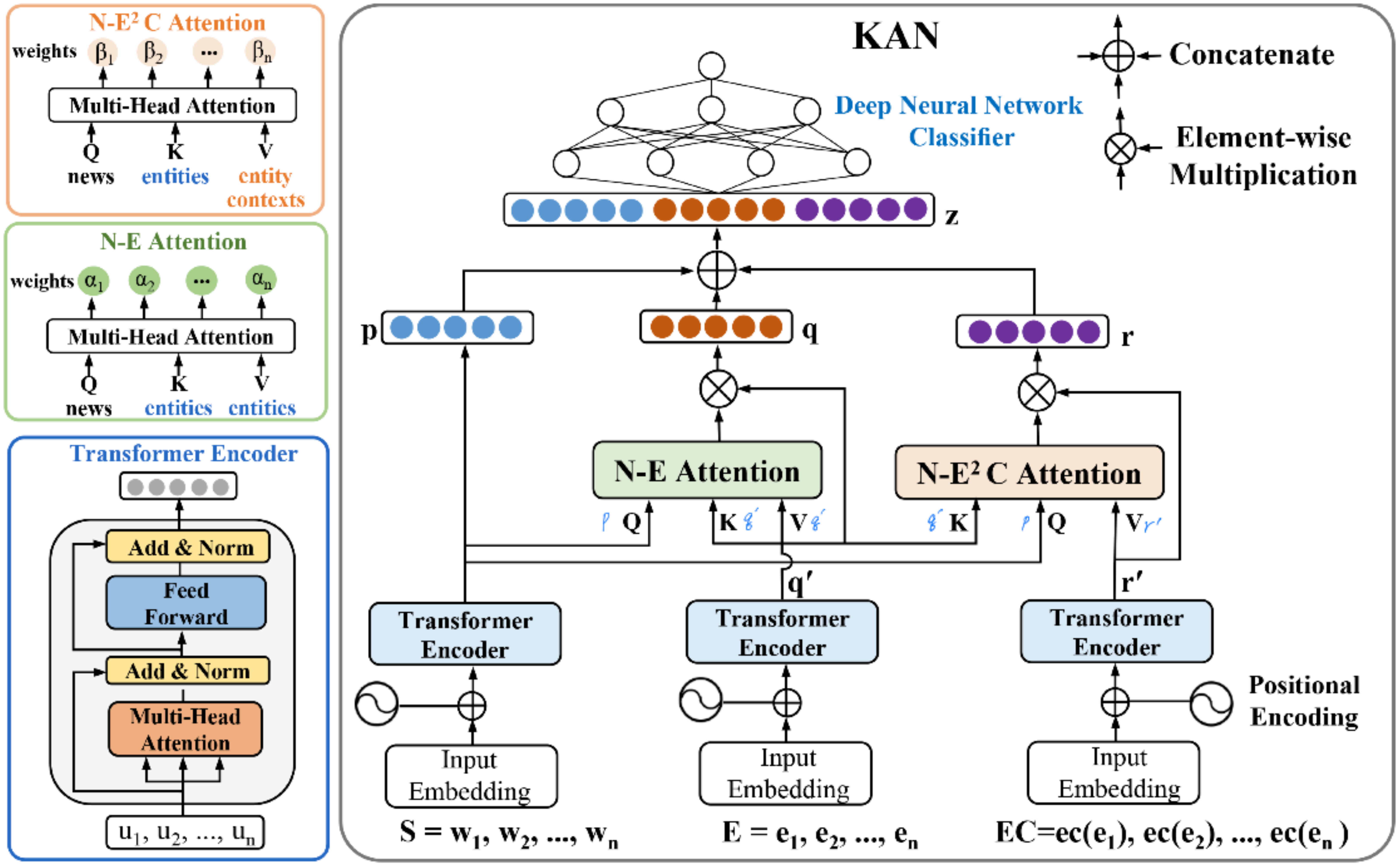
- Given a news article $\boldsymbol{S} = \{w_i\}$ as well as the relevant entities $\boldsymbol{E} = \{e_i\}$ and entity contexts $\boldsymbol{EC} = \{ec(e_i)\}$.
- Aim to learn such a fake news detection function $F : F(\boldsymbol{S}, \boldsymbol{E}, \boldsymbol{EC}) \rightarrow y$
  - $y \in \{0,1\}$ is the ground-truth label of news.
- Practically, use the average value of the vector representations of the neighbor entities to represent $ec(e_i)$.



Proposed Model

Framework overview

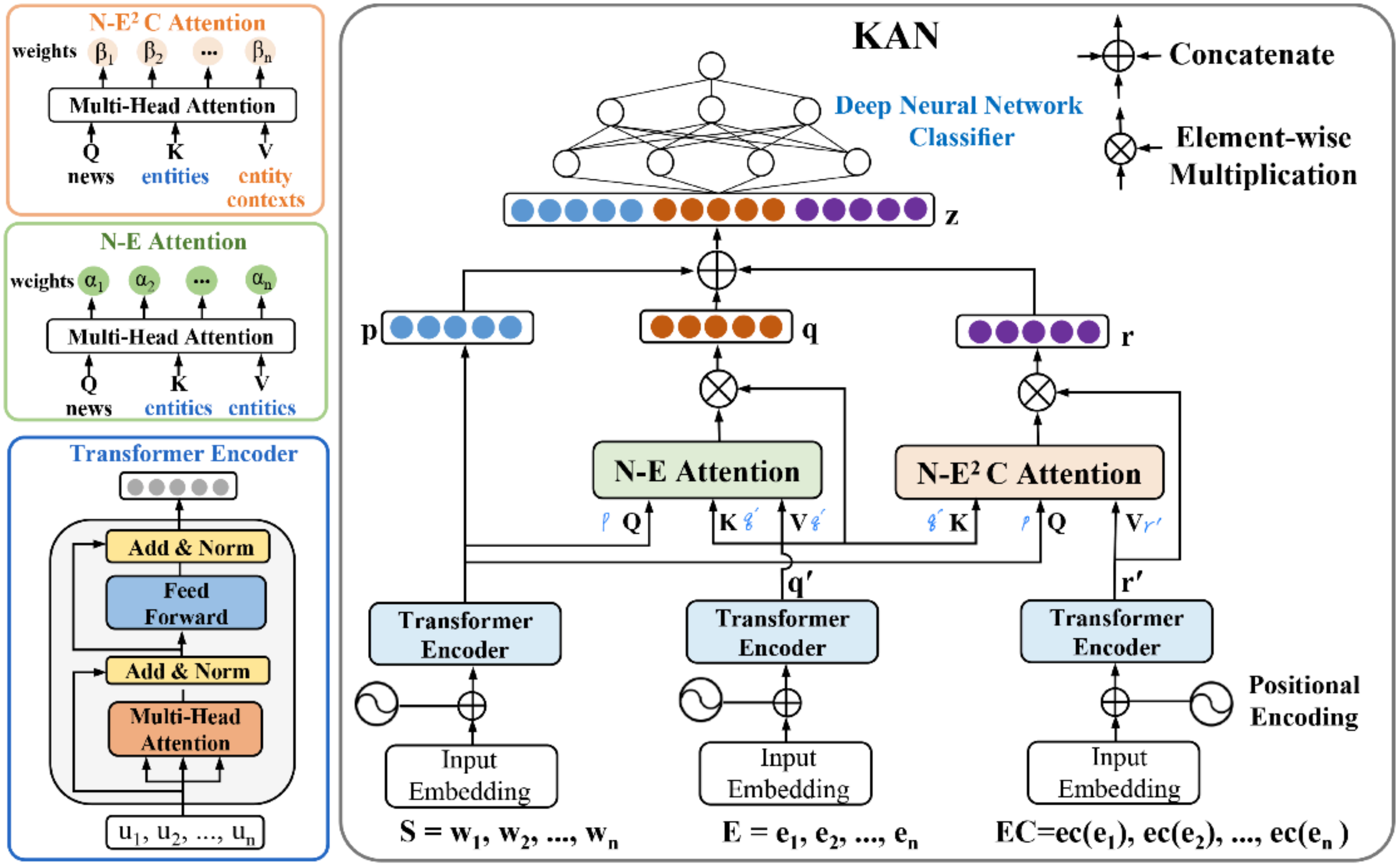
- Text encoding
- Knowledge Extraction
- Knowledge Encoder
- Knowledge-aware Attention
- Deep Neural Network Classifier



Proposed Model

KAN Framework

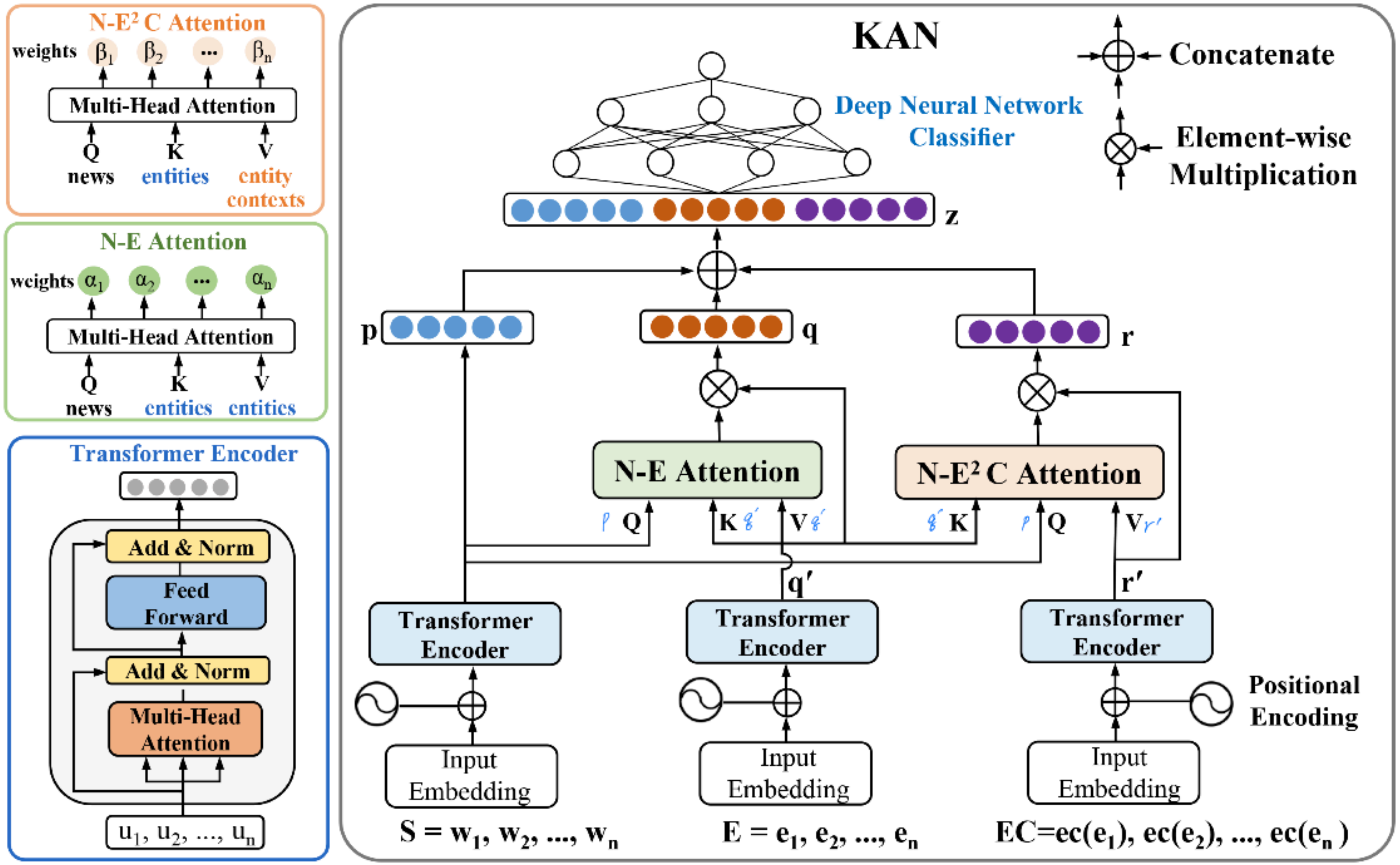
- Input to KAN consists of news contents, entities, and entity contexts.
- Output of KAN is the probability distribution of labels over classes.



