An LLM-Enabled Knowledge Elicitation and Retrieval Framework for Zero-Shot Cross-Lingual Stance Identification

Ruike Zhang[1,2], **Yuan Tian**[1,2], **Penghui Wei**[1,2*], **Dajun Zeng**[1,2], **Wenji Mao**[1,2*]
[1]State Key Laboratory of Multimodal Artificial Intelligence Systems,
Institute of Automation, Chinese Academy of Sciences
[2]School of Artificial Intelligence, University of Chinese Academy of Sciences
{zhangruike2020,tianyuan2021,wenji.mao}@ia.ac.cn,weipenghui2016@gmail.com

Abstract

Stance detection aims to identify the attitudes toward specific targets from text, which is an important research area in text mining and social media analytics. Existing research is mainly conducted in monolingual setting on English datasets. To tackle the data scarcity problem in low-resource languages, cross-lingual stance detection (CLSD) transfers the knowledge from high-resource (source) language to low-resource (target) language. The CLSD task is the most challenging in zero-shot setting when no training data is available in target language, and transferring stance-relevant knowledge learned from high-resource language to bridge the language gap is the key for improving the performance of zero-shot CLSD. In this paper, we leverage the capability of large language model (LLM) for stance knowledge acquisition, and propose KEAR, a knowledge elicitation and retrieval framework. The knowledge elicitation module in KEAR first derives different types of stance knowledge from LLM's reasoning process. Then, the knowledge retrieval module in KEAR matches the target language input to the most relevant stance knowledge for enhancing text representations. Experiments on multilingual datasets show the effectiveness of KEAR compared with competitive baselines as well as the CLSD approaches trained with labeled data in target language[1].

1 Introduction

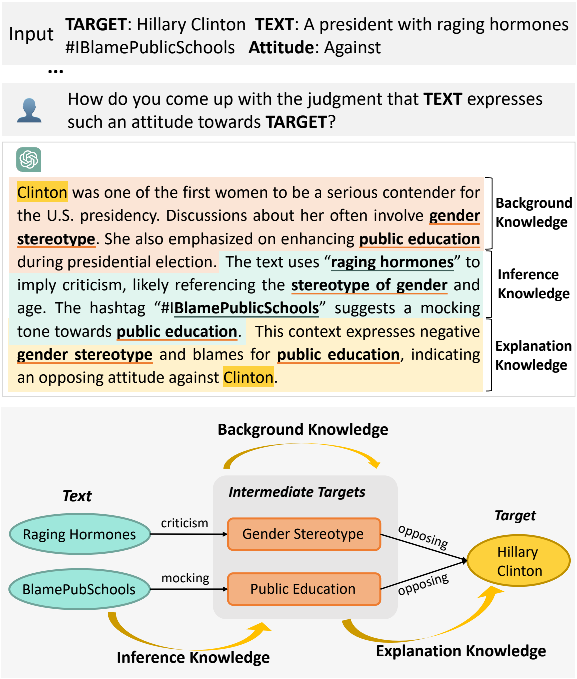
Stance detection aims to determine the attitudes (e.g., in favor of, against or neutral) toward predefined targets (e.g., entities, controversial topics or events) from a given text. It has attracted significant research attention and can facilitate critical applications such as market analysis, veracity checking and public opinion mining (Küçük and Can, 2020). Recently, a variety of monolingual methods were proposed for different settings, including *in-domain* methods (Mohammad et al., 2016; Augenstein et al., 2016), *cross-target* methods (Xu et al., 2018; Wei and Mao, 2019) and *zero-shot* methods (Allaway and Mckeown, 2020; Liang et al., 2022; Li et al., 2023). The majority of them are conducted on English datasets, whereas in other low-resource languages, it lacks sufficient data for training quality stance detection models.

To alleviate the data scarcity issue, cross-lingual stance detection (CLSD) transfers the knowledge learned from high-resource (source) language to low-resource (target) one. Recent approaches usually heavily rely on labeled or unlabeled data in target language (Mohtarami et al., 2019; Zhang et al., 2023b). In extreme data-scarce practical applications, CLSD is the most challenging in zero-shot setting. **Zero-shot cross-lingual stance detection** (Vamvas and Sennrich, 2020) aims to identify the stance toward certain targets when no training data is available (i.e., even without unlabeled training data) in target language. The state-of-the-art approach (Hardalov et al., 2022) proposes to pretrain language models with sentiment-based data and transfer the learned knowledge to target languages with prompt-based fine-tuning.

In this paper, we focus on the challenging task of zero-shot CLSD. Since there is no training data for target language, due to the unique linguistic and cultural nuances in target language, the disparity among languages cannot be overcome via training on source language data alone. In such situation, incorporating language-agnostic external knowledge for zero-shot CLSD can serve as a feasible scheme to enhance cross-lingual representations and bridge the language gap. Although existing monolingual methods have incorporated external knowledge including sentiment lexicons (Li and Caragea, 2019; Zhang et al., 2020), knowledge graph (Liu et al., 2021) and Wikipedia (He et al., 2022; Zhu et al., 2022; Li et al., 2023), they lack the proper consideration on the acquisition of stance-relevant knowl-

[*] Corresponding author
[1] Source code: https://github.com/ALUKErnel/KEAR.

Findings of the Association for Computational Linguistics: EMNLP 2024, pages 12253–12266
November 12-16, 2024 ©2024 Association for Computational Linguistics



Input TARGET: Hillary Clinton TEXT: A president with raging hormones #IBlamePublicSchools Attitude: Against

...

How do you come up with the judgment that TEXT expresses such an attitude towards TARGET?

Clinton was one of the first women to be a serious contender for the U.S. presidency. Discussions about her often involve gender stereotype. She also emphasized on enhancing public education during presidential election. The text uses "raging hormones" to imply criticism, likely referencing the stereotype of gender and age. The hashtag "#IBlamePublicSchools" suggests a mocking tone towards public education. This context expresses negative gender stereotype and blames for public education, indicating an opposing attitude against Clinton.

Figure 1: The stance inferential process with an LLM, where stance knowledge including background knowledge, inference knowledge, and explanation knowledge can be elicited to facilitate standpoint identification.

edge that reveals the implicit inferential relationships underlying stance identification.

Given that different types of knowledge are involved in identifying the stance toward certain targets, we propose to leverage the capability of large language model (LLM) for stance knowledge acquisition. Figure 1 illustrates an LLM-generated stance identification example. First, **background knowledge** facilitates stance detection by providing the knowledge on factual statements as well as intermediate entities or claims (viewed as intermediate targets) relevant to the *destination target*, i.e., "Hillary Clinton" in Figure 1. Then, on top of background knowledge, **inference knowledge** provides the knowledge on reasoning about the attitudes toward *intermediate targets* (e.g., "gender stereotype" and "public education" in Figure 1) from textual expressions. Further, on the basis of background and inference knowledge, **explanation knowledge** concludes the standpoint toward the destination target based on its intrinsic relationship with intermediate targets. As shown in Figure 1, based on the above considerations, we can better leverage the inferential capabilities of LLMs to elicit stance-relevant knowledge and facilitate standpoint identification. Such stance knowledge is largely language-agnostic, therefore it is suitable to bridge the language gap.

To this end, we propose a **K**nowledge **E**licitation **A**nd **R**etrieval (KEAR) framework for zero-shot CLSD, comprised of knowledge elicitation, verification and retrieval modules. The *knowledge elicitation module* first distinguishes different types of stance knowledge from LLM's reasoning process via intermediate target mining and speech act lexicon construction. Specifically, guided by speech act theory (Searle, 1969), a speech act lexicon is constructed with performative verbs to support stance knowledge partition. Meanwhile, intermediate targets are mined with topic modeling to construct the target hierarchy for inference knowledge and explanation knowledge discrimination. Then, the *knowledge verification module* refines the acquired stance knowledge via a multi-agent collaborative process. Finally, the *knowledge retrieval* module matches the target language input to the most relevant stance knowledge for enhancing text representations, which in turn can also provide interpretable information for stance detection.

