Shufeng XIONG, Guipei ZHANG, Xiaobo FAN, et al., 2025. MAL: multilevel active learning with BERT for Chinese textual affective structure analysis. *Frontiers of Information Technology & Electronic Engineering,* 26(6):833-846. https://doi.org/10.1631/FITEE.2400242

MAL: multilevel active learning with BERT for Chinese textual affective structure analysis

Key words: Sentiment analysis; Sequence labeling; Active learning (AL); Bidirectional encoder representations from Transformers (BERT)

Haiping SI
E-mail: haiping@henau.edu.cn
ORCID: https://orcid.org/0000-0001-8430-149X

Motivation

1. Chinese textual affective structure analysis (CTASA) is a complex and costly task: CTASA requires identifying eight sentiment-related components (e.g., cause, trigger, and holder), making manual annotation time-consuming and resource-intensive.

2. Limitations of existing active learning (AL) strategies: most AL methods rely solely on sentence-level uncertainty or word-level diversity, which fails to capture the linguistic complexity of Chinese.

3. Diverse and informal expressions in Chinese social media: for example, the sentence “内实在太痛苦鸟!” contains non-standard expressions for “那” and “了.” Capturing such variation requires strategies that consider both semantic structure and lexical diversity.

Main idea

1. We propose a novel multilevel active learning (MAL) strategy that integrates sentence-level semantic similarity and word-level label divergence to select valuable samples for annotation.

2. Key elements:

(1) Sentence-level features extracted from bidirectional encoder representations from Transformers (BERT) ([CLS] embeddings) to evaluate structural similarity.

(2) Word-level probability distributions generated by a conditional random field (CRF) layer, compared using cross-KL divergence.

3. By jointly considering these two levels, MAL enables more precise and informative sample selection to improve performance with fewer labeled data.

Method

The proposed MAL consists of four modules: feature extraction module, filter module, word-level divergence calculator module, and update module.