Proposed Model

KAN Framework

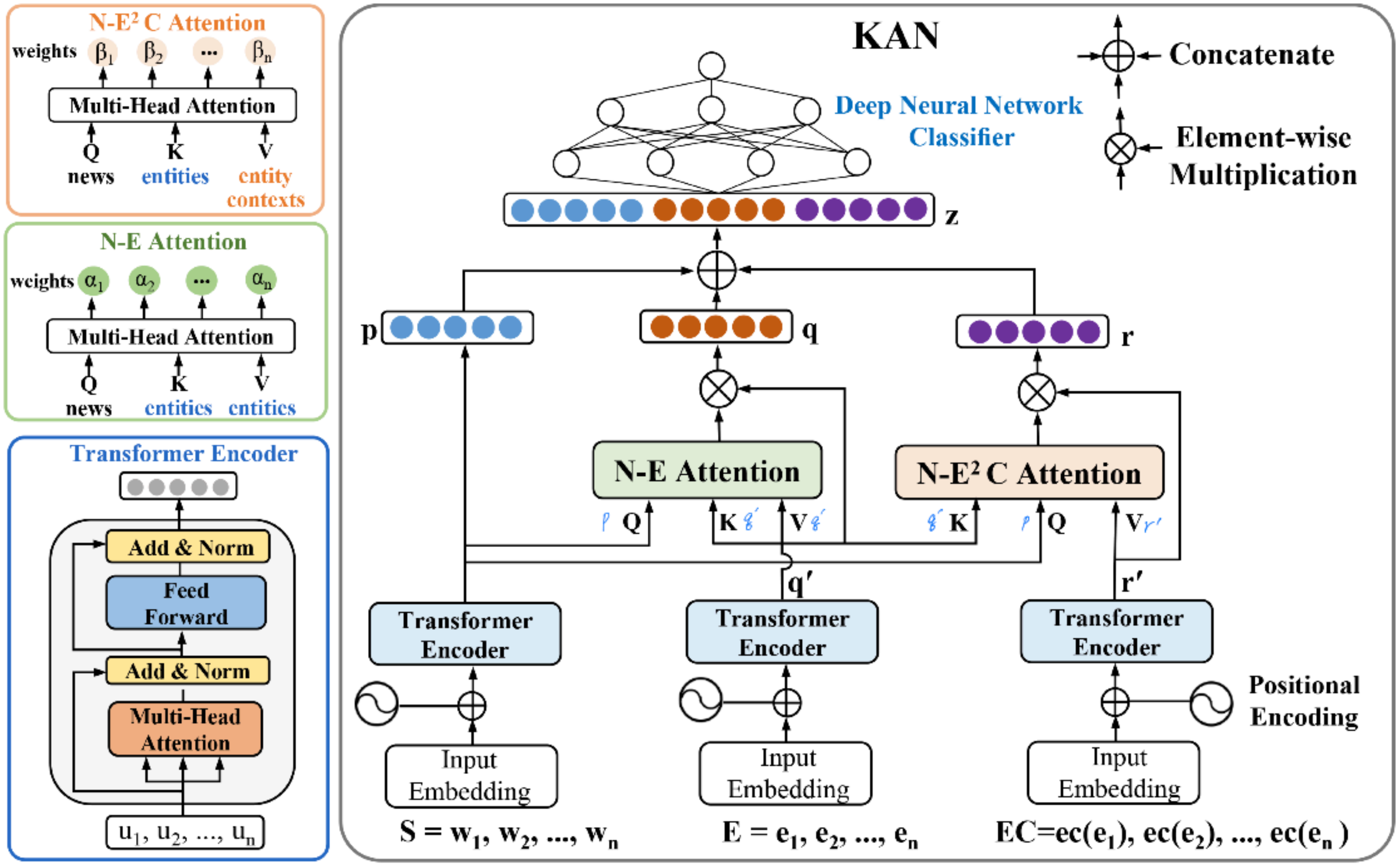
- For each piece of news, a Transformer Encoder is used to encode news contents and generate the representation of news.



Proposed Model

KAN Framework

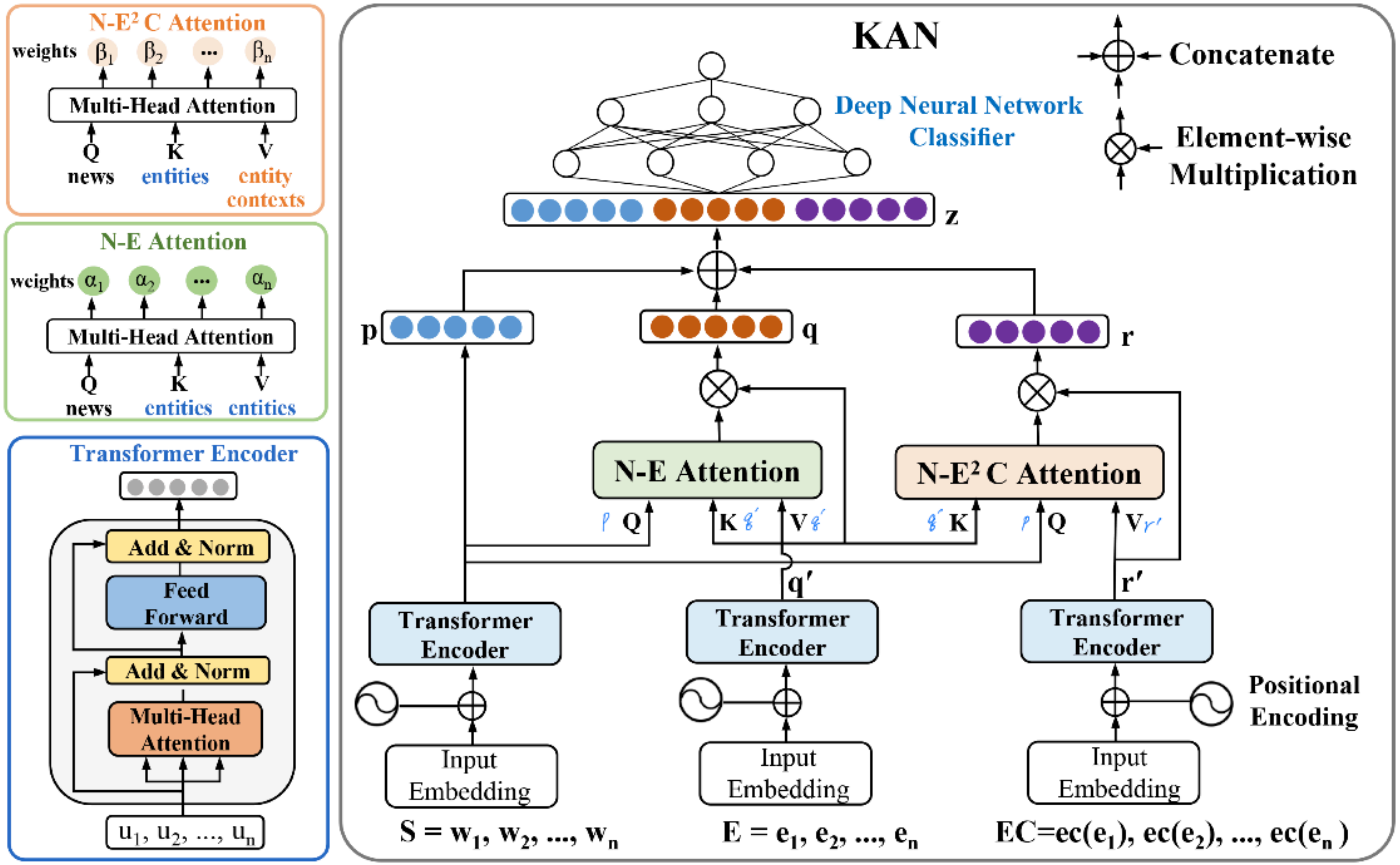
- Extraction of entities and entity contexts from knowledge graph.
- Then, these two kinds of extracted external knowledge are encoded by transformer encoders respectively to produce the knowledge encoding.



Proposed Model

KAN Framework

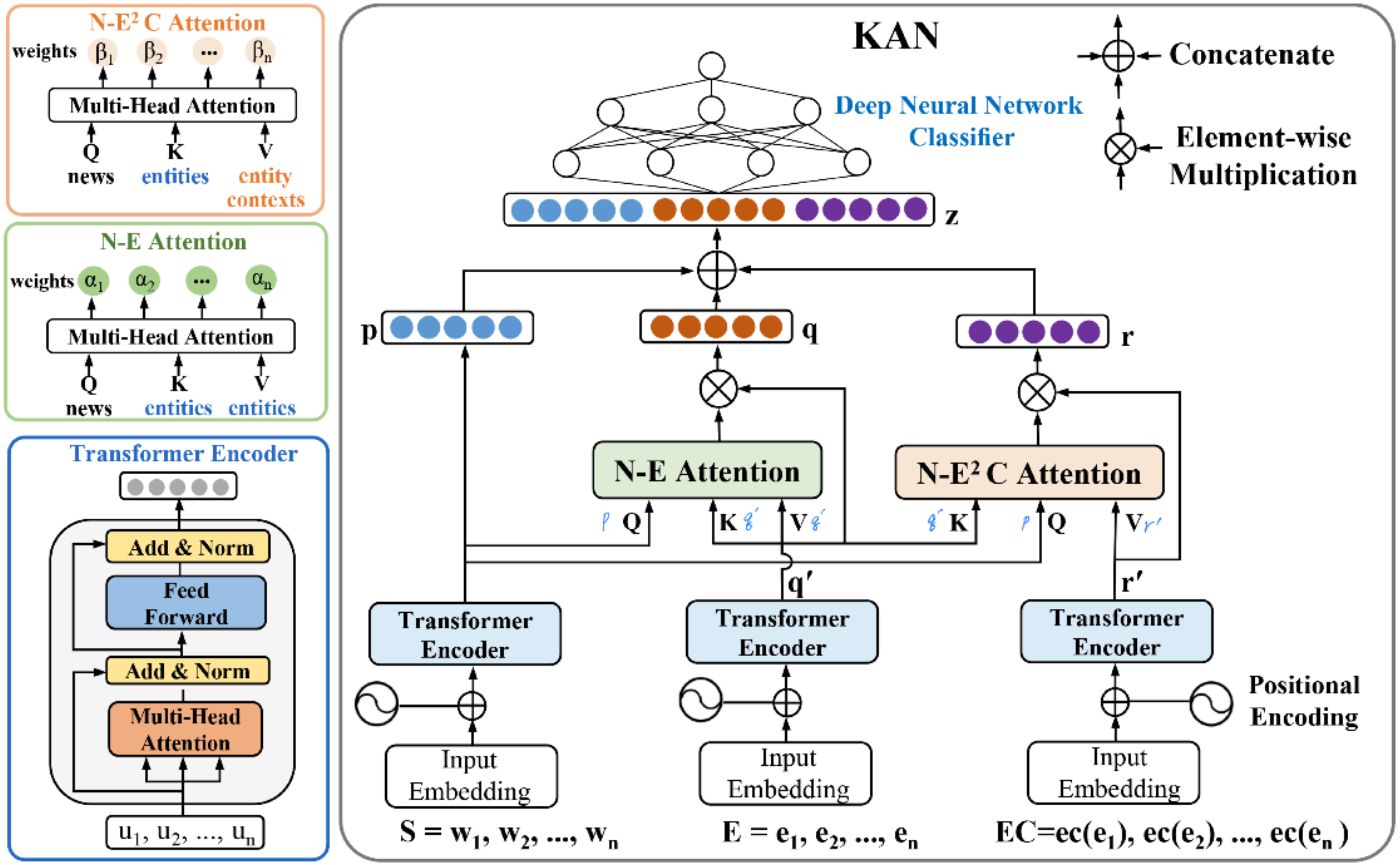
- To fuse knowledge encoding into the model effectively, design two attention mechanisms to measure the relative importance of entities and entity contexts, and then aggregate their vector representations with different weights.



Proposed Model

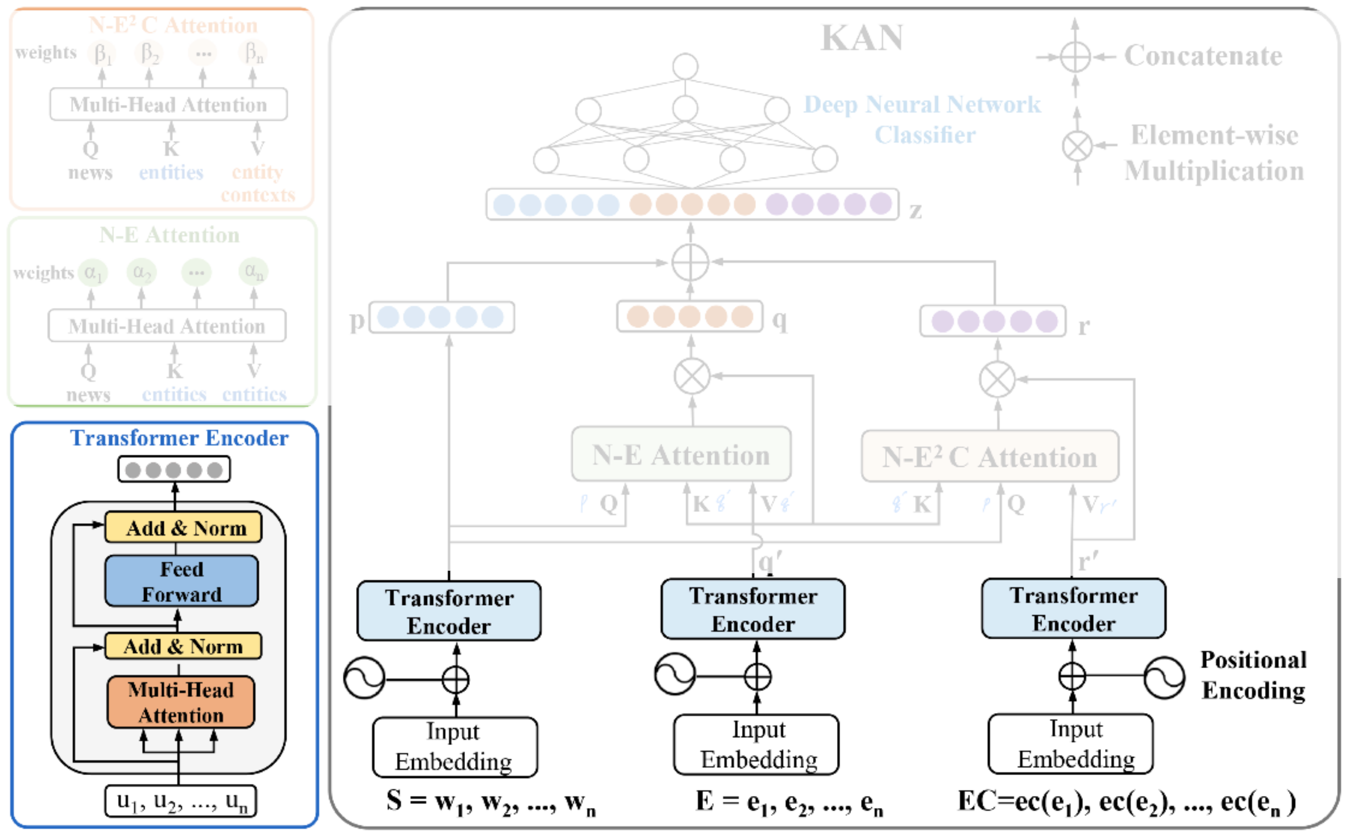
KAN Framework

- Representation of news, entities, and their contexts are concatenated and fed into a fully-connected network to predict the veracity.