The contributions of our work are as follows:

- We make the first attempt to explicitly elicit different types of stance knowledge critical for stance identification, and propose an LLM-enabled knowledge elicitation and retrieval framework for zero-shot CLSD.

- Our framework conducts stance knowledge acquisition and verification through target structure and semantic lexicon based knowledge partition, as well as LLM agent collaboration.

- The knowledge retrieval process retrieves the most relevant stance knowledge based on the target language input as the transferrable knowledge to bridge the language gap.

- Experimental results on multilingual datasets verify the effectiveness of our method compared to competitive baselines, and its superiority over the approaches trained with the labeled data of target language.

2 Task Formulation

For the task of zero-shot cross-lingual stance detection (CLSD), ***no training data in target language is available***. The training set (source language) with N_s samples is denoted as $\boldsymbol{D_s} = \{(t_i^s, c_i^s), y_i^s\}_{i=1}^{N_s}$, where t_i^s, c_i^s are the pre-defined

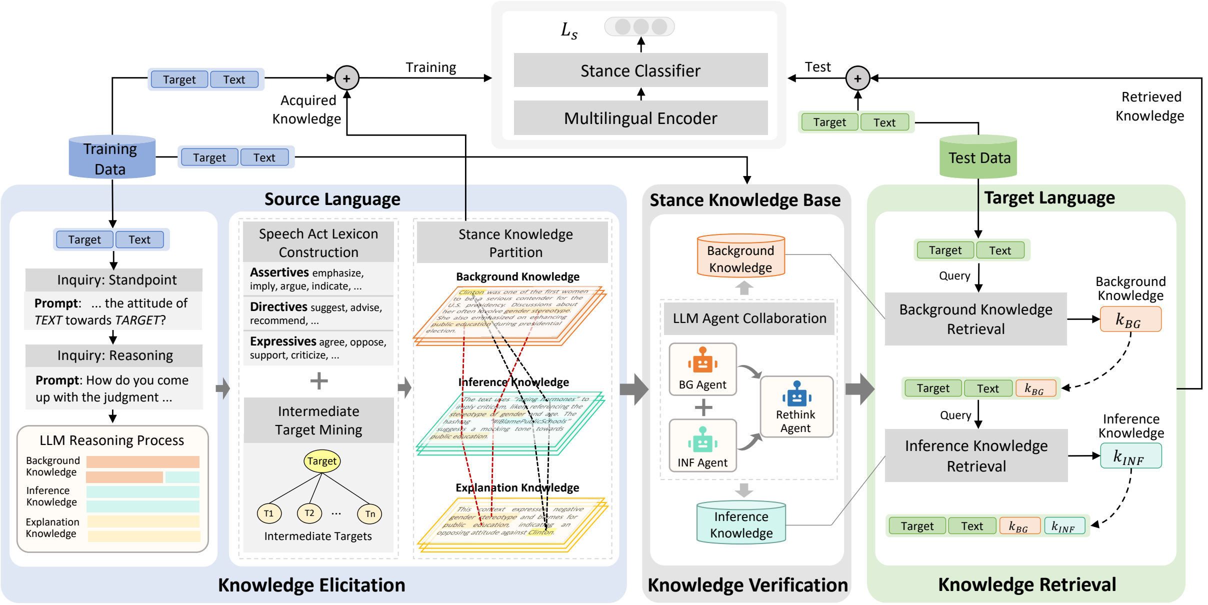


Figure 2: Overview of our LLM-enabled knowledge elicitation and retrieval framework KEAR for zero-shot CLSD.

target and text of the i-th sample, and y_i^s is the stance label. The test set (target language) is denoted as $\boldsymbol{D}_t = \{(t_i^t, c_i^t), y_i^t\}_{i=1}^{N_t}$.

We aim to learn a projection from the target-text pair to stance label, and we introduce external knowledge as an additional input to improve performance. Specifically, $\boldsymbol{D}_s$ is used to acquire language-agnostic background knowledge (BG), inference knowledge (INF), and explanation knowledge (EXP) from LLM's reasoning process. Each type of knowledge is represented as $\boldsymbol{K}_* = \{k_*^{(j)}\}$, where $* \in \{\text{BG}, \text{INF}, \text{EXP}\}$, and $k_*^{(j)}$ is the j-th knowledge (sentence) of the specific type $*$. All the knowledge is represented in the middle language.

3 Our Proposed Method

The overall structure of the proposed method KEAR is shown in Figure 2. *Knowledge elicitation module* first derives different types of stance knowledge from the LLM-generated reasoning process. The acquired knowledge is then validated by *knowledge verification module* and transferred to the source language through *knowledge retrieval module* hierarchically. Finally, the retrieved stance knowledge is concatenated with the input and then fed into the *classifier module*.

Stance identification:

TARGET: [t_i^s] TEXT: [c_i^s] What is the *attitude* of *TEXT* toward *TARGET*? Select from Favor, Against and None.

Stance reasoning:

How do you come up with the judgment that *TEXT* expresses such an *attitude* toward *TARGET*?

Figure 3: Prompt design for LLM stance reasoning.

3.1 Knowledge Elicitation Module

The knowledge elicitation module first acquires the reasoning process with LLM prompting (Section 3.1.1). After acquiring the reasoning process from LLM, we mine hierarchical target structure (Section 3.1.2) for each sample including destination target and intermediate targets. Since intermediate targets are related to the given target, they play a key role in revealing the implicit inferential relationship for stance identification. Also, we construct a speech act lexicon (Section 3.1.3) with performative verbs that possibly express standpoints for attitude detection. Based on these, we devise a knowledge partition (Section 3.1.4) algorithm to acquire stance knowledge step by step.

3.1.1 Stance Reasoning with Prompting

We first inquire about the stance of the given sample in source language. Specifically, given a target-text pair (t_i^s, c_i^s), we feed it into LLM with the stance identification prompt in Figure 3. If the stance identified by LLM does not match the ground truth y_i^s, such samples will be excluded from further consideration. Then, we obtain the LLM-generated reasoning process P_i for stance justification using the stance reasoning prompt in Figure 3.

Category	Sub-Category	Examples
Assertives	Subjective	accept, affirm, agree, decline, withdraw, doubt, refuse, ...
	Descriptive	imply, emphasize, argue, indicate, infer, justify, mention, state, highlight, ...
Directives	Advice	advise, suggest, urge, recommend, ...
	Command	order, command, request, ... (not relevant)
Expressives	Attitude	advocate, favor, oppose, approve, blame, accuse, condemn, criticize, support, mock, ...
	Behabitives	appreciate, thank, congratulate, welcome, greet, bless, congratulate, praise, apologize, ...

Table 1: Categories and examples of speech act lexicon that include standpoint expressions.

3.1.2 Target Structure Mining

We construct target architectures including (destination) targets and intermediate targets, which are implicitly related to targets, for the successive knowledge partition. We employ phrase-level topic modeling that yields latent topics as intermediate targets from the reasoning process acquired with LLM prompting. Specifically, for the reasoning process P_i, all the noun phrases w are identified and divided into two clusters, one cluster C_t that contains various mentions of the given (destination) targets, and the other cluster C_i that contains candidates possibly related to the given targets. We represent each sentence in P_i with the noun phrases w in C_i alone and utilize phrase-level topic modeling for intermediate target mining.