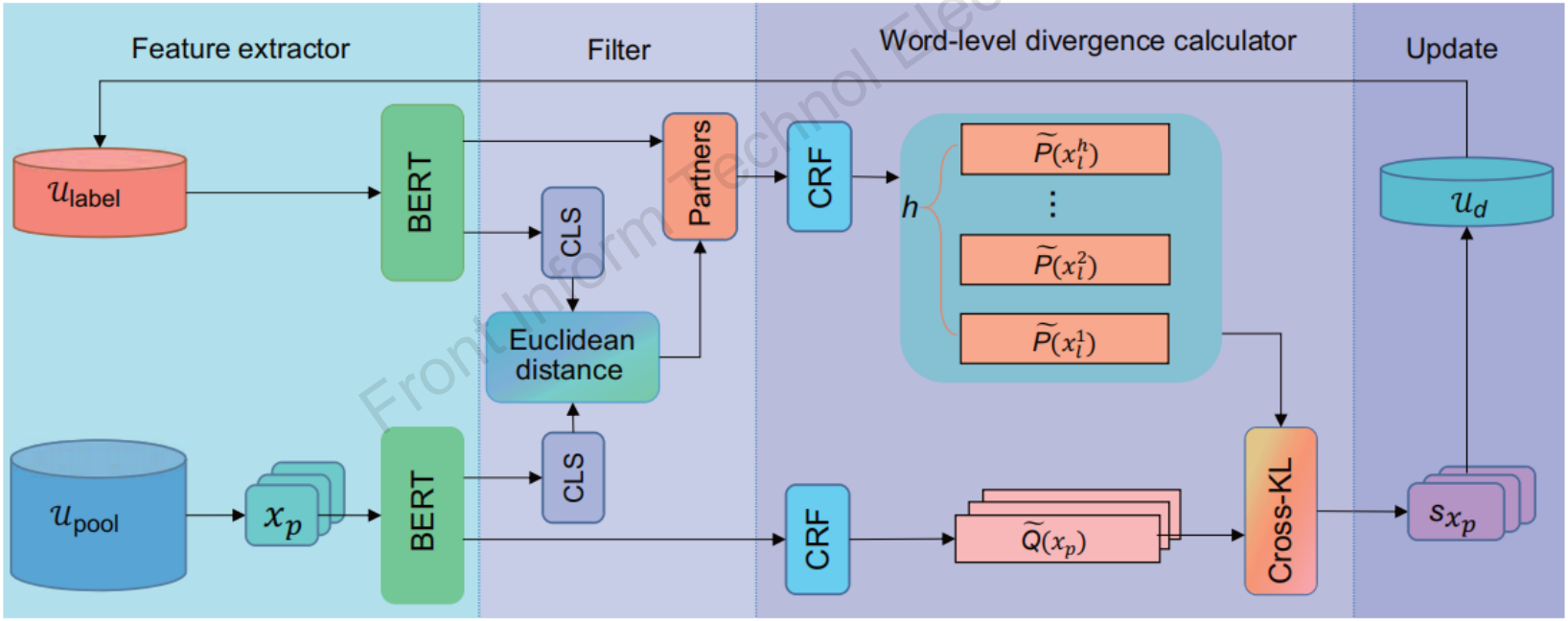

Fig. 1 Structure of multilevel active learning (MAL). BERT: bidirectional encoder representations from Transformers; CLS: classification; CRF: conditional random field; KL: Kullback–Leibler

Method (Cont'd)

1. Feature extraction module

Use BERT [CLS] embedding to represent sentence-level semantics:

$$\text{Feature}_{\text{sentence}} = \text{BERT}_{[\text{CLS}]}$$

2. Filter module

Calculate Euclidean distance between unlabeled sample x_p and labeled samples x_l:

$$\text{Distance}(s_1, s_2) = \sqrt{\sum_{i=1}^{n} (s_{1i} - s_{2i})^2}$$

Select h nearest neighbors (partners) for comparison.

Method (Cont'd)

3. Word-level divergence calculator module

Compute cross-KL divergence between label distributions of x_p and its h partners:

$$\text{cross-KL}(P\|Q) = \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}\text{KL}(p_i\|q_j)$$

Capture token-level structural differences.

Method (Cont'd)

4. Update module

For each x_p, compute score:

$$s_{x_p} = \sum_{k=1}^{h} \text{cross-KL}(P^{(x_l^k)} \| Q^{(x_p)})$$

Select *h* lowest-scoring samples for annotation and update the training set.

Major results

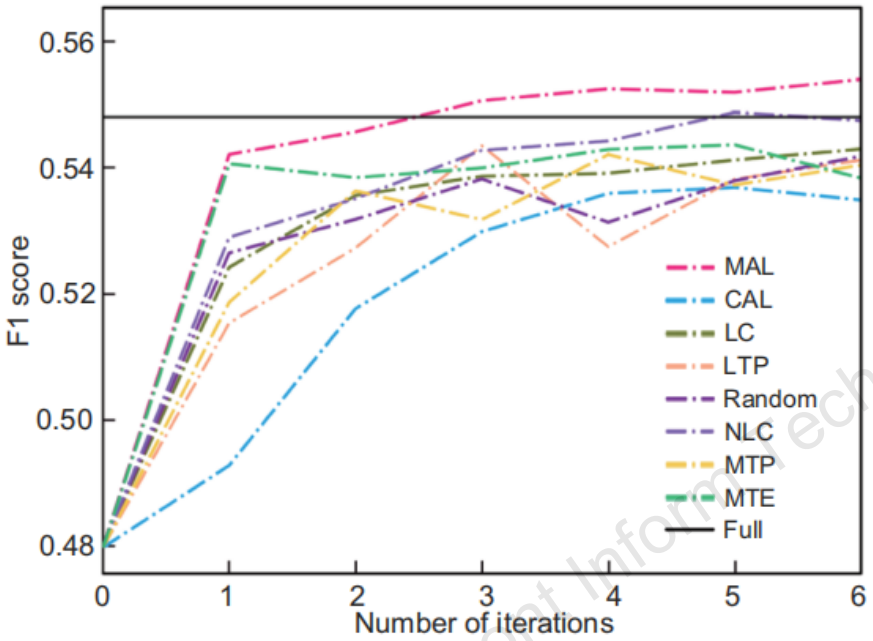

Fig. 2 Performance of MAL and baselines on the CTAS dataset. CTAS: Chinese textual affective structure; MAL: multilevel active learning; CAL: contrastive active learning; LC: least confidence; LTP: lowest token probability; Random: select samples in a random manner; NLC: normalized LC; MTP: minimum token probability; MTE: maximum token entropy; Full: training model with all of the available data. References to color refer to the online version of this figure

MAL achieves comparable performance to or better performance than the fully supervised model after just three iterations.