Proposed Model

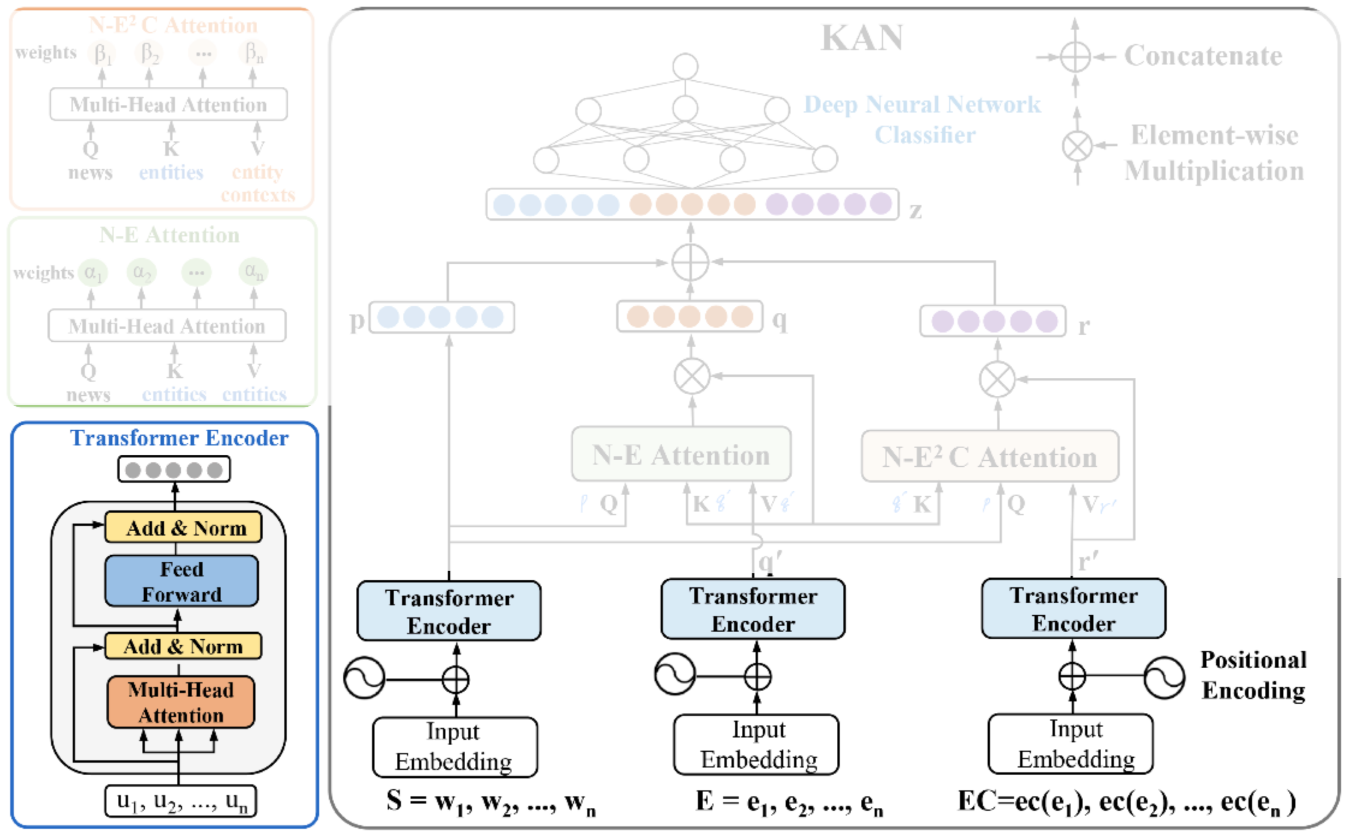
Text Encoding



- This module aims to produce the news content representation p.
- To capture the representation of news contents, employ Transformer Encoder as the core of the module.
- Transformer contains self-attention layers which can learn long-term dependency.
- Meanwhile, it's able to capture the sequence information through positional encoding and has a strong ability to extract semantic features.

Proposed Model

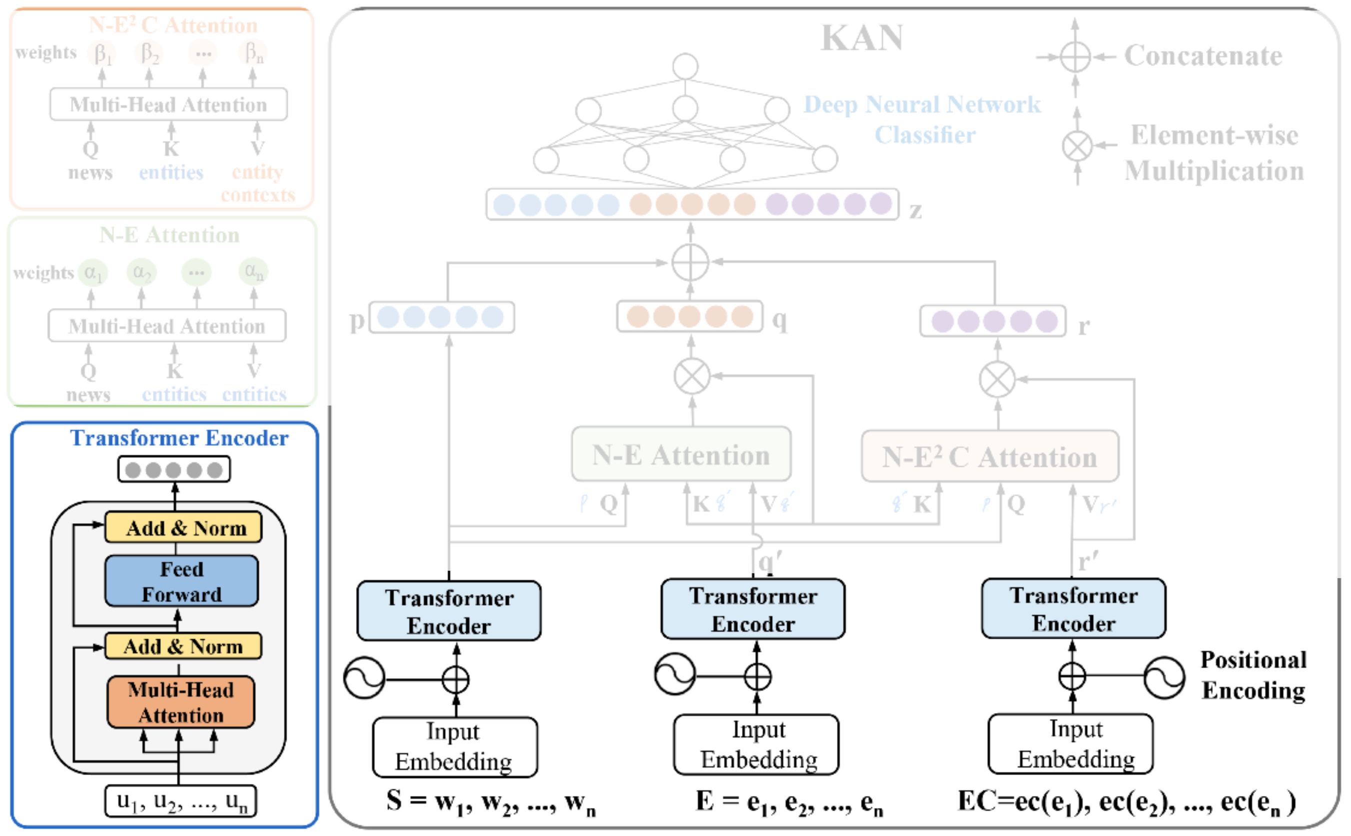
Text Encoding



- Transformer encoder generates the text encoding from the original word sequence and positional encoding.
- Given a piece of news $\boldsymbol{S} = \{w_1, w_2, \cdots, w_n\}$ of length n, each word w_i is projected into a continuous word embedding w_i' from a continuous word embedding matrix $\boldsymbol{M} \in \mathbb{R}^{V \times d}$.
- Then, obtain the news vectors $\boldsymbol{S}' = \{w_1', w_2', \cdots, w_n'\}$, $\boldsymbol{S}' \in \mathbb{R}^{n \times d}$.
- Moreover, in order to make use of the word order in the news, positional encodings are used and combined with the word embeddings:
    - $u_t = w_t' + pos_t$, pos_t: position encoding for t-th word in the news