Adopting Latent Dirichlet Allocation (LDA), from the distribution of noun phrases in the reasoning process $\mathcal{P}(w_k|P_i)$, we can estimate the phrase distribution of the topic $\mathcal{P}(w_k|z_j)$ (where z_j represents the j-th topic) and the topic distribution of the reasoning process $\mathcal{P}(z_j|P_i)$:

$$\mathcal{P}(w_k|P_i) = \sum_{j=1}^{N_{tp}} \mathcal{P}(w_k|z_j)\mathcal{P}(z_j|P_i) \qquad (1)$$

where N_{tp} denotes the number of topics. For each topic z_j in the i-th reasoning process P_i, we select noun phrases with the top n_{tp} highest probabilities as intermediate targets $\boldsymbol{T}_i$.

$$\boldsymbol{T}_i = \text{argmax}_{n_{tp}} \mathcal{P}(w_k|z_j) \qquad (2)$$

Algorithm 1: Stance Knowledge Partition

```
Input: (1) The reasoning process $P_i = \{l_k\}_{k=1}^{n_i}$ for
       sample $(t_i^s, c_i^s)$; (2) Target hierarchy
       $\mathcal{H} = \{t_i^s, T_i\}$ that includes target $t_i^s$ and
       intermediate targets $T_i$; (3) Speech act lexicon
       $L_{sa} = \{v_j\}_{j=1}^{n_V}$; (4) Attitude detector $D$ that
       can determine attitude expression.
Output: Knowledge type of each sentence $l_k$ in $P_i$.
foreach $l_k \in P_i$ do
    /* 1: Target Hierarchy */
    if $l_k$ contains target hierarchy then
        /* 2: Speech Act Lexicon */
        if $\exists v_j \in L_{sa}, v_j \in l_k$ then
            /* 3: Attitude Detection */
            if $l_k$ has an attitude detected by $D$ then
                /* 4: Intermediate Target */
                if $v_j$'s object $\in T_i$ then
                    $l_k \rightarrow$ INF knowledge
                else
                    if $v_j$'s object is a $t_i^s$ mention
                      then
                        $l_k \rightarrow$ EXP knowledge
            else
                $l_k \rightarrow$ BG knowledge
        else
            $l_k \rightarrow$ BG knowledge
    else
        $l_k \rightarrow$ discard
```

3.1.3 Speech Act Lexicon Construction

To differentiate stance knowledge, our work relies on the well-founded speech act theory (Searle, 1969, 1979) in linguistic pragmatics as the guideline for identifying illocutionary acts (i.e. performative verbs), so as to discriminate stance knowledge. Table 1 shows our design of the speech act category structure, including Assertives, Directives, Expressives, and their sub-categories.

Specifically, *Assertives* convey information like statements and claims to support the speaker's standpoint. *Directives* represent the speaker's request and desire, which are also common for convincing others to approve one's proposition. *Expressives* express the speaker's attitudes and emotions on the specific objects. The details of lexicon construction are provided in Appendix C.

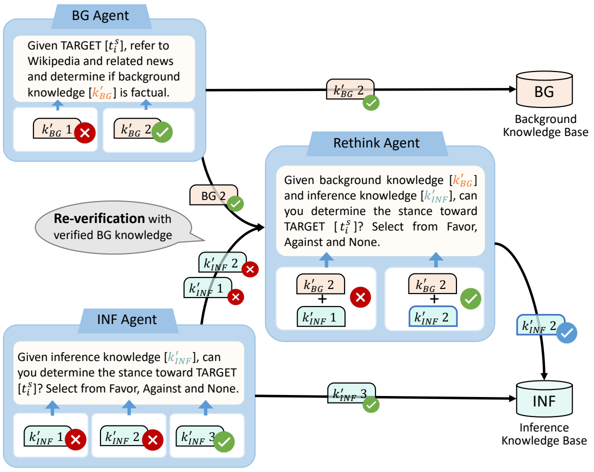
3.1.4 Stance Knowledge Partition

The knowledge partition is primarily based on relevance judgment and subjective detection. Relevance judgment utilizes the target hierarchical structure to eliminate irrelevant information generated by LLM. Subjective-objective detection is based on the characteristics of the background knowledge, which is mainly an objective statement without attitude expressing, consisting of lexicon-based judgment and attitude detection. At the sentence level, we initially filter out purely objective descriptions without performative verbs in the speech act lexicon. Further, we employ an attitude detector to determine standpoint expressions based on the semantics of the sentences.

The specific procedures of knowledge partition include four steps as shown in Algorithm 1. For each sentence l_k in the reasoning process P_i for sample (t_i^s, c_i^s), we first determine if there exists the mined target structure (step 1). l_k is then checked whether matches with the speech act lexicon (step 2) to differentiate background knowledge coarsely. If so, it proceeds to detect standpoint expression with an attitude detector (step 3) for fine-grained background knowledge partition. Finally, sentence l_k is further differentiated between inference knowledge and explanation knowledge according to whether the attitudes are expressed toward an intermediate target or (destination) target (step 4). Each stance knowledge is appended to the temporary stance knowledge base $\boldsymbol{K}'_*$.

3.2 Knowledge Verification Module

To alleviate the hallucination problems in LLM, the knowledge elicitation module has preliminarily verified the knowledge through stance identification in stance reasoning with prompting (Figure 3), as well as the knowledge partition process. In this section, we conduct an elaborative verification of the elicited stance knowledge with LLM agent collaboration, which can significantly enhance the usability of knowledge and further facilitate stance identification on target language. The proposed knowledge verification module includes a BG agent for verifying background knowledge, an INF agent for verifying inference knowledge, as well as a rethink agent for further verification of these two types of knowledge, as shown in Figure 4.



Given TARGET [t_i^s], refer to Wikipedia and related news and determine if background knowledge [k'_{BG}] is factual.

Given background knowledge [k'_{BG}] and inference knowledge [k'_{INF}], can you determine the stance toward TARGET [t_i^s]? Select from Favor, Against and None.

Given inference knowledge [k'_{INF}], can you determine the stance toward TARGET [t_i^s]? Select from Favor, Against and None.

Figure 4: The collaborative knowledge verification process for the sample (t_i^s, c_i^s).

For each training sample in the source language, the background knowledge in $\boldsymbol{K}'_{\mathrm{BG}}$ is input into the BG agent to assess its factuality. Simultaneously, the temporary inference knowledge in $\boldsymbol{K}'_{\mathrm{INF}}$ is input into the INF agent to testify whether it can infer the correct stance toward destination target. If it fails, we hypothesize that the knowledge might lack sufficient information for accurate stance determination. We give it another chance by supplementing the background knowledge for re-verification. The combination of inference knowledge and verified background knowledge is fed into the rethink agent. If it succeeds, the inference knowledge is accepted. The verified knowledge is appended to the knowledge base $\boldsymbol{K}_{\mathrm{BG}}$ and $\boldsymbol{K}_{\mathrm{INF}}$ respectively for subsequent knowledge retrieval. With the re-verification process, both background and inference knowledge are rigorously evaluated, thereby enhancing the reliability of the elicited knowledge in stance identification.

3.3 Knowledge Retrieval Module

To transfer the stance knowledge elicited from source language to target language for zero-shot CLSD, we devise a hierarchical knowledge retrieval module that retrieves the most relevant background knowledge and inference knowledge from the stance knowledge base. We adopt dense passage retrieval (DPR) method (Karpukhin et al., 2020) to build a cross-lingual retriever that contains one encoder $\mathrm{E_I}$ for input target-text pair and another encoder $\mathrm{E_K}$ for stance knowledge. To map the input target-text pair in target language and the candidate stance knowledge in the middle language to the same space, we utilize a cross-lingual retriever through which the relevant input pair and knowledge have a smaller distance so that the most related knowledge can be retrieved and transferred from source to target language.