Major results (Cont'd)

Filter module effectiveness

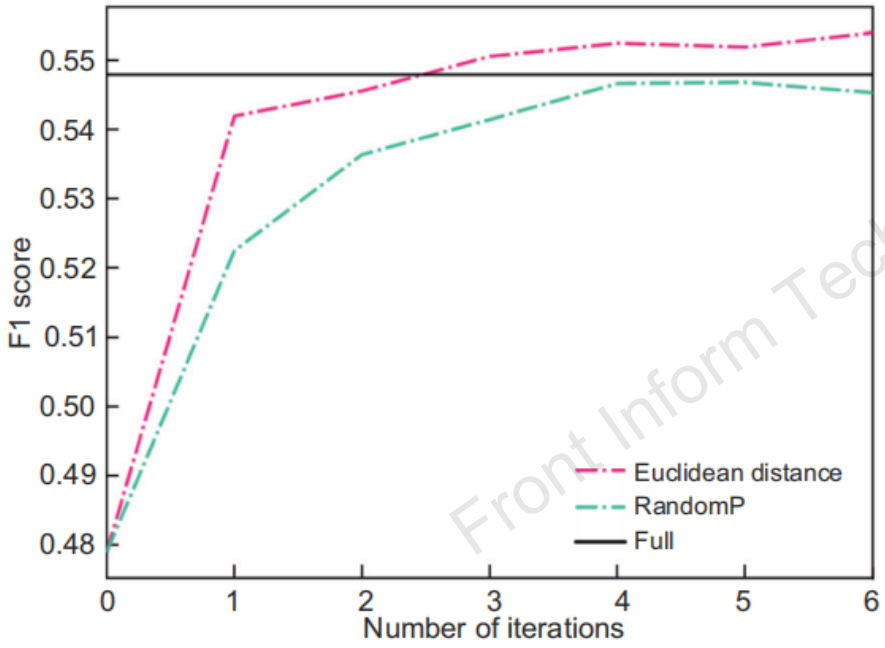

Fig. 3 Impact of partner selection on the performance, where RandomP refers to selecting partners in a random manner. References to color refer to the online version of this figure

Number of partners

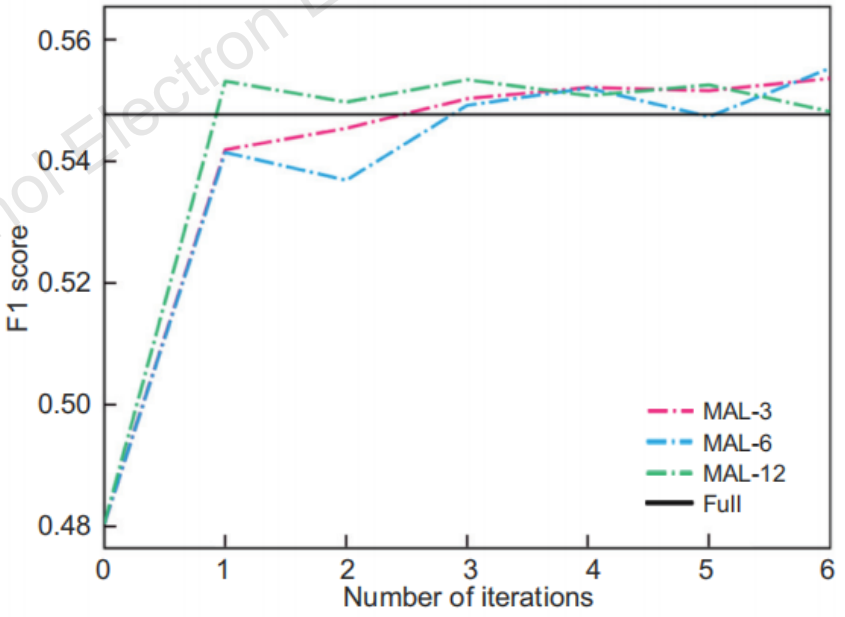

Fig. 4 Impact of the number of partners on performance. MAL-3, MAL-6, and MAL-12: multilevel active learning method with 3, 6, and 12 partners, respectively. References to color refer to the online version of this figure

Major results (Cont'd)