Proposed Model

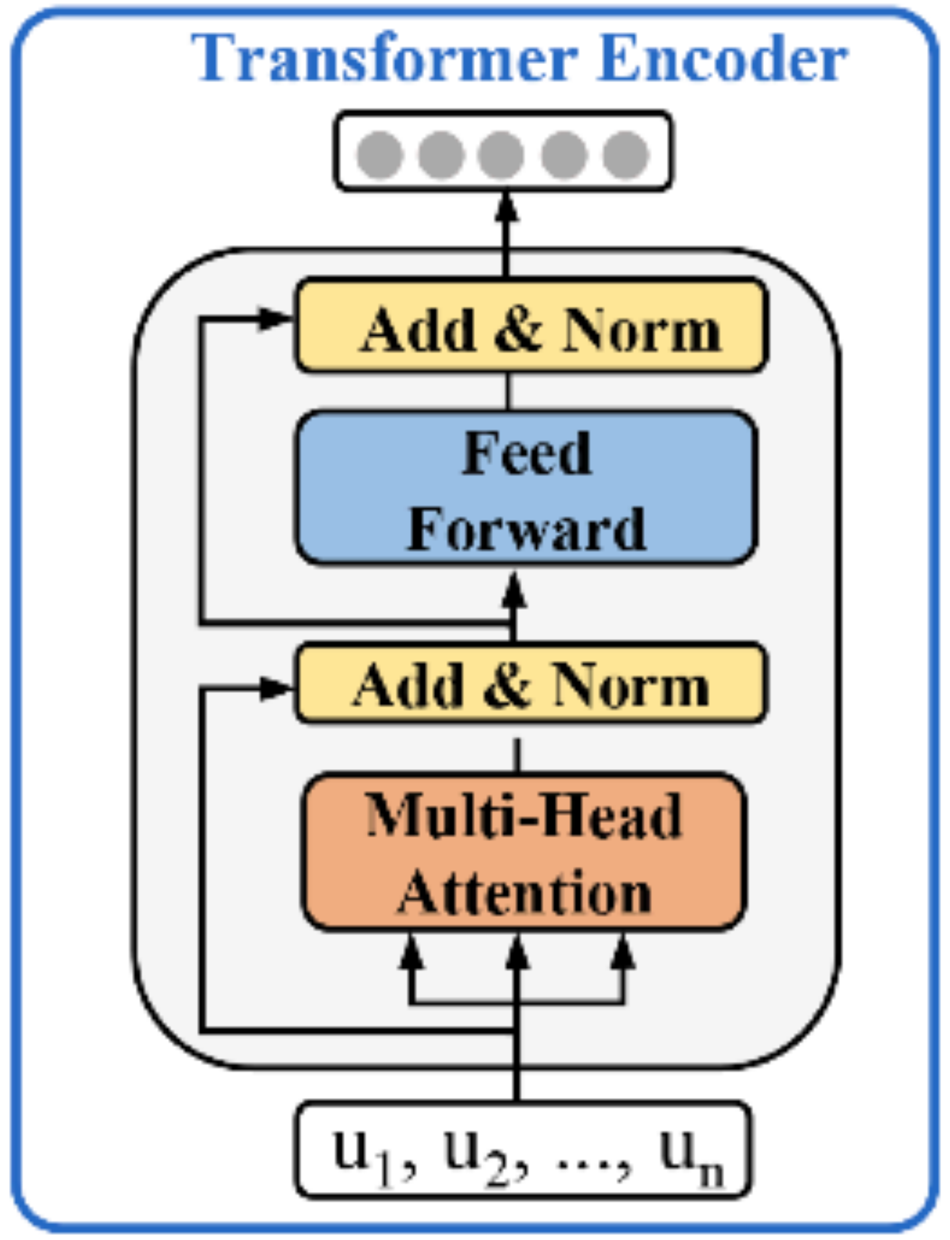
Text Encoding

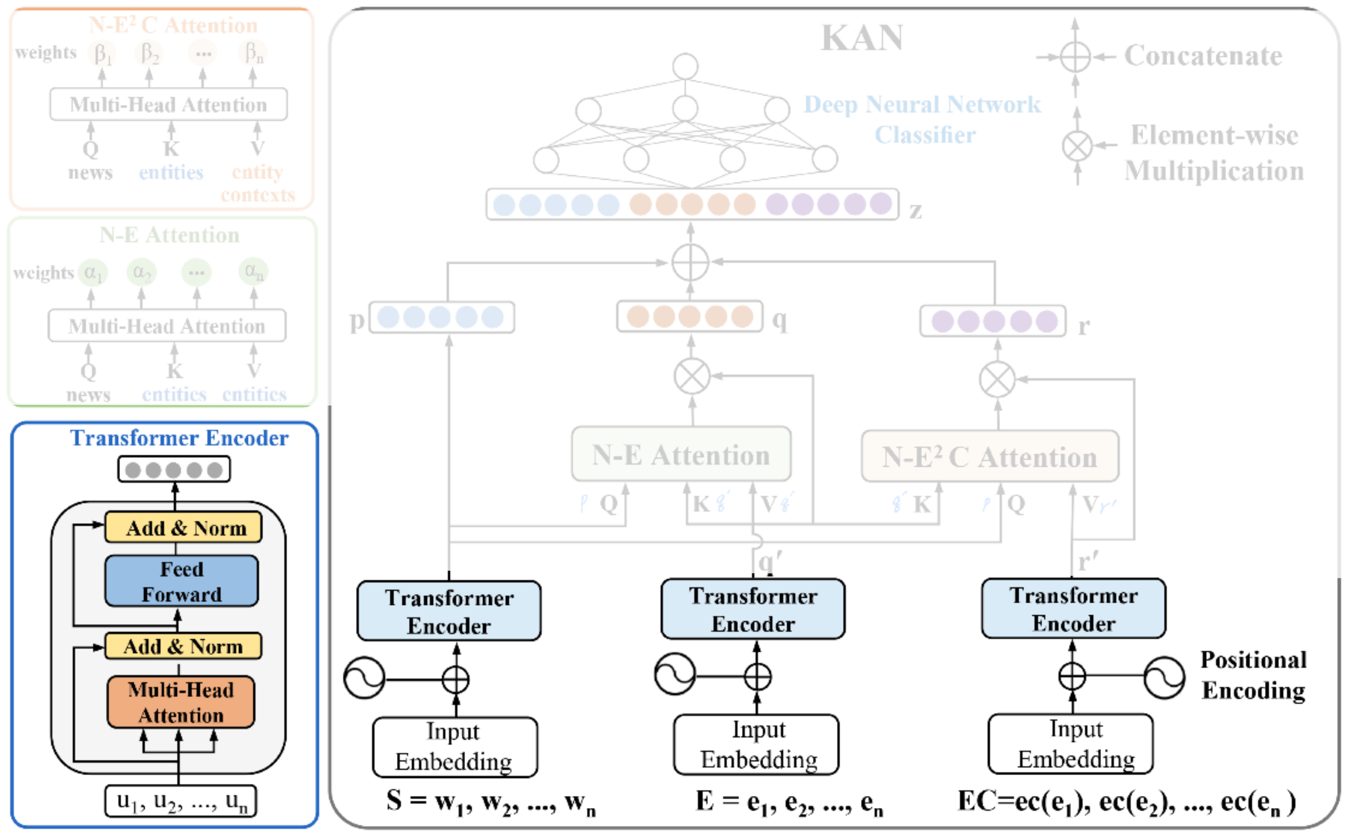


- Denote $u = u_0, \cdots, u_n \in \mathbb{R}^{n \times d}$ as input encodings to the bottoms of transformer encoder.

- In general, the architecture of encoder is stacked with identical layers.

- Each layer is constructed by multi-head self-attention mechanism, residual connection, layer normalization and fully connected feed-forward network.

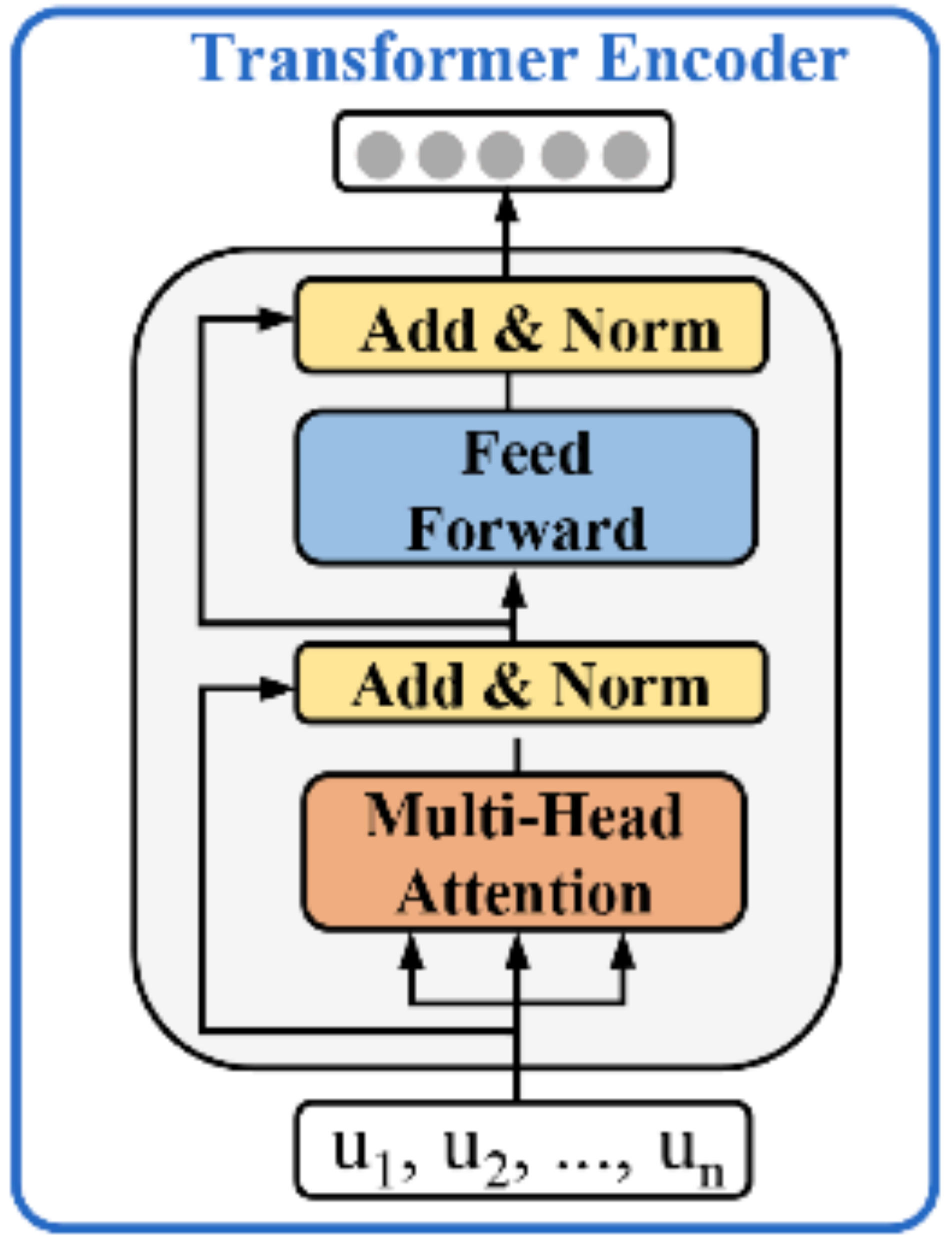




Proposed Model

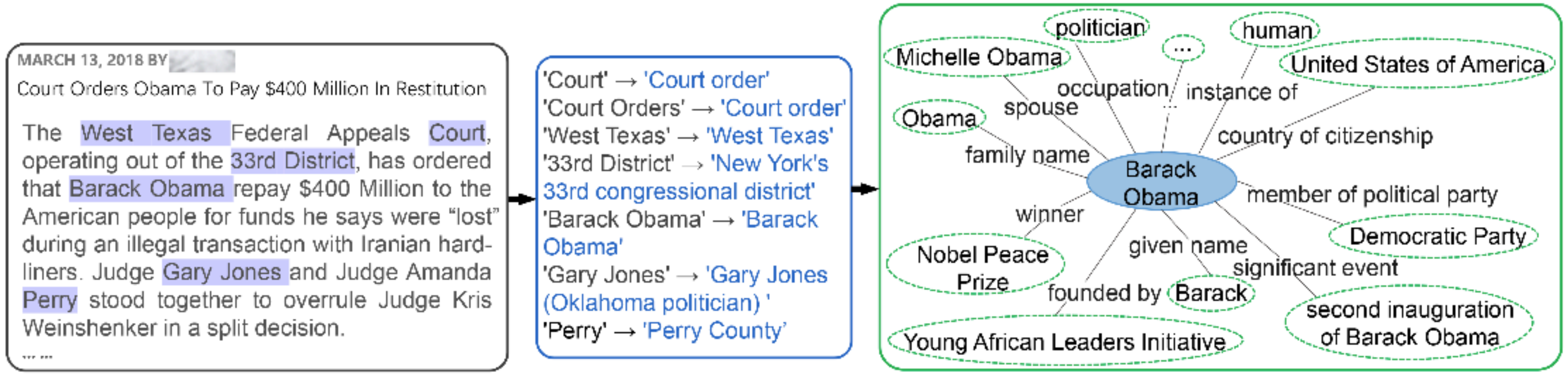
Text Encoding

- In this paper, employ one layer Transformer Encoder to process the input encodings u:
  - $\tilde{a} = \mathrm{MultiHeadAttention}(u)$
  - $a = \mathrm{LayerNorm}(\tilde{a} + u)$
  - $\tilde{u} = \mathrm{FeedForwardNetwork}(a)$
  - $p = \mathrm{LayerNorm}(\tilde{u} + a)$



Proposed Model

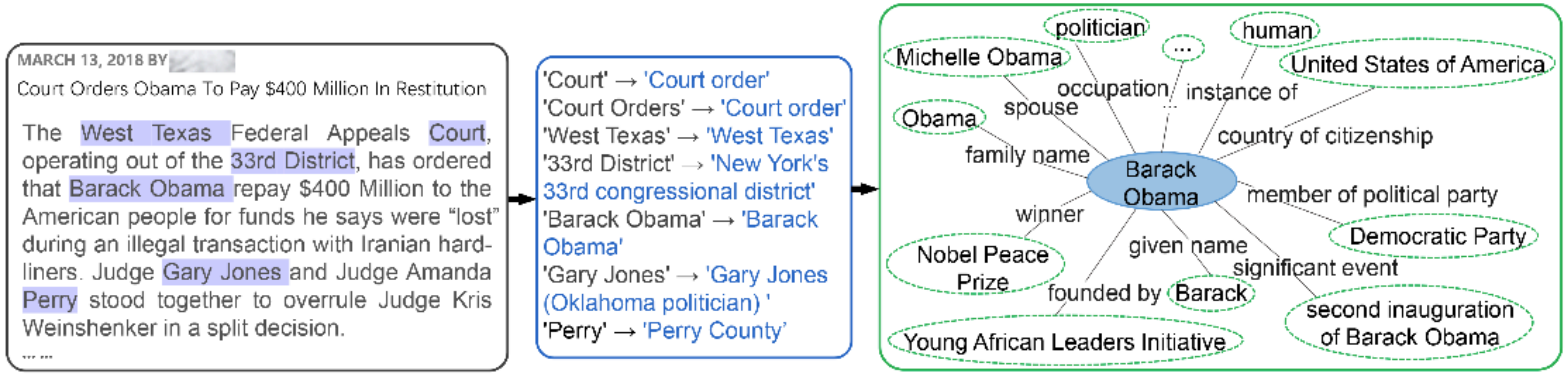
Knowledge Extraction



(a) A Piece of News (b) Entity Linking (c) Entity Context Extraction

- This module aims to retrieve relevant entities from the knowledge graph.
- Through entity linking, the entity mentions in the news contents are identified and aligned with their counterpart entities in the knowledge graph.
- After that, can acquire entities sequence $\boldsymbol{E} = \{e_1, e_2, \cdots, e_n\}$.
- The entity contexts are chosen out according to the linked entities in the former step.