Specifically, given an input target-text pair (t^t, c^t) in target language, it retrieves the most relevant background knowledge $\hat{k}_{\mathrm{BG}}$ from the knowledge base $\boldsymbol{K}_{\mathrm{BG}}$, based on inner dot product score:

$$score_j = \mathrm{E_I}\left(t^t, c^t\right)^\top \mathrm{E_K}\left(k_{\mathrm{BG}}^{(j)}\right) \tag{3}$$

$$\hat{k}_{\mathrm{BG}} = \operatorname{argmax}(\{score_j\}_{j=1}^{n_{\mathrm{BG}}}) \tag{4}$$

where n_{BG} is the size of $\boldsymbol{K}_{\mathrm{BG}}$. Hierarchically, based on the input and retrieved background knowledge $\hat{k}_{\mathrm{BG}}$, it retrieves the most relevant inference knowledge $\hat{k}_{\mathrm{INF}}$ from $\boldsymbol{K}_{\mathrm{INF}}$.

3.4 Stance Classification

The retrieved stance knowledge is concatenated together with target-text pair (t_i, c_i) and fed into the multilingual encoder. The last hidden state of [CLS] is taken as the knowledge-enhanced representation $\boldsymbol{h}_i \in \mathbb{R}^d$:

$$\overline{k}_i = \left[\hat{k}_{\mathrm{BG}_i}; \hat{k}_{\mathrm{INF}_i}\right] \tag{5}$$

$$\boldsymbol{h}_i = \mathrm{Encoder}(\texttt{[CLS]}t_i\texttt{[SEP]}c_i; \overline{k}_i\texttt{[SEP]}) \tag{6}$$

The acquired knowledge-enhanced representation $\boldsymbol{h}_i$ is fed into the classifier for cross-lingual stance detection, which is a two-layer feed-forward network followed by $\mathrm{Softmax}$. Note that the classifier is only trained using the source language data and its corresponding stance knowledge by minimizing cross-entropy loss $\mathcal{L}_S$:

$$\tilde{\boldsymbol{y}}_i^s = \mathrm{Softmax}(\mathrm{FFN}(\boldsymbol{h}_i^s)) \tag{7}$$

$$\mathcal{L}_S = -\frac{1}{N_s}\sum_{i=1}^{N_s} \boldsymbol{y}_i^s \log(\tilde{\boldsymbol{y}}_i^s) \tag{8}$$

where $\tilde{\boldsymbol{y}}_i^s$ is the predicted stance and $\boldsymbol{y}_i^s$ is the ground truth stance label.

4 Experiments

4.1 Experimental Setups

Datasets We evaluate the proposed method on three multilingual stance datasets. **Politics** (Zhang et al., 2023b) is a multilingual stance dataset constructed from X-stance (Vamvas and Sennrich, 2020), regarding "Foreign Policy" and "Immigration" in Swiss. Each sample can be classified into "Favor" or "Against". We take German as source language and French as target language, resulting in 5926 instances as source training data and 232 instances as test data in target language, with 31 different targets in total. **CIC** (Zotova et al., 2020) contains tweets on the target "Independence of Catalonia" in Spanish and Catalan. The categories of stances include "Favor", "Against" and "Neutral". Spanish is considered as the source language and Catalan is the target language. In this way, there are 6046 source training data and 2010 target test data. **VaxxStance** (Agerri et al., 2021) provides data in Spanish and Basque referring to "Vaccines". Each sample can be classified into "Favor", "Against" or "None". There are 1602 training instances in source language Spanish and 312 test instances in target language Basque.

Implementation Details All the experiments are conducted on GPUs of NVIDIA GeForce RTX 3090. The LLM for reasoning and knowledge verification is GPT-4 with on `gpt-4-1106-preview` using OpenAI API[1]. We use the spacy model `en_core_web_md`[2] for part-of-speech tagging and dependency parsing in the knowledge elicitation module. The topic model LDA is implemented with gensim[3], where both N_{tp} and n_{tp} are set to 2. The cross-lingual retriever is `mcontriever-msmarco`[4] which is a pre-trained model for information retrieval with contrastive learning (Izacard et al., 2021). The multilingual encoder is `xlm-roberta-base`[5], which contains 12 hidden layers and 12 attention heads, and the hidden size of d_h is 768. The parameters of the multilingual encoder and stance classifier are optimized by Adam, with a learning rate of $1e^{-5}$. The batch size is 32 for Politics, and 16 for CIC and VaxxStance. We train our method for 14 epochs and use the model of the last epoch for testing on target language data.

4.2 Comparative Methods

We compare our proposed method with representative monolingual stance detection methods, existing cross-lingual stance detection (CLSD) methods and zero-shot CLSD methods, as well as LLMs including GPT-3.5 and GPT-4.

Monolingual stance detection method (adapted to cross-lingual stance detection by replacing the embeddings or encoder with XLM-Roberta): **BiCond** (Augenstein et al., 2016) incorporates target information into text encoding with conditional biLSTMs; **TAN** (Du et al., 2017) learns target-specific representations with attention mechanism; **TGMN** (Wei et al., 2018) develops a multi-hop memory network and mines critical clues iteratively for stance detection; **CrossNet** (Xu et al., 2018) learns target-independent text

[1] https://platform.openai.com/docs/models

[2] https://spacy.io/models/en

[3] https://radimrehurek.com/gensim/

[4] https://huggingface.co/facebook/mcontriever-msmarco

[5] https://huggingface.co/FacebookAI/xlm-roberta-base

Method	**Resource**	**Politics** (de→fr)		**CIC** (es→ca)		**VaxxStance** (es→eu)	
		Acc (%)	**F1** (%)	**Acc** (%)	**F1** (%)	**Acc** (%)	**F1** (%)
Monolingual Stance Detection Method							
BiCond	-	60.9 ± 1.6	58.9 ± 1.8	46.7 ± 2.6	42.4 ± 2.5	43.1 ± 4.4	41.0 ± 3.1
TAN	-	60.2 ± 1.9	59.9 ± 1.9	48.1 ± 1.5	42.2 ± 3.4	45.4 ± 5.9	42.7 ± 4.5
TGMN	-	63.3 ± 1.2	62.1 ± 1.6	50.1 ± 1.9	44.9 ± 3.1	44.6 ± 4.8	42.6 ± 3.2
CrossNet	-	59.7 ± 2.5	57.5 ± 1.0	47.2 ± 1.4	42.2 ± 3.9	37.9 ± 2.0	35.0 ± 3.0
JointCL	-	72.3 ± 2.4	72.2 ± 2.4	49.8 ± 2.1	45.1 ± 4.3	43.9 ± 1.5	40.0 ± 2.1
Zero-Shot Cross-Lingual Stance Detection Method							
mWiki	zero-shot	-	58.8 ± 0.0†	-	21.7 ± 0.0†	-	-
enstance	zero-shot	-	61.1 ± 0.0†	-	22.3 ± 0.0†	-	-
mBERT-ft	zero-shot	67.7 ± 2.6	67.0 ± 2.7	51.0 ± 1.0	45.2 ± 2.5	41.8 ± 3.5	38.1 ± 1.6
XLM-R-ft	zero-shot	74.1 ± 0.4	73.7 ± 0.7	50.3 ± 2.4	45.3 ± 4.5	49.0 ± 3.5	45.0 ± 3.0
Large Language Model (Zero-Shot)							
GPT-3.5	-	73.8 ± 0.5	73.2 ± 0.4	34.7 ± 0.6	31.0 ± 0.6	51.0 ± 1.1	38.9 ± 0.8
GPT-4	-	78.8 ± 1.4	78.7 ± 1.4	51.2 ± 2.1	47.6 ± 2.0	46.3 ± 1.6	47.9 ± 1.5
KEAR (Ours)	zero-shot	**79.3 ± 1.9**	**79.2 ± 1.8**	**54.0 ± 0.6**	**52.5 ± 0.5**	**55.5 ± 1.7**	**53.1 ± 1.1**
Cross-Lingual Stance Detection Method							
TaRA	32-shot	79.3 ± 1.4‡	79.0 ± 1.4‡	53.1 ± 2.2	51.8 ± 1.3	53.8 ± 2.5	49.1 ± 4.2
CCSD	full unlabeled	70.1 ± 0.0‡	69.9 ± 0.0‡	43.2 ± 0.4	43.1 ± 0.4	42.1 ± 1.4	41.0 ± 1.2
mWiki	32-shot	-	57.7 ± 0.0†	-	42.3 ± 0.0†	-	-
enstance	32-shot	-	64.6 ± 0.0†	-	44.3 ± 0.0†	-	-
KEAR (Ours)	zero-shot	**79.3 ± 1.9**	**79.2 ± 1.8**	**54.0 ± 0.6**	**52.5 ± 0.5**	**55.5 ± 1.7**	**53.1 ± 1.1**