Label distribution analysis

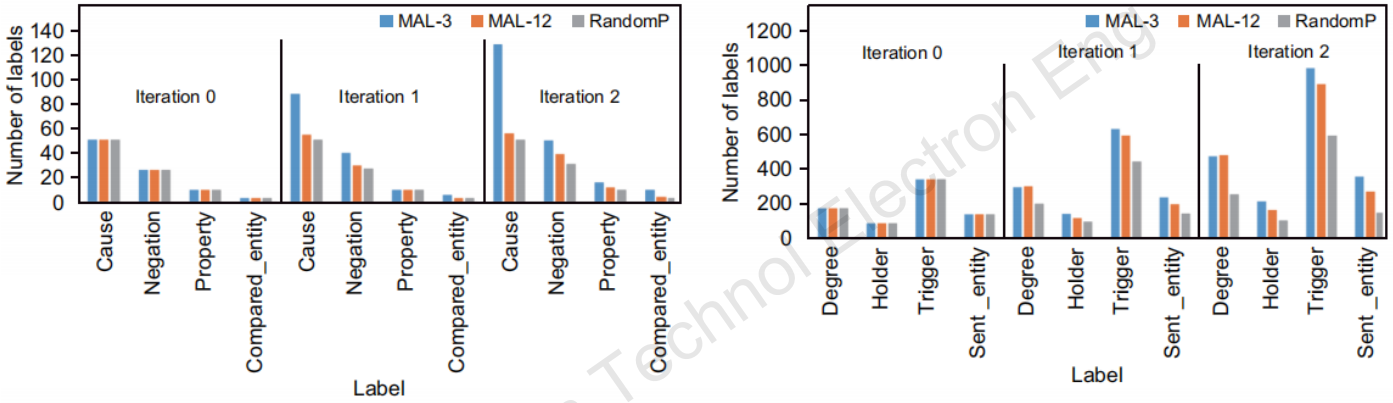

Fig. 5 Distribution trends of different labels in selected samples when using different methods. RandomP refers to selecting partners in a random manner. MAL: multilevel active learning. References to color refer to the online version of this figure

Table 2 Number of labels in the training set

Label	Number of labels								
	Initial	LC	NLC	MTP	MTE	LTP	RandomP	CAL	MAL
Cause	46	73	203	104	73	155	93	199	70
Degree	211	397	431	426	383	476	413	490	327
Holder	86	175	204	194	159	204	187	217	129
Negation	19	39	38	47	37	39	33	45	27
Property	10	22	18	27	21	26	20	30	13
Trigger	373	744	707	809	729	752	750	826	626
Compared_entity	3	4	9	6	3	6	6	12	6
Sent_entity	134	255	257	288	238	326	265	405	204

CTAS: Chinese textual affective structure; MAL: multilevel active learning; CAL: contrastive active learning; LC: least confidence; LTP: lowest token probability; NLC: normalized LC; MTP: minimum token probability; MTE: maximum token entropy

Conclusions

1. We introduce MAL, a multilevel active learning framework for CTASA.

2. By combining sentence-level representations from BERT with word-level uncertainty from CRF predictions, MAL achieves superior performance with minimal annotation.

3. This is the first AL strategy tailored for CTASA, offering a scalable, cost-effective solution for sentiment structure tasks.

4. Future work includes combining MAL with stronger pretrained models and adapting it to other sequence labeling scenarios under low-resource constraints.

Shufeng Xiong is recognized as a Young Backbone Teacher of Henan Province, a Top Talent at Henan Agricultural University, and a Young Science and Technology Expert of Pingdingshan City. His primary research spans natural language processing, multimodal fusion analysis, agricultural information mining, food safety information monitoring, and industrial applications of machine learning.

Haiping Si specializes in computer science applied to agricultural informatics. As a principal investigator (PI), he has spearheaded over 10 national and international projects, including EU-China collaborations on molecular breeding computation, China's National Spark Program for germplasm information systems, and major Henan Province initiatives on grain traceability and crop resource sharing. He has published extensively on agricultural data intelligence, with key works in top journals: blockchain-based IoT security (*Fut Gener Comput Syst*), real-time harvester scheduling algorithms (*Comput Electron Agric*). His contributions advance digital solutions for crop germplasm management and agricultural traceability. His research focuses on big data analytics, data visualization, crop information science, and software engineering.