Proposed Model
Knowledge Extraction

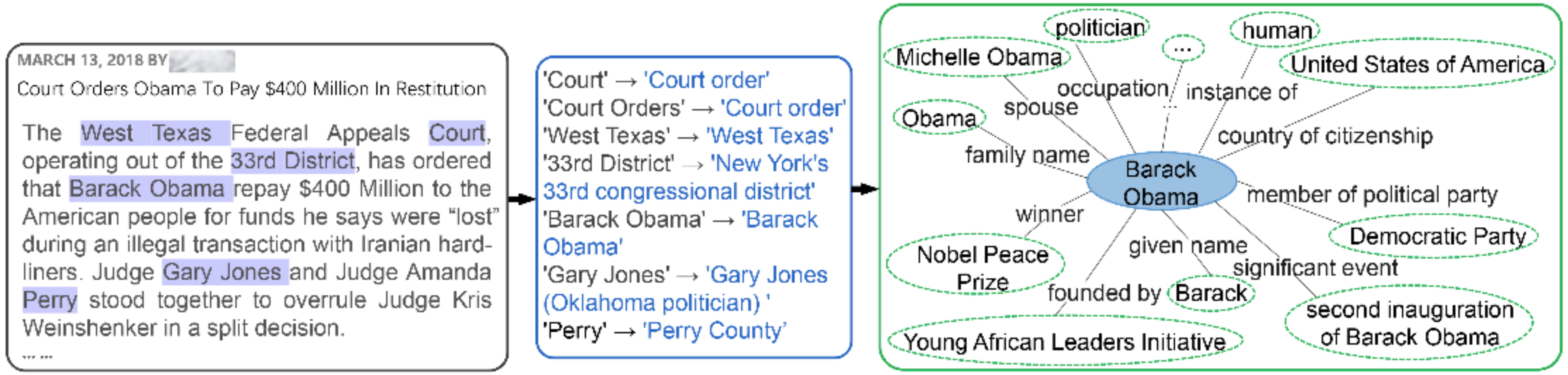


(a) A Piece of News (b) Entity Linking (c) Entity Context Extraction

- The "entity context" of entity e_i is defined as the immediate neighbors in the knowledge graph.
- Extract neighbors entities with one-hop distance related to the current entity:
  - $\boldsymbol{ec}(e_i) = \{e \mid (e, rel, e_i) \in G \ \text{ or } \ (e_i, rel, e) \in G\}$
  - rel: relation between two entities, G: knowledge graph
- After the entity context is distilled from a knowledge graph, each entity is associated with entity context set, then can obtain entity contexts sequence $\boldsymbol{EC} = \{ec(e_1), ec(e_2), \cdots, ec(e_n)\}$.

Proposed Model

Knowledge Encoder



(a) A Piece of News (b) Entity Linking (c) Entity Context Extraction

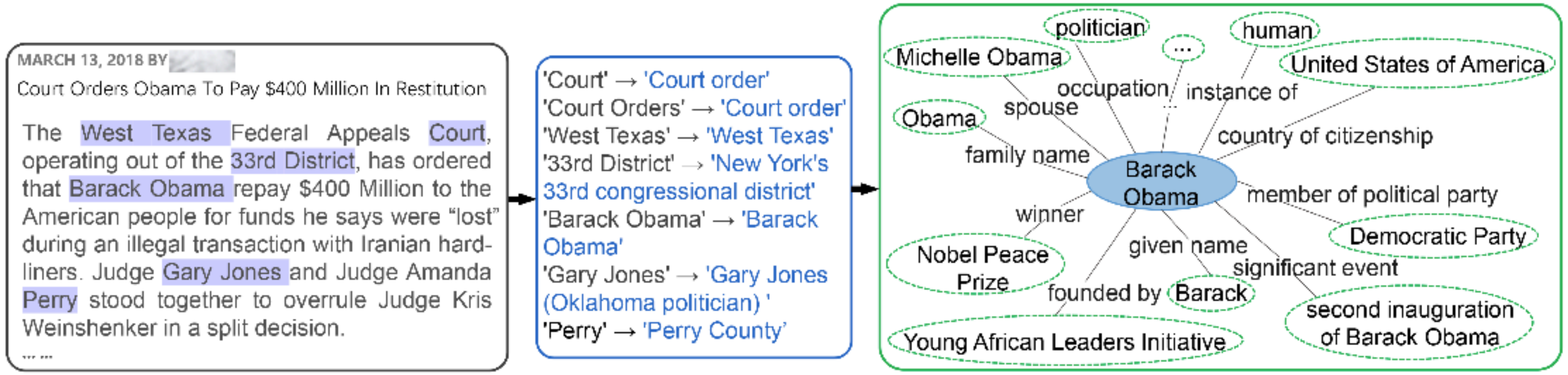
- Introduction of external knowledge can provide more complementary information and reduce the ambiguity caused by entity mentions in news.

- Given a piece of news, entities and entity contexts related to this news can help to boost the detection performance.

- Extracted entities sequence E and entity contexts sequence EC are embedded by word2vec, and then obtain the entities embedding $E' = \{e'_1, e'_2, \cdots, e'_n\}, E' \in \mathbb{R}^{n \times d}$ and entity contexts embedding $EC' = \{ec'_1, ec'_2, \cdots, ec'_n\}, EC' \in \mathbb{R}^{n \times d}$.

Proposed Model

Knowledge Encoder

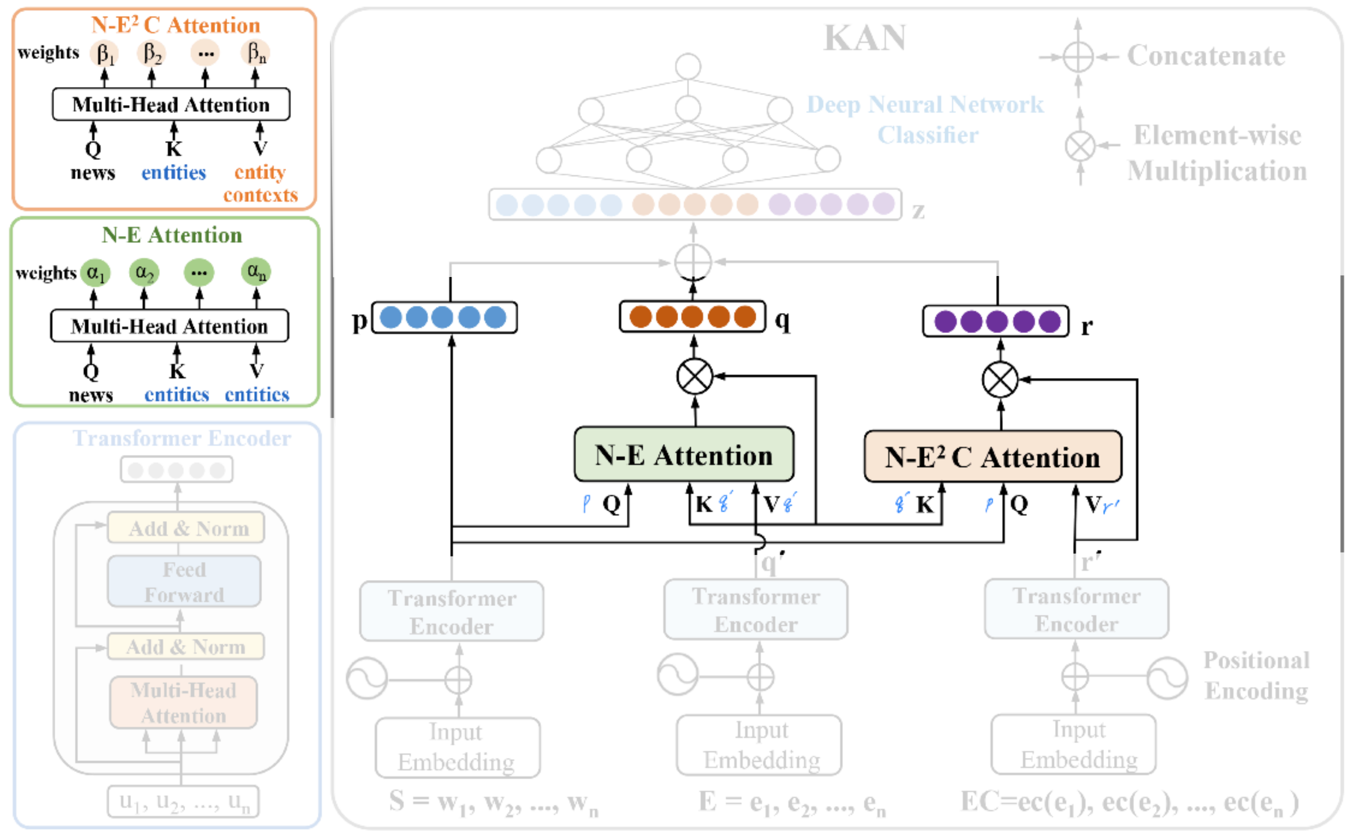


(a) A Piece of News (b) Entity Linking (c) Entity Context Extraction

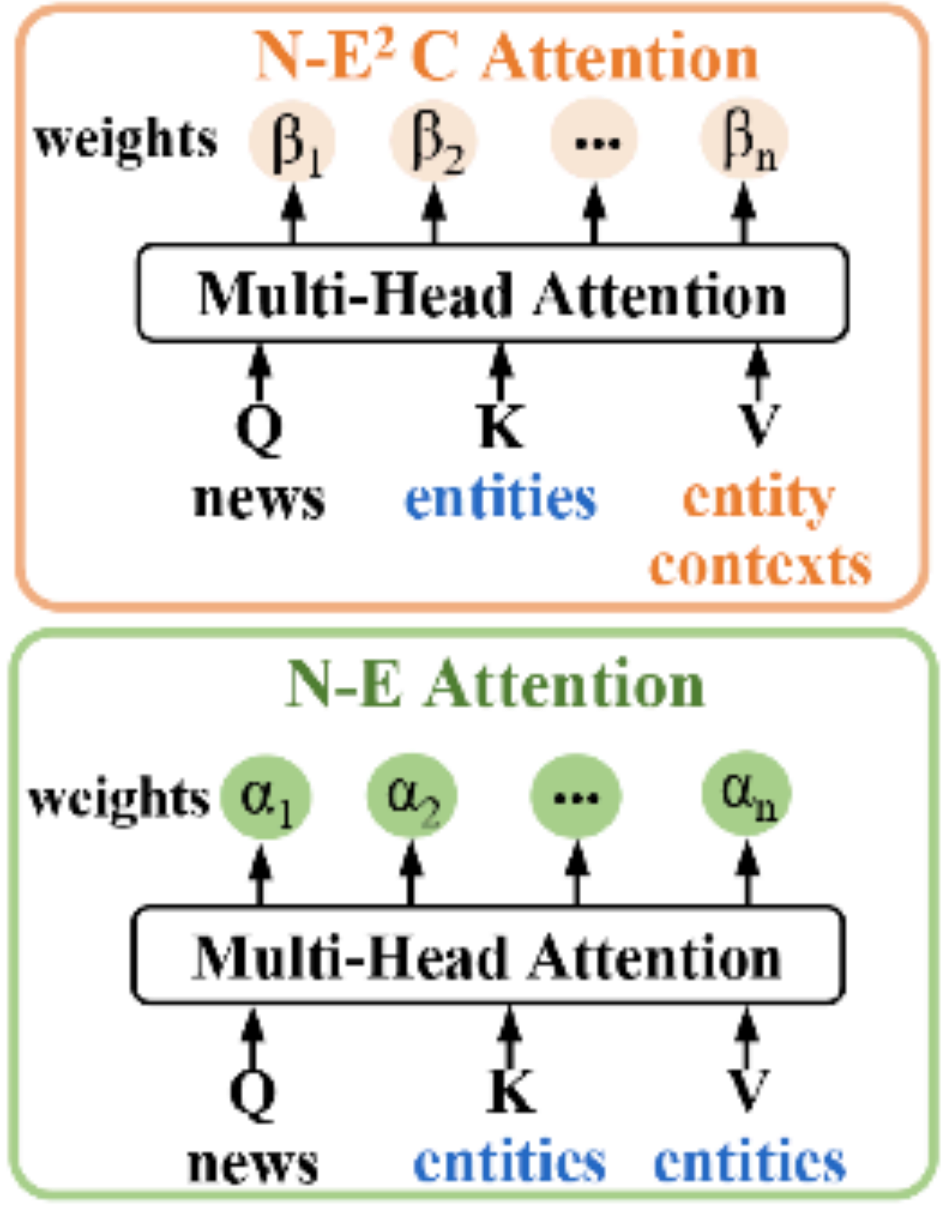
- Entity context embedding ec_i' is calculated as the average of its context entities:
  - $$ec_i' = \frac{1}{|ec(e_i)|} \sum_{e_t \in ec(e_i)} e_t'$$
  - e_t': entity embedding
  - $ec(e_i)$: neighbors entities set with one hop distance of e_i in knowledge graph.
- After the acquisition of the embeddings of entities and entity contexts, encode each of them with a transformer encoder and take the outputs q' and r' as the intermediate encoding of entity and entity contexts.