Table 2: Experimental results of comparative baselines and our proposed method KEAR on the three datasets. Column "Resource" denotes training data resource in target language of each method. † and ‡ denote that the results are taken from Hardalov et al. (2022) and Zhang et al. (2023a,b), respectively. All the trained models (except for mBERT-FT) are based on XLM-Roberta (Conneau et al., 2020) for CLSD. We report the average scores and standard deviations of 5 runs in percentage. The best performances are marked in bold.

representations with self-attention for cross-target stance detection; **JointCL** (Liang et al., 2022) proposes a prototypical contrastive learning strategy for zero-shot stance detection.

Zero-shot cross-lingual stance detection method (without target language data): **mWiki** (Hardalov et al., 2022) pre-trains XLM-R (Conneau et al., 2020) on additional multilingual Wikipedia data with sentiment-based stance task, and predicts the stance label in target language; **enstance** (Hardalov et al., 2022) is similar to mWiki and pre-trains XLM-R (Hardalov et al., 2022) on all the English stance datasets; both **mBERT-ft** (Devlin et al., 2019) and **XLM-R-ft** (Conneau et al., 2020) are fine-tuned on the source language data.

Cross-lingual stance detection method (trained with target language data): **TaRA** (Zhang et al., 2023a) devises target-level target relation alignment using labeled data in target language; **CCSD** (Zhang et al., 2023b) develops dual knowledge distillation framework using unlabeled data in target language.

4.3 Main Results

We use accuracy and macro F1 as the evaluation metrics. Table 2 reports the experimental results of the proposed method KEAR and baselines on the three datasets. For monolingual stance detection methods, learning target-specific representation based on cross-lingual word embeddings is not sufficient for cross-lingual stance detection. This is mainly because the language gap causes word vectors to be too separated in the representation space. For the zero-shot stance detection method in monolingual setting JointCL, the generalization ability is improved by modeling the correlation between known targets, which still does not completely fit the zero-shot CLSD task due to language disparity. As for zero-shot cross-lingual stance detection methods, the superior performances of fine-tuning mPLMs demonstrate their cross-lingual abilities in stance detection tasks. mWiki and enstance, pre-trained with sentiment-based tasks, obtain suboptimal performances on target language data in zero/few-shot setting, which is mainly due to the

Variant	**Politics** (de→fr)		**CIC** (es→ca)		**VaxxStance** (es→eu)	
	Acc (%)	**F1** (%)	**Acc** (%)	**F1** (%)	**Acc** (%)	**F1** (%)
KEAR (Ours)	79.3 ± 1.9	79.2 ± 1.8	54.0 ± 0.6	52.5 ± 0.5	55.5 ± 1.7	53.1 ± 1.1
Knowledge Elicitation						
w/o Intermediate Target	75.3 ± 0.4	75.2 ± 0.4	53.0 ± 1.9	48.7 ± 2.6	49.3 ± 2.9	47.3 ± 2.8
w/o Speech Act Lexicon	74.6 ± 1.7	74.5 ± 1.8	53.1 ± 0.5	49.8 ± 1.8	51.9 ± 2.3	49.3 ± 2.4
w/o Knowledge Partition	73.3 ± 1.3	73.1 ± 1.4	52.2 ± 1.3	47.3 ± 2.4	50.7 ± 2.2	47.1 ± 2.2
Knowledge Verification						
w/o Knowledge Verification	74.2 ± 2.0	74.1 ± 2.0	52.4 ± 2.7	46.8 ± 3.9	53.8 ± 1.2	50.6 ± 2.0
Knowledge Retrieval						
w/o Sequential Retrieval	76.3 ± 1.1	76.1 ± 1.1	53.0 ± 1.0	49.2 ± 1.9	50.7 ± 4.1	49.2 ± 3.1
w/o BG Knowledge Retrieval	75.4 ± 2.7	75.0 ± 2.4	53.3 ± 0.6	49.1 ± 1.1	51.3 ± 3.1	48.7 ± 3.0
w/o INF Knowledge Retrieval	75.0 ± 2.5	74.9 ± 2.5	52.3 ± 0.7	47.6 ± 1.0	49.3 ± 3.8	47.8 ± 2.5

Table 3: Ablation results of the variants of our proposed method KEAR on the three datasets. For each variant of KEAR, we report the average scores and standard deviations of 5 runs in percentage.

disparity between sentiment and stance detection tasks. Our proposed method KEAR performs 5-8% (macro F1) better than the above baseline methods on the three datasets. This improvement demonstrates that KEAR effectively reduces language disparities by utilizing knowledge elicited from LLM as a bridge and by transferring the knowledge across languages via fine-grained knowledge retrieval.

We also compare our method with existing CLSD methods trained with target language data, as shown in Table 2. Without using target language data for training, our method still outperforms the baselines, further demonstrating the effectiveness of our method incorporated with stance-relevant inferential knowledge. When comparing with the LLM baselines[1] in zero-shot setting, KEAR outperforms both GPT-3.5 and GPT-4 on the three benchmark datasets, indicating the effectiveness of the proposed knowledge elicitation and retrieval framework. Another advantage of KEAR is that its inference speed is much faster than that of LLMs, within 1 minute for each test dataset on GPU.

4.4 Ablation Study

Table 3 gives experimental results of the variants of our proposed method KEAR. For the knowledge elicitation module, the performance decreases more without intermediate target mining, underscoring its crucial role in revealing implicit inferential relationships for stance identification. Excluding the speech act lexicon results in a drop in performance, highlighting the importance of the constructed lexicon containing standpoint expressions for effective knowledge partition. Without discriminating the types of knowledge, the serious decrease in both accuracy and F1 indicates the effectiveness of fine-grained knowledge elicitation and retrieval. We can also see from the performance drop that the verification module is necessary for high-quality knowledge elicitation. As for knowledge retrieval, the hierarchical retrieval strategy aligns with the associations between different types of stance knowledge. Moreover, excluding the retrieval of any specific type of knowledge results in a performance drop, especially for inference knowledge. This indicates that inference knowledge with attitude expressions toward intermediate targets is more important for revealing implicit inferential relationship for stance judgement. Detailed analysis of stance knowledge is provided in Figure 5, Appendix A.2.