Proposed Model

Knowledge-aware Attention

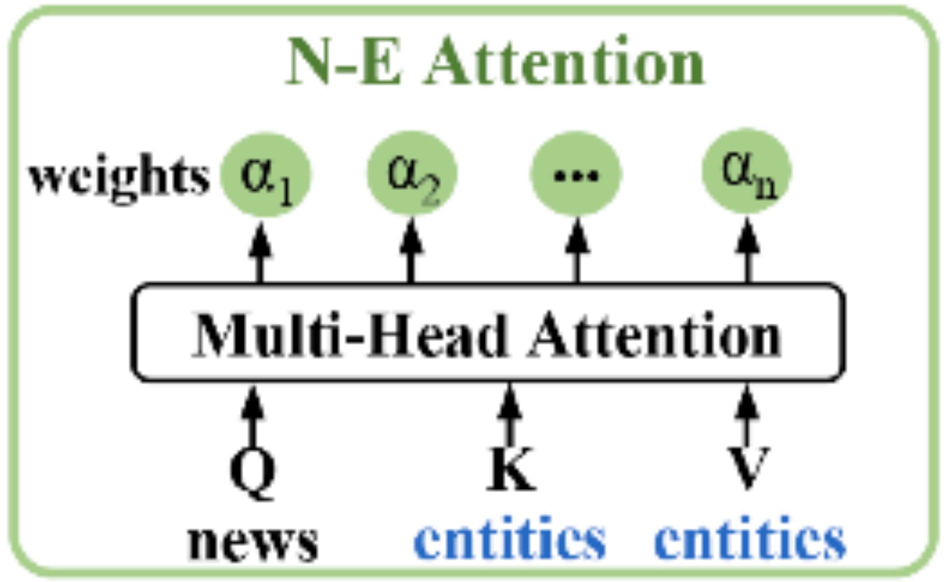


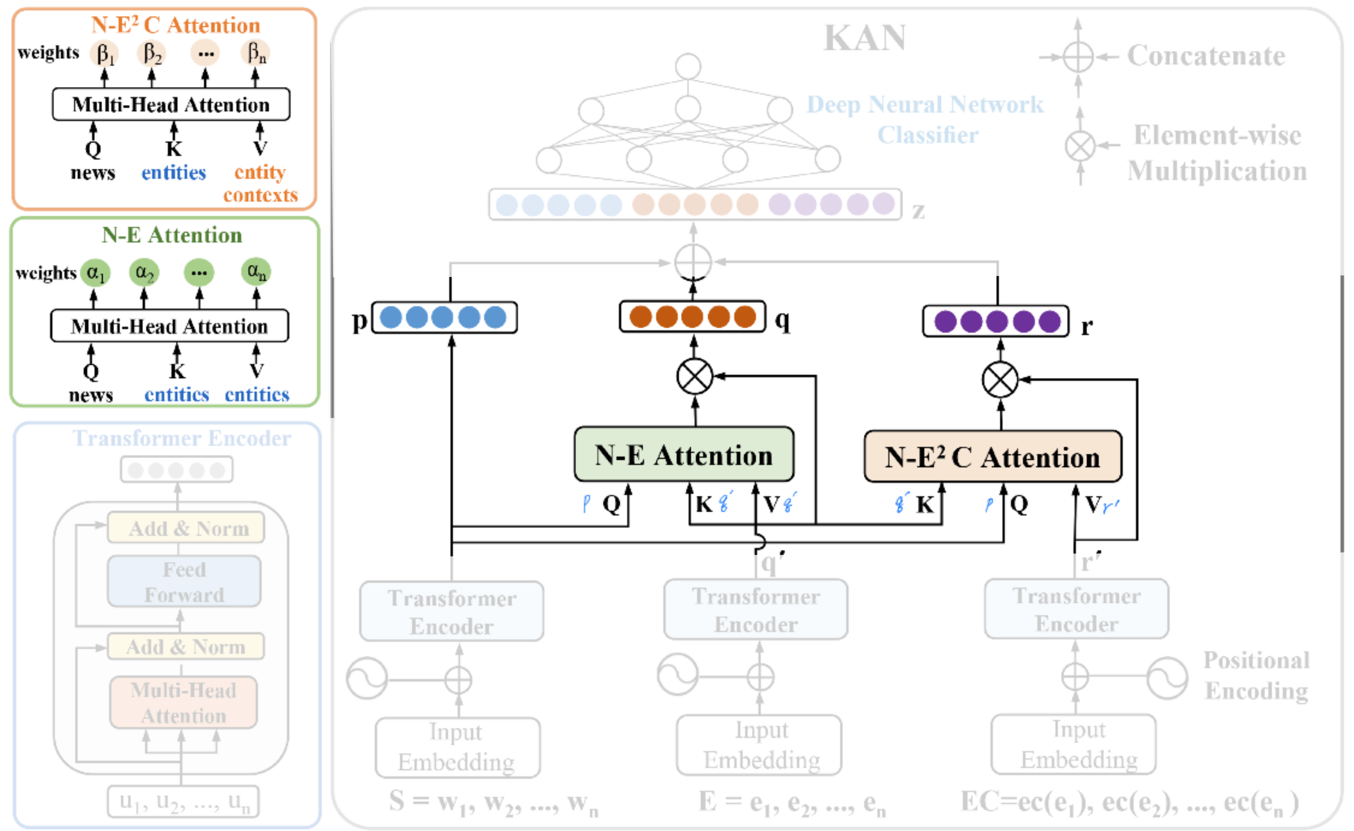
- The external knowledge obtained from knowledge graph provides rich information to help detect the class labels for news.
- To characterize the relative importance of external knowledge, design two attention networks based on multi-head attention.
  - Allows the model to consider information from different representation subspaces at different positions.



Proposed Model

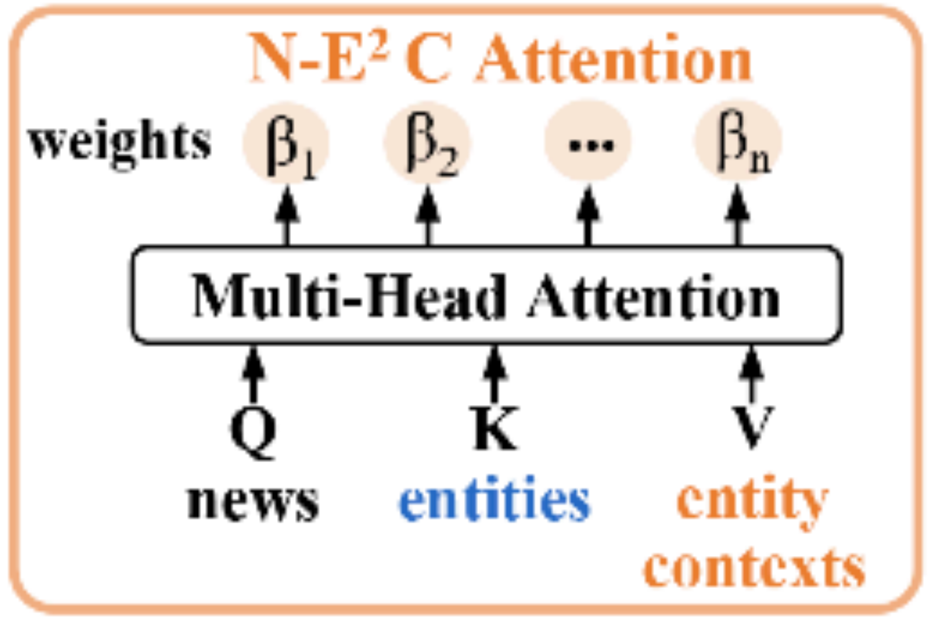
Knowledge-aware Attention

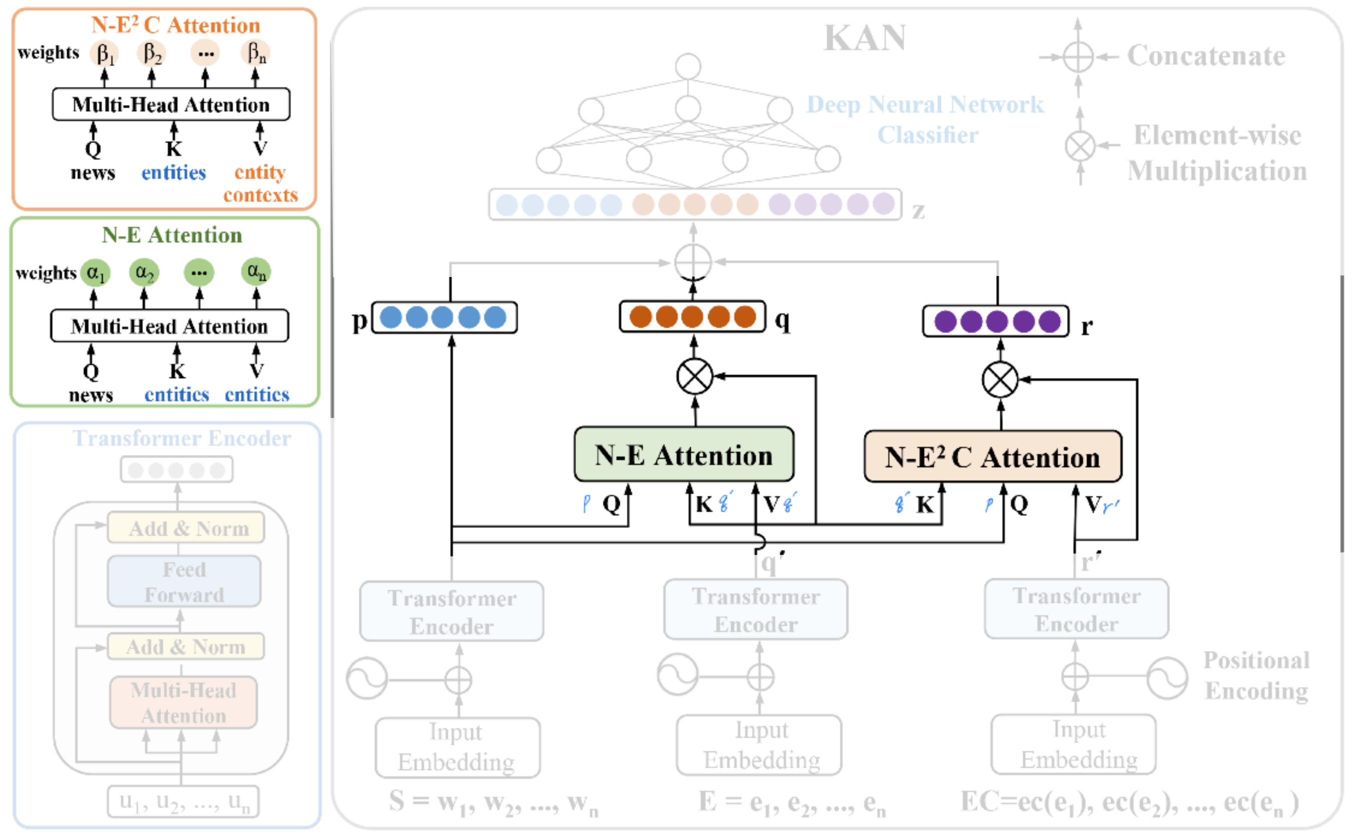




- Since not all entities contribute equally to the meaning of a news, design News towards Entities (N-E) attention to measure the importance of each entity with respect to the news content.
  - Queries: news representation p, Key and values: entities intermediate encoding q'
- By calculating the semantic similarity between news and its corresponding entities, each entity is assigned a weight α_i to represent its importance:
  - $Q = W_Q\boldsymbol{p}, K = W_K\boldsymbol{q}', V = W_V\boldsymbol{q}', \boldsymbol{\alpha} = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right), \boldsymbol{q} = \boldsymbol{\alpha}V$

Proposed Model
Knowledge-aware Attention



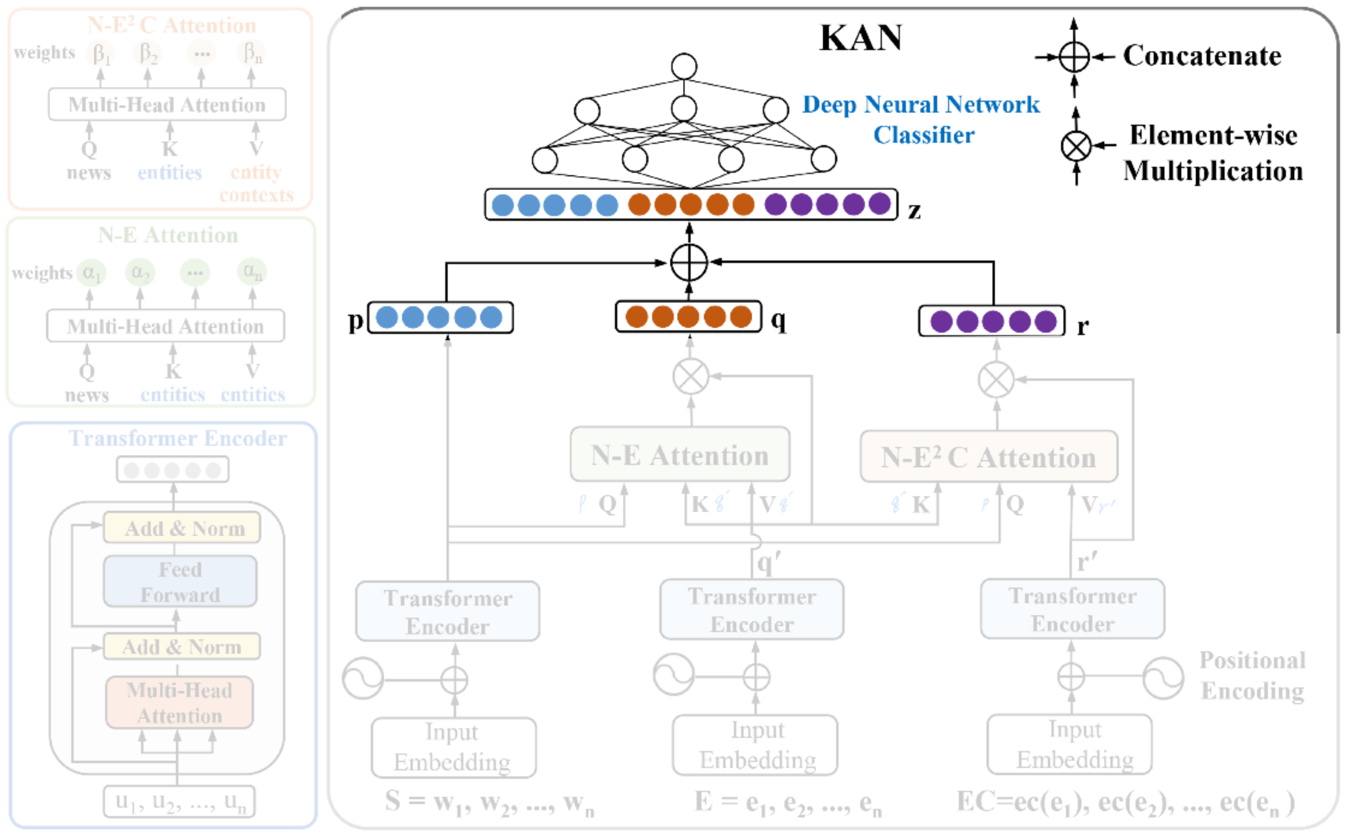


- In order to take into account the relative importance of entity contexts, propose News towards Entities and Entity Contexts (N-E²C) attention to measure the importance of each entity context according to news and its entities.

  - Queries: news representation p, keys: entities intermediate encoding q', values: entity context intermediate encoding r'

- Through calculating the semantic similarity between news and its corresponding entities, the weight β_i is assigned to each entity context according to the vitality of the corresponding entity:

  - $Q = W_Q \boldsymbol{p}, K = W_K \boldsymbol{q}', V = W_V \boldsymbol{r}', \boldsymbol{\beta} = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right), \boldsymbol{r} = \boldsymbol{\beta} V$



Proposed Model

Deep Neural Network Classifier

- Final representation of news $z = \boldsymbol{p} \oplus \boldsymbol{q} \oplus \boldsymbol{r}$.
- After that, z is fed into a fully connected layer followed by a softmax function to predict the distribution $\boldsymbol{P}$ over news labels on the target:
  - $\boldsymbol{P} = \text{softmax}(\boldsymbol{W_o} z + \boldsymbol{b_o})$
- It's trained to minimize the cross entropy loss function:
  - $J = -\sum_{i \in D} \log \boldsymbol{P}_i\left(c_i\right) + \frac{\lambda}{2} \|\Theta\|_2^2$

Experiments

Dataset & Experimental Setup

Statistic	PolitiFact	GossipCop	PHEME
# True news	443	4219	1886
# Fake news	372	3393	856
# Total news	815	7612	2742
avg.# words per news	1427	705	410
avg.# entities per news	55	36	20

- PolitiFact & GossipCop of FakeNewsNet
- PHEME
- Knowledge Extraction: use entity linking tools TagMe to disambiguate entity mentions in news contents and link them to corresponding entities in the knowledge graph Wikidata. In the procedure of entity contexts extraction, retrieve neighbors entities from Wikidata.
- Hold out 10% of dataset for validation, rest of the dataset conduct 5-fold cross-validation.

Experiments

Baselines

- SVM: is utilized to detect fake news based on features extracted from the news.
- RFC: random forest classifier using identified characteristics of news to detect whether news is fake or true.
- DTC: the news information credibility model using decision tree classifier based on various hand-crafted feature.
- GRU – 2: model based on GRUs by adding a second GRU layer that captures higher level feature interaction between different time steps.
- B – TransE: model combines positive and negative single models to detect fake news based on news content and knowledge graphs.
- KCNN: model that utilize CNN to learn representation of news. Consist of three parts: news embeddings, entity embeddings and contexts embeddings.

Experiments

Result and Analysis : Existing Methods

Datasets	Metric	SVM	RFC	DTC	GRU-2	B-TransE	KCNN	KAN
PolitiFact	Precision	0.746	0.7470	0.7476	0.7083	0.7739	0.7852	0.8687
	Recall	0.6826	0.7361	0.7454	0.7048	0.7658	0.7824	0.8499
	F1	0.6466	0.7362	0.7450	0.7041	0.7641	0.7804	0.8539
	Accuracy	0.6694	0.7406	0.7486	0.7109	0.7694	0.7827	0.8586
	AUC	0.6826	0.8074	0.7454	0.7896	0.834	0.8488	0.9197
GossipCop	Precision	0.7493	0.7015	0.6921	0.7176	0.7369	0.7483	0.7764
	Recall	0.6254	0.6707	0.6922	0.7079	0.733	0.7422	0.7696
	F1	0.5955	0.6691	0.6919	0.7079	0.734	0.7433	0.7713
	Accuracy	0.6643	0.6918	0.6959	0.718	0.7394	0.7491	0.7766
	AUC	0.6253	0.7389	0.6929	0.7516	0.7995	0.8125	0.8435
PHEME	Precision	0.7357	0.6602	0.648	0.7003	0.6834	0.6832	0.7593
	Recall	0.6116	0.6090	0.6541	0.6901	0.6061	0.6419	0.7437
	F1	0.6120	0.6138	0.6499	0.6917	0.6074	0.6489	0.7461
	Accuracy	0.7379	0.7128	0.6909	0.7371	0.72	0.7265	0.783
	AUC	0.6115	0.6833	0.6541	0.7552	0.7278	0.745	0.8373

- For content-based methods, such as SVM, RFC, DTC, GRU-2.
- SVM performs the worst among all the methods.
- DTC and RFC don't achieve good performance on three datasets.
- This's because they are built with such hand-crafted features or rules that are inferior to latent features learned by deep learning methods.

Experiments

Result and Analysis
: Existing Methods

Datasets	Metric	SVM	RFC	DTC	GRU-2	B-TransE	KCNN	KAN
PolitiFact	Precision	0.746	0.7470	0.7476	0.7083	0.7739	0.7852	0.8687
	Recall	0.6826	0.7361	0.7454	0.7048	0.7658	0.7824	0.8499
	F1	0.6466	0.7362	0.7450	0.7041	0.7641	0.7804	0.8539
	Accuracy	0.6694	0.7406	0.7486	0.7109	0.7694	0.7827	0.8586
	AUC	0.6826	0.8074	0.7454	0.7896	0.834	0.8488	0.9197
GossipCop	Precision	0.7493	0.7015	0.6921	0.7176	0.7369	0.7483	0.7764
	Recall	0.6254	0.6707	0.6922	0.7079	0.733	0.7422	0.7696
	F1	0.5955	0.6691	0.6919	0.7079	0.734	0.7433	0.7713
	Accuracy	0.6643	0.6918	0.6959	0.718	0.7394	0.7491	0.7766
	AUC	0.6253	0.7389	0.6929	0.7516	0.7995	0.8125	0.8435
PHEME	Precision	0.7357	0.6602	0.648	0.7003	0.6834	0.6832	0.7593
	Recall	0.6116	0.6090	0.6541	0.6901	0.6061	0.6419	0.7437
	F1	0.6120	0.6138	0.6499	0.6917	0.6074	0.6489	0.7461
	Accuracy	0.7379	0.7128	0.6909	0.7371	0.72	0.7265	0.783
	AUC	0.6115	0.6833	0.6541	0.7552	0.7278	0.745	0.8373

- GRU-2 performs better than hand-crafted models in GossipCop and PHEME.
  - Superiority of feature extraction of deep neural networks.
- However, GRU-2 achieve slightly lower results on PolitiFact.
  - This's probably because GRU-2 is limited to deal with long sentences in dataset.

Experiments

Result and Analysis : Existing Methods

Datasets	Metric	SVM	RFC	DTC	GRU-2	B-TransE	KCNN	KAN
PolitiFact	Precision	0.746	0.7470	0.7476	0.7083	0.7739	0.7852	**0.8687**
	Recall	0.6826	0.7361	0.7454	0.7048	0.7658	0.7824	**0.8499**
	F1	0.6466	0.7362	0.7450	0.7041	0.7641	0.7804	**0.8539**
	Accuracy	0.6694	0.7406	0.7486	0.7109	0.7694	0.7827	**0.8586**
	AUC	0.6826	0.8074	0.7454	0.7896	0.834	0.8488	**0.9197**
GossipCop	Precision	0.7493	0.7015	0.6921	0.7176	0.7369	0.7483	**0.7764**
	Recall	0.6254	0.6707	0.6922	0.7079	0.733	0.7422	**0.7696**
	F1	0.5955	0.6691	0.6919	0.7079	0.734	0.7433	**0.7713**
	Accuracy	0.6643	0.6918	0.6959	0.718	0.7394	0.7491	**0.7766**
	AUC	0.6253	0.7389	0.6929	0.7516	0.7995	0.8125	**0.8435**
PHEME	Precision	0.7357	0.6602	0.648	0.7003	0.6834	0.6832	**0.7593**
	Recall	0.6116	0.6090	0.6541	0.6901	0.6061	0.6419	**0.7437**
	F1	0.6120	0.6138	0.6499	0.6917	0.6074	0.6489	**0.7461**
	Accuracy	0.7379	0.7128	0.6909	0.7371	0.72	0.7265	**0.783**
	AUC	0.6115	0.6833	0.6541	0.7552	0.7278	0.745	**0.8373**

- Using both news content and external knowledge achieve consistently better results than the methods which are purely based on news contents.
  - KAN > KCNN > B-TransE > GRU-2, SVM, RFC, DTC
- Indicates that models can successfully incorporate the external knowledge and significantly boost the detection performance.

Experiments

Result and Analysis
: Existing Methods

Datasets	Metric	SVM	RFC	DTC	GRU-2	B-TransE	KCNN	KAN
PolitiFact	Precision	0.746	0.7470	0.7476	0.7083	0.7739	0.7852	**0.8687**
	Recall	0.6826	0.7361	0.7454	0.7048	0.7658	0.7824	**0.8499**
	F1	0.6466	0.7362	0.7450	0.7041	0.7641	0.7804	**0.8539**
	Accuracy	0.6694	0.7406	0.7486	0.7109	0.7694	0.7827	**0.8586**
	AUC	0.6826	0.8074	0.7454	0.7896	0.834	0.8488	**0.9197**
GossipCop	Precision	0.7493	0.7015	0.6921	0.7176	0.7369	0.7483	**0.7764**
	Recall	0.6254	0.6707	0.6922	0.7079	0.733	0.7422	**0.7696**
	F1	0.5955	0.6691	0.6919	0.7079	0.734	0.7433	**0.7713**
	Accuracy	0.6643	0.6918	0.6959	0.718	0.7394	0.7491	**0.7766**
	AUC	0.6253	0.7389	0.6929	0.7516	0.7995	0.8125	**0.8435**
PHEME	Precision	0.7357	0.6602	0.648	0.7003	0.6834	0.6832	**0.7593**
	Recall	0.6116	0.6090	0.6541	0.6901	0.6061	0.6419	**0.7437**
	F1	0.6120	0.6138	0.6499	0.6917	0.6074	0.6489	**0.7461**
	Accuracy	0.7379	0.7128	0.6909	0.7371	0.72	0.7265	**0.783**
	AUC	0.6115	0.6833	0.6541	0.7552	0.7278	0.745	**0.8373**

- For using both news contents and knowledge methods, KAN achieves better performance than KCNN and B-TransE.
- KAN consistently outperform KCNN on three datasets.
  - F1 score (PolitiFact/GossipCop/PHEME): 7.4% ↑/ 2.8% ↑/ 9.7% ↑
  - Accuracy (PolitiFact/GossipCop/PHEME): 7.6% ↑/ 2.8% ↑/ 5.7% ↑