4.5 Human Evaluation

We conduct human evaluation to assess the quality of the elicited background knowledge and inference knowledge. For the three datasets, we randomly extract 200 pieces of knowledge. Specifically, for *background knowledge*, we examine (1) whether the knowledge is objective and contains complete information; (2) whether it is a factual statement about destination target. For *inference knowledge*, we examine (1) whether the knowledge expresses viewpoints toward the intermediate target other than the final target; (2) whether the knowledge provides reasoning process on how to come into the attitude toward the intermediate. As shown in Table 4, we calculate the proportion of knowledge that meets the requirements determined

[1] The prompt for the three datasets is the same as the stance identification prompt in knowledge elicitation module.

Knowledge	Politics			CIC			VaxxStance		
	Ratio 1	Ratio 2	Avg	Ratio 1	Ratio 2	Avg	Ratio 1	Ratio 2	Avg
Background	96.5	97.0	96.8	97.0	97.0	97.0	97.5	98.0	97.8
Inference	92.5	90.5	91.5	93.0	91.5	92.3	93.0	93.0	93.0

Table 4: Human evaluation on the efficacy of each type of stance knowledge. Ratio 1 and Ratio 2 denote the correct ratio of two evaluators respectively. The average Cohen's kappa coefficients (Cohen, 1960) κ of the inter-rater agreement for human evaluation on background knowledge and inference knowledge are 0.67 and 0.81, respectively (note that $0.6 \leq \kappa \leq 0.8$ means substantial agreement and $\kappa \geq 0.8$ means almost perfect agreement).

by two evaluators (Ratio 1 and Ratio 2), and report the average rate (Avg). We also provide the average kappa coefficients that reflect the inter-rater agreement on the knowledge evaluation.

From the results in Table 4, we can see that the majority of the stance knowledge satisfies the standards, demonstrating the effectiveness of our proposed knowledge elicitation module. Besides, we can see that the accuracy of inference knowledge is slightly lower compared with background knowledge. This is because inference knowledge reflects the implicit inferential relationship of stance identification, and is determined based on the intermediate targets with diverse forms and contents. Thus, its acquisition and discrimination are more difficult than those of background knowledge.

5 Related Work

Monolingual Stance Detection *In-domain stance detection* aims to identify the stance toward pre-defined targets (Wei et al., 2018; Chai et al., 2022; Zheng et al., 2022). *Zero-shot stance detection* predicts stance on unseen targets without training data. The mainstream work adopts transfer learning (Wei and Mao, 2019; Allaway et al., 2021; Hardalov et al., 2021) and mines implicit associations across targets (Allaway and Mckeown, 2020; Liang et al., 2022; Liu et al., 2022).

As for external knowledge, some work uses sentiment lexicon to guide knowledge transfer between targets (Sun et al., 2018; Li and Caragea, 2019; Zhang et al., 2020). Liu et al. (2021); Luo et al. (2022) introduce commonsense knowledge from ConceptNet to improve the generalization ability. Besides, He et al. (2022); Zhu et al. (2022) incorporate target-related knowledge from Wikipedia to further enhance stance detection. Further, Li et al. (2023) utilizes LLM to filter the knowledge retrieved from Wikipedia for stance detection augment. However, their work fails to consider the stance-relevant inferential knowledge which is important for stance identification.

Cross-Lingual Stance Detection Compared with English, data resources in most other languages are rather limited (Lai et al., 2018; Cignarella et al., 2020; Baly et al., 2018; Khouja, 2020). To address this problem, cross-lingual stance detection (CLSD) transfers knowledge from high-resource source language to low-resource target language. Mohtarami et al. (2019) proposes contrastive language adaptation to align representations in source and target languages. Based on this, Zhang et al. (2023a) further develops a target-level contrastive learning method for fine-grained alignment. However, both methods rely on annotated data in target language. Zhang et al. (2023b) utilizes unlabeled data in target language via dual knowledge distillation to bridge the language gap.

To tackle the situation of extremely scarce data resources, zero-shot CLSD (Vamvas and Sennrich, 2020) identifies the stance toward targets with no training data available in target language. Hardalov et al. (2022) pre-train multilingual PLM with sentiment-based corpus and transfer the knowledge to target languages via prompt-based tuning.

6 Conclusion

We propose an LLM-enabled knowledge elicitation and retrieval framework for zero-shot cross-lingual stance detection, which explicitly elicits stance knowledge critical for stance detection from LLM. The knowledge elicitation module acquires stance knowledge with target structure and lexicon based knowledge partition, and the verification module further verifies the knowledge with collaborative agents. The knowledge retrieval process matches the target language data with the most relevant stance knowledge to bridge the language gap. Experimental results on the multilingual stance datasets verify the effectiveness of our method.

Limitations

Although our proposed framework is intended to be developed for zero-shot cross-lingual stance identification, it can also be utilized in monolingual stance detection setting. In such setting, the LLM-enabled knowledge elicitation module in our framework is utilized to acquire stance-relevant knowledge for enhancing stance detection. The knowledge retrieval module in such monolingual setting functions as a means to select the most relevant stance knowledge among the volume of the acquired stance knowledge. Nonetheless, since we have not conducted a thorough experimental study to verify the performance of our method in monolingual setting, we shall leave the exploration of this issue to our future work.

Acknowledgments

This work is supported in part by the National Natural Science Foundation of China under Grants #72293575 and #62206287, and the Beijing Nova Program 20230484430.

References

Rodrigo Agerri, Roberto Centeno, Marıa Espinosa, Joseba Fernandez de Landa, and Alvaro Rodrigo. 2021. VaxxStance: A dataset for cross-lingual stance detection on vaccines.

Emily Allaway and Kathleen Mckeown. 2020. Zero-shot stance detection: A dataset and model using generalized topic representations. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 8913–8931.

Emily Allaway, Malavika Srikanth, and Kathleen Mckeown. 2021. Adversarial learning for zero-shot stance detection on social media. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4756–4767.

Isabelle Augenstein, Tim Rocktäschel, Andreas Vlachos, and Kalina Bontcheva. 2016. Stance detection with bidirectional conditional encoding. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 876–885.

John Langshaw Austin. 1975. *How to do things with words*. Harvard university press.

Ramy Baly, Mitra Mohtarami, James Glass, Lluís Màrquez, Alessandro Moschitti, and Preslav Nakov. 2018. Integrating stance detection and fact checking in a unified corpus. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 21–27.

Heyan Chai, Siyu Tang, Jinhao Cui, Ye Ding, Binxing Fang, and Qing Liao. 2022. Improving multi-task stance detection with multi-task interaction network. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 2990–3000.

Alessandra Teresa Cignarella, Mirko Lai, Cristina Bosco, Viviana Patti, Rosso Paolo, et al. 2020. SardiStance@EVALITA2020: Overview of the task on stance detection in italian tweets. In *Proceedings of the Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2020)*, pages 1–10.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Édouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186.

Jiachen Du, Ruifeng Xu, Yulan He, and Lin Gui. 2017. Stance classification with target-specific neural attention networks. In *Proceedings of the International Joint Conferences on Artificial Intelligence*, pages 3988–3994.

Momchil Hardalov, Arnav Arora, Preslav Nakov, and Isabelle Augenstein. 2021. Cross-domain label-adaptive stance detection. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 9011–9028.

Momchil Hardalov, Arnav Arora, Preslav Nakov, and Isabelle Augenstein. 2022. Few-shot cross-lingual stance detection with sentiment-based pre-training. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 10729–10737.

Zihao He, Negar Mokhberian, and Kristina Lerman. 2022. Infusing knowledge from Wikipedia to enhance stance detection. In *Proceedings of the 12th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis*, pages 71–77.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2021. Unsupervised dense information retrieval with contrastive learning. *arXiv preprint arXiv:2112.09118*.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 6769–6781.