Experiments

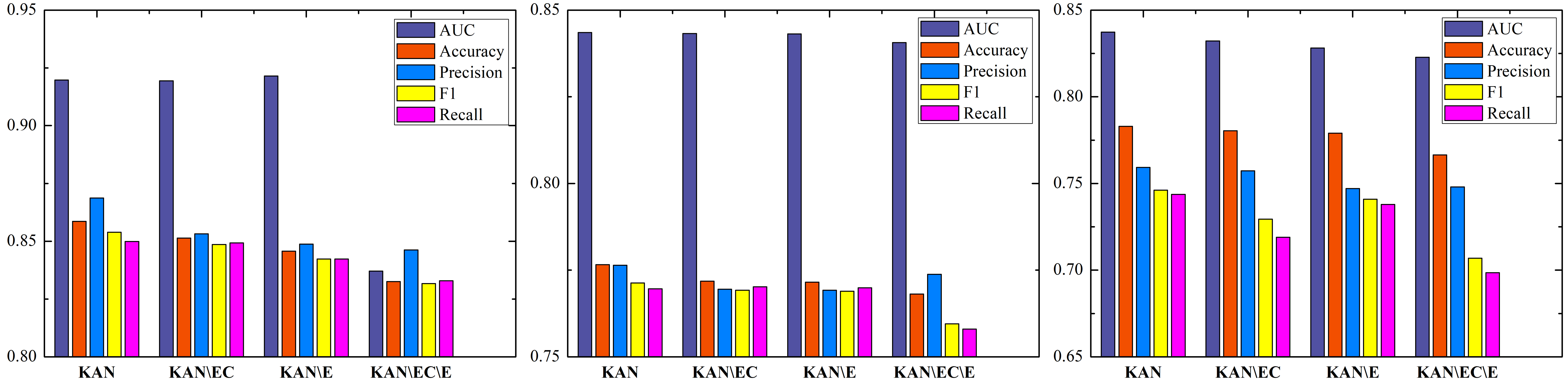
Result and Analysis : Existing Methods

Datasets	Metric	SVM	RFC	DTC	GRU-2	B-TransE	KCNN	KAN
PolitiFact	Precision	0.746	0.7470	0.7476	0.7083	0.7739	0.7852	**0.8687**
	Recall	0.6826	0.7361	0.7454	0.7048	0.7658	0.7824	**0.8499**
	F1	0.6466	0.7362	0.7450	0.7041	0.7641	0.7804	**0.8539**
	Accuracy	0.6694	0.7406	0.7486	0.7109	0.7694	0.7827	**0.8586**
	AUC	0.6826	0.8074	0.7454	0.7896	0.834	0.8488	**0.9197**
GossipCop	Precision	0.7493	0.7015	0.6921	0.7176	0.7369	0.7483	**0.7764**
	Recall	0.6254	0.6707	0.6922	0.7079	0.733	0.7422	**0.7696**
	F1	0.5955	0.6691	0.6919	0.7079	0.734	0.7433	**0.7713**
	Accuracy	0.6643	0.6918	0.6959	0.718	0.7394	0.7491	**0.7766**
	AUC	0.6253	0.7389	0.6929	0.7516	0.7995	0.8125	**0.8435**
PHEME	Precision	0.7357	0.6602	0.648	0.7003	0.6834	0.6832	**0.7593**
	Recall	0.6116	0.6090	0.6541	0.6901	0.6061	0.6419	**0.7437**
	F1	0.6120	0.6138	0.6499	0.6917	0.6074	0.6489	**0.7461**
	Accuracy	0.7379	0.7128	0.6909	0.7371	0.72	0.7265	**0.783**
	AUC	0.6115	0.6833	0.6541	0.7552	0.7278	0.745	**0.8373**

- Attribute the superiority of KAN to two reasons:
  - KAN uses the knowledge-aware network which can eliminate the ambiguity caused by the entity mentions in the news and learn knowledge-level connections among news entities.
  - KAN employs the attention network which can measure the importances of entity and entity context knowledge and effectively fuse them into news representation.

Experiments

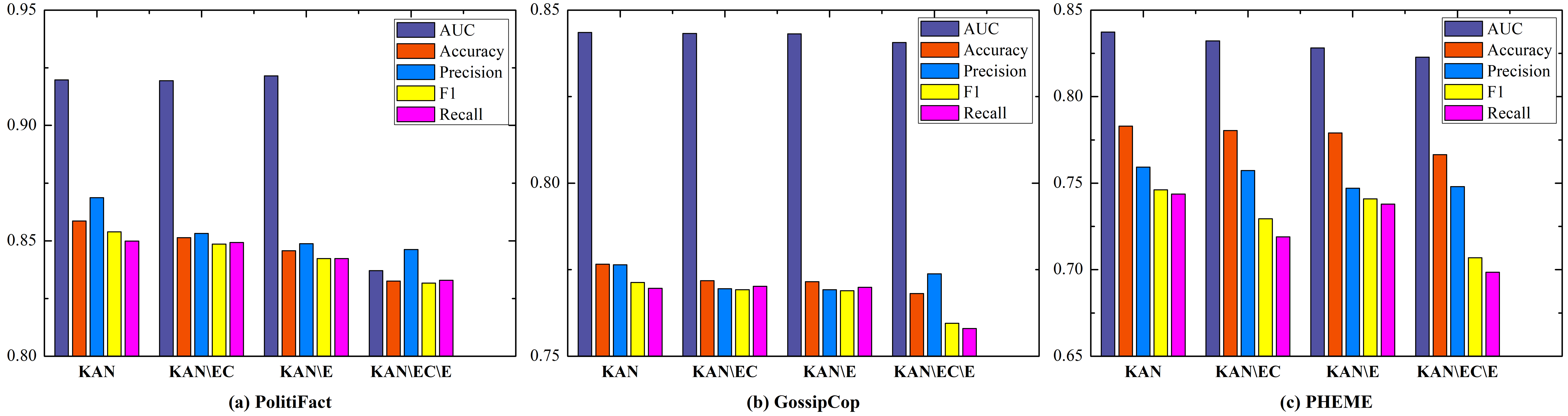
Result and Analysis
: KAN variants



(a) PolitiFact (b) GossipCop (c) PHEME

- **KAN**: proposed model.
- **KAN/EC**: without entity contexts sequence when information is fed into the model.
- **KAN/E**: without entities sequence when information is fed into the model.
- **KAN/EC/E**: only detect fake news by news contents.

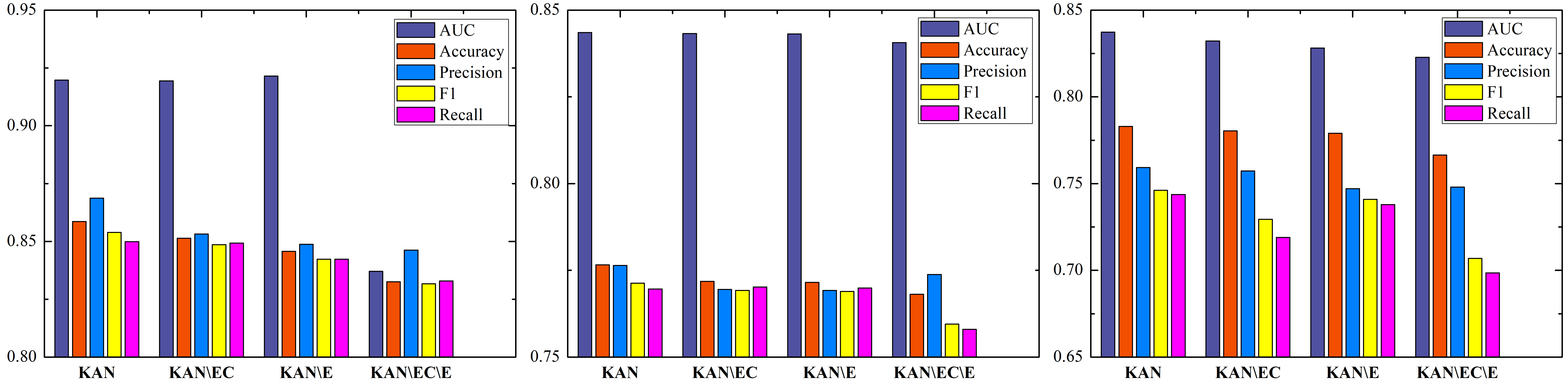
Experiments
Result and Analysis
: KAN variants



- When eliminate the entity contexts knowledge, the result are reduced.
  - Suggests that the comprehensive information of entity contexts is helpful for understanding entities in news.
- When disregard the entities sequence, the performance of **KAN/E** degrades in comparison with **KAN** on three datasets.
  - Suggest the entities play an important role in disambiguation of entity mentions in the news, also provides the basis for effectively incorporating entity contexts.

Experiments

Result and Analysis
: KAN variants

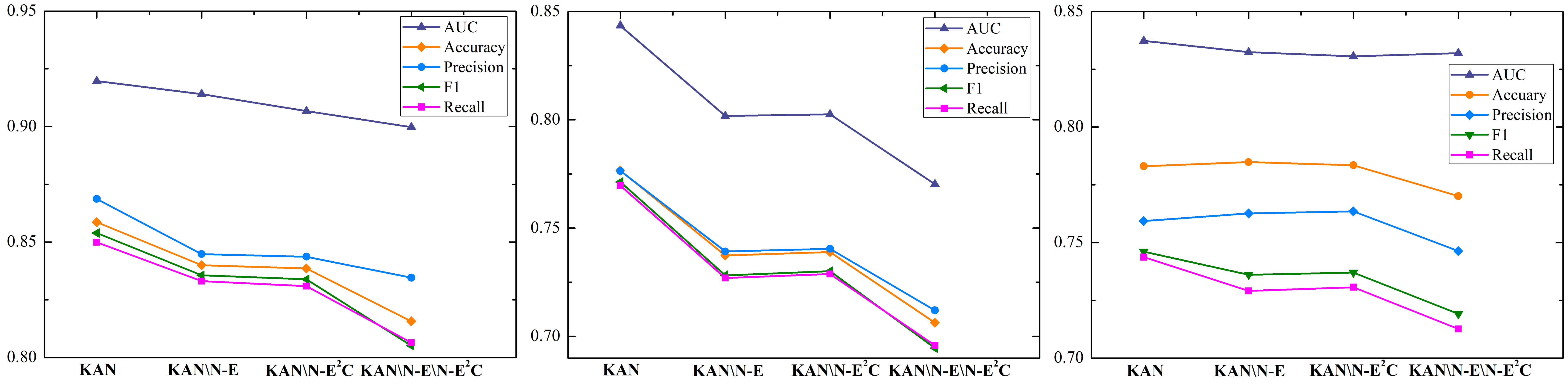


(a) PolitiFact (b) GossipCop (c) PHEME

- When the external knowledge is removed from KAN, the results of **KAN/EC/E** degrade in comparison with KAN in terms of F1 scores.
    - Performance (PolitiFact/GossipCop/PHEME): 2.2% ↓/ 1.2% ↓/ 1.3% ↓
    - Suggest the importance to consider knowledge of news to guide fake news detection in KAN.

Experiments

Result and Analysis : KAN variants

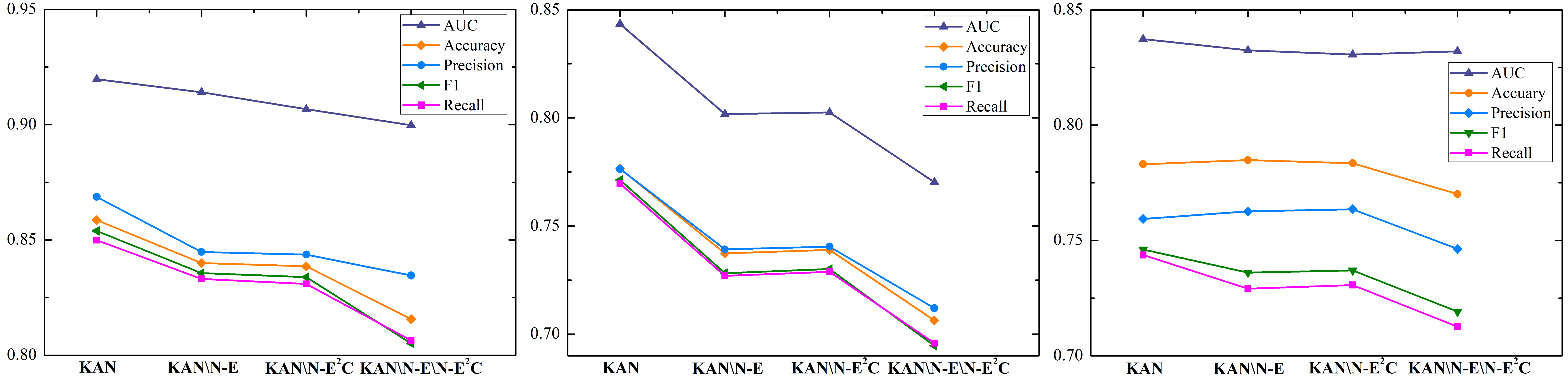


(a) Politifact (b) GossipCop (c) PHEME

- $\mathbf{KAN}$: proposed model.
- $\mathbf{KAN/N-E}$: without counting News contents towards Entities (N-E) attention.
- $\mathbf{KAN/N-E^2C}$: without considering News contents towards Entities and Entity Contexts (N-E^{2}C) attention.
- $\mathbf{KAN/N-E/N-E^2C}$: eliminates both N-E attention and N-E^{2}C attention.

Experiments

Result and Analysis
: KAN variants



(a) Politifact **(b) GossipCop** **(c) PHEME**

- Usage of N-E attention and N-E^{2}C attention can improve performance respectively, and can achieve even better performance by using them together.

- Result of using N-E attention and N-E^{2}C attention together improved
  - 2.2% on PolitiFact, 6.2% on GossipCop in terms of Accuracy.
  - Validate the effectiveness of proposed attention mechanisms.

Conclusion and Future work

- This work attempts to incorporate entities and entity context knowledge from knowledge graph for fake news detection.
- Propose Knowledge-aware Attention Network that effectively integrates the two kinds of knowledge with news through attention mechanisms.
- For future work, the authors will search for better representation form of knowledge to incorporate it into neural networks as explicit features to further boost fake news detection performance.

Comments
of KAN

- Knowledge graph limited problem (like OOV).
  - Some newly-emergence event cannot derive from knowledge graph.
- May can try to find other complement information to learn comprehensive news representation.
  - Like user profile information, news image…etc.
- Multi-head attention also can employ to other complement informations.