Jude Khouja. 2020. Stance prediction and claim verification: An Arabic perspective. In *Proceedings of the Third Workshop on Fact Extraction and VERification*, pages 8–17.

Dilek Küçük and Fazli Can. 2020. Stance detection: A survey. *ACM Computing Surveys*, 53(1):1–37.

Mirko Lai, Viviana Patti, Giancarlo Ruffo, and Paolo Rosso. 2018. Stance evolution and twitter interactions in an italian political debate. In *International Conference on Applications of Natural Language to Information Systems*, pages 15–27. Springer.

Ang Li, Bin Liang, Jingqian Zhao, Bowen Zhang, Min Yang, and Ruifeng Xu. 2023. Stance detection on social media with background knowledge. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 15703–15717.

Yingjie Li and Cornelia Caragea. 2019. Multi-task stance detection with sentiment and stance lexicons. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing and the International Joint Conference on Natural Language Processing*, pages 6299–6305.

Bin Liang, Qinglin Zhu, Xiang Li, Min Yang, Lin Gui, Yulan He, and Ruifeng Xu. 2022. JointCL: A joint contrastive learning framework for zero-shot stance detection. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 81–91.

Rui Liu, Zheng Lin, Huishan Ji, Jiangnan Li, Peng Fu, and Weiping Wang. 2022. Target really matters: target-aware contrastive learning and consistency regularization for few-shot stance detection. In *Proceedings of the International Conference on Computational Linguistics*, pages 6944–6954.

Rui Liu, Zheng Lin, Yutong Tan, and Weiping Wang. 2021. Enhancing zero-shot and few-shot stance detection with commonsense knowledge graph. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3152–3157.

Yun Luo, Zihan Liu, Yuefeng Shi, Stan Z Li, and Yue Zhang. 2022. Exploiting sentiment and common sense for zero-shot stance detection. In *Proceedings of the International Conference on Computational Linguistics*, pages 7112–7123.

Saif Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016. SemEval-2016 task 6: Detecting stance in tweets. In *Proceedings of the International Workshop on Semantic Evaluation*, pages 31–41.

Mitra Mohtarami, James Glass, and Preslav Nakov. 2019. Contrastive language adaptation for cross-lingual stance detection. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing and the International Joint Conference on Natural Language Processing*, pages 4442–4452.

John R Searle. 1969. *Speech acts: An essay in the philosophy of language*, volume 626. Cambridge university press.

John R Searle. 1979. *Expression and meaning: Studies in the theory of speech acts*. Cambridge University Press.

Qingying Sun, Zhongqing Wang, Qiaoming Zhu, and Guodong Zhou. 2018. Stance detection with hierarchical attention network. In *Proceedings of the International Conference on Computational Linguistics*, pages 2399–2409.

Jannis Vamvas and Rico Sennrich. 2020. X-stance: A multilingual multi-target dataset for stance detection. In *Proceedings of the SwissText & KONVENS Joint Conference*, pages 1–9.

Penghui Wei and Wenji Mao. 2019. Modeling transferable topics for cross-target stance detection. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1173–1176.

Penghui Wei, Wenji Mao, and Daniel Zeng. 2018. A target-guided neural memory model for stance detection in twitter. In *Proceedings of the International Joint Conference on Neural Networks*, pages 1–8.

Chang Xu, Cecile Paris, Surya Nepal, and Ross Sparks. 2018. Cross-target stance classification with self-attention networks. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 778–783.

Bowen Zhang, Min Yang, Xutao Li, Yunming Ye, Xiaofei Xu, and Kuai Dai. 2020. Enhancing cross-target stance detection with transferable semantic-emotion knowledge. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 3188–3197.

Ruike Zhang, Nan Xu, Hanxuan Yang, Yuan Tian, and Wenji Mao. 2023a. Target-oriented relation alignment for cross-lingual stance detection. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6391–6404.

Ruike Zhang, Hanxuan Yang, and Wenji Mao. 2023b. Cross-lingual cross-target stance detection with dual knowledge distillation framework. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 10804–10819.

Kai Zheng, Qingfeng Sun, Yaming Yang, and Fei Xu. 2022. Knowledge stimulated contrastive prompting for low-resource stance detection. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 1168–1178.

Qinglin Zhu, Bin Liang, Jingyi Sun, Jiachen Du, Lanjun Zhou, and Ruifeng Xu. 2022. Enhancing zero-shot stance detection via targeted background knowledge. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2070–2075.

Elena Zotova, Rodrigo Agerri, Manuel Nuñez, and German Rigau. 2020. Multilingual stance detection in tweets: The catalonia independence corpus. In *Proceedings of the Language Resources and Evaluation Conference*, pages 1368–1375.

A Analysis of Knowledge Elicitation

A.1 Analysis of Knowledge Source

We extract target-related textual knowledge from Wikipedia and GPT-4 to compare the impact of knowledge sources, and also compare the ways of incorporating the knowledge into stance detection. Following He et al. (2022), we input knowledge together with the target-text pair into stance detection model. Experimental results in Table 5 show that Wikipedia and GPT-4 have similar performances, yet both lower than our method KEAR because the knowledge in our method is extracted from the stance reasoning process which is closely related to stance determination. In addition, the performance of directly combining the extracted knowledge with each sample is lower than that of the retrieval method due to the irrelevant information.

A.2 Analysis of Stance Knowledge

Figure 5 shows the results of our method using background knowledge, inference knowledge and both two types of knowledge in knowledge retrieval module, in accordance with the results of the last two lines of ablation study in Table 3. Only using inference knowledge in our knowledge retrieval gains a higher performance on the three datasets compared to only using background knowledge. We speculate the reason for this phenomenon is that inference knowledge expressing attitudes toward intermediate targets, vital for revealing the implicit inferential relationship for stance detection in data-scarce languages. We also observe that individual knowledge is not as effective as combined knowledge, further demonstrating the effectiveness of our proposed method.

Source	Retri.	**Politics**	**CIC**	**VaxxStance**
		F1 (%)	F1 (%)	F1 (%)
Wikipedia	w/o	69.0 ± 2.2	44.1 ± 5.0	42.7 ± 4.7
Wikipedia	w/	73.2 ± 2.6	45.0 ± 4.8	45.5 ± 3.6
GPT-4	w/o	68.9 ± 3.8	45.7 ± 3.4	40.3 ± 2.4
GPT-4	w/	**79.2** ± **1.8**	**52.5** ± **0.5**	**53.1** ± **1.1**

Table 5: The impact of knowledge sources and their utilizations on the three datasets. "Retri." denotes retriever, and "w/" and "w/o" denote using the retriever or not for matching the most relevant knowledge.

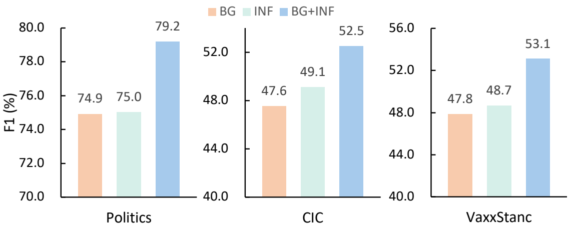


Figure 5: The comparisons of background knowledge and inference knowledge on the three datasets.

B Examples of Stance Knowledge

Table 6 provides cases of background knowledge and inference knowledge. We can see from the table that the acquired background knowledge contains objective factual statements. Inference knowledge reveals attitudes and opinions toward intermediate targets related to destination target based on the textual expressions.

C Speech Act Lexicon Construction

Specifically, *Assertives* convey information like statements and claims to support the speaker's standpoint. We classify its sub-categories as "Subjective" and "Descriptive". *Directives* represent the speaker's request and desire, which is also common for convincing others to approve one's proposition. We classify its sub-categories as "Advice" and "Command", where the sub-category "Command" should be excluded. *Expressives* express the speaker's attitudes and emotions on the specific objects, which are the most important signals in stance expressions. Following Austin (1975), we classify its sub-categories as "Attitude" and "Behabitives".

To construct the speech act lexicon, we randomly select 5000 samples from X-stance (Vamvas and Sennrich, 2020), and acquire the reasoning process via LLM prompting. We then extract all the verbs (1000+) with part-of-speech tagging and filter out irrelevant ones based on the above taxonomy. We finally construct a lexicon with 321 performative verbs for stance knowledge acquisition.

Knowledge Type	Content
Background	The **UNSC** is one of the six principal organs of **the United Nations**, charged with ensuring international peace and **security**, accepting new members to the UN, and approving any changes to its charter.
	Vox is a **political party** in Spain known for its strong Spanish nationalism and **opposition to regional separatism**, including the independence of Catalonia.
	The **SCC** has historically positioned itself in favor of **unity with Spain** and has been an active participant in the political discourse surrounding Catalan independence.
	Factual knowledge about the **Pfizer vaccine** is that it is one of the vaccines authorized for emergency use to prevent **COVID-19** and is considered safe and effective by health authorities worldwide.
	The **Moderna vaccine** has been through clinical trials and authorized for emergency use in various countries to combat the pandemic.
Inference	The author fears that Swiss farmers could not compete with cheaper imports and suggests that Switzerland should focus on strengthening self-sufficiency and ***supporting*** **local farmers** with fair prices.
	The criticism of Ciudadanos' ineffectiveness implies ***disapproval*** of their **handling of the independence issue**.
	Since the Constitution currently maintains the unity of Spain, defending it suggests ***opposition*** to any **separatist movements** that would break this unity.
	The use of hashtags such as PorEspaña and EspañaViva, which translate to "For Spain" and "Lively Spain," respectively, alongside the support for @vox_es, suggests a nationalistic sentiment and a ***desire*** for **a unified Spain**.
	The TEXT criticizes those who mocked the government and Health Minister Salvador Illa for announcing a vaccine by December, referring to political figures such as Ayuso and Bonilla, who are implied to have been ***skeptical*** or ***negative*** about the **vaccine rollout** and **public health measures**.

Table 6: Examples of the elicited stance knowledge generated by LLM. The intermediate targets and (destination) targets are in bold font. The textual expressions that express attitudes toward them are highlighted in italics.

D Attitude Detector

The attitude detector in our work consists of an encoder of pre-trained language model BERT (Devlin et al., 2019) and a classifier with two-layer feed-forward network. Each sentence l_k is encoded as a hidden vector. It is combined with an additional feature computed with the speech act lexicon and fed into the classifier. Specifically, the additional feature is a one-hot vector $h_{sa} \in \mathbb{R}^{d_{sa}}$, where d_{sa} is the size of the speech act lexicon. If any word in the sentence l_k matches with the verbs in the speech act lexicon, the corresponding dimension of h_{sa} is set to 1. The attitude detector is optimized with cross-entropy loss.

We randomly select 300 samples from Politics and acquire 250 LLM-generated reasoning processes whose stance prediction is correct. Each reasoning process is split into sentences, resulting in 1389 sentences. We utilize the existing stance model to label the stance expression automatically. "Favor" and "Against" are considered as expressing attitude, and "Neutral" is considered as no attitude expression. We split the 1389 labeled sentences into train/valid/test sets with the ratio of "60%-20%-20%" for training the attitude detector. We select the model that performs best on the valid set for further attitude detection.

E Prompt Engineering

We use GPT-4 for knowledge elicitation and verification. Below we provide the prompt engineering steps in KEAR.

E.1 Knowledge Elicitation

Step 1. Design candidate prompts for stance identification and stance reasoning. Candidate prompts are listed in Table 7.

Step 2. Evaluate candidate prompts by calculating the average accuracy of each prompt. Take the evaluation of stance identification prompts as an example. We randomly select 100 samples from source language data in the benchmark datasets and calculate the accuracy of each prompt based on the ground truth of stance label.

Step 3. Select the prompt with the highest performance. Selected prompts for stance identification:

Stance Identification Prompts
P1. TARGET: [target] TEXT: [text] What is the attitude of TEXT toward TARGET? Select from Favor, Against and None. **P2.** What is the attitude of [text] toward [target]? Select from Favor, Against and None. **P3.** The attitude of [text] toward [target] is [MASK]. Select from Favor, Against and None.
Stance Reasoning Prompts
P1. How do you come up with the judgment that TEXT expresses such an attitude toward TARGET? **P2.** How do you infer the judgment that TEXT expresses such an attitude toward TARGET? **P3.** How do you come up with the judgment that TEXT expresses such an attitude toward TARGET? Think step by step.

Table 7: Candidate prompts for stance identification and stance reasoning with LLM.

> TARGET: [target] TEXT: [text] What is the attitude of TEXT toward TARGET? Select from Favor, Against and None.

Selected prompts for stance reasoning:

> How do you come up with the judgment that TEXT expresses such an attitude toward TARGET?

Step 4. Determine zero-shot vs. 1-shot prompting by calculating the average accuracy of selected prompts. Take the evaluation of stance identification prompts as an example. We randomly select 100 samples from source language data in the benchmark datasets and calculate the accuracy of each prompt based on the trade-off between accuracy and efficiency.

E.2 Knowledge Partition and Verification

The prompt engineering steps in knowledge partition and verification are as follows:

Step 1. Design stance knowledge partition strategies. Table 8 shows detailed strategies, which are the main clues for the design of the knowledge partition algorithm in Algorithm 1.

Step 2. Design LLM agent collaboration strategies for knowledge verification. Table 9 shows detailed strategies, which are the main clues for the design of the knowledge verification process in Section 3.2.

Knowledge Partition Strategy
Strategy 1. For each sentence in the reasoning process for the target-text pair, we first determine if there exists the mined target structure. **Strategy 2.** The sentence is then checked whether it matches the speech act lexicon to differentiate background knowledge coarsely. **Strategy 3.** The sentence proceeds to detect standpoint expression with an attitude detector for fine-grained background knowledge partition. **Strategy 4.** The sentence is further differentiated between inference knowledge and explanation knowledge according to whether the attitudes are expressed toward an intermediate target or destination target.

Table 8: Strategies for knowledge partition.

Knowledge Verification Strategy
Strategy 1. For each training sample in source language, the background knowledge is input into the BG agent to assess its factuality. **Strategy 2.** The inference knowledge is input into the INF agent to testify whether it can infer the correct stance toward destination target. **Strategy 3.** If Strategy 2 fails, the specific inference knowledge is supplemented with verified background knowledge for re-verification in the Rethink agent.

Table 9: Strategies for knowledge verification.

Step 3. Design prompts for collaborative knowledge verification. To evaluate the quality of the designed prompt, we calculate the average accuracy based on the ground truth of human rating and stance labels. Selected prompts for BG agent:

> Given TARGET [target], refer to Wikipedia and related news and determine if background knowledge [k_{BG}] is factual.

Selected prompts for INF agent:

> Given inference knowledge [k_{INF}], can you determine the stance toward TARGET [target]? Select from Favor, Against and None.

Selected prompts for EXP agent:

> Given background knowledge [k_{BG}] and inference knowledge [k_{INF}], can you determine the stance toward TARGET [target]? Select from Favor, Against and None.

where k_{BG} and k_{INF} are background knowledge and inference knowledge respectively.