BKEE: Pioneering Event Extraction in the Vietnamese Language

Thi-Nhung Nguyen[1], Tien-Bang Tran[2], **Trong-Nghia Luu[2]**,
Thien Huu Nguyen[3], Kiem-Hieu Nguyen[2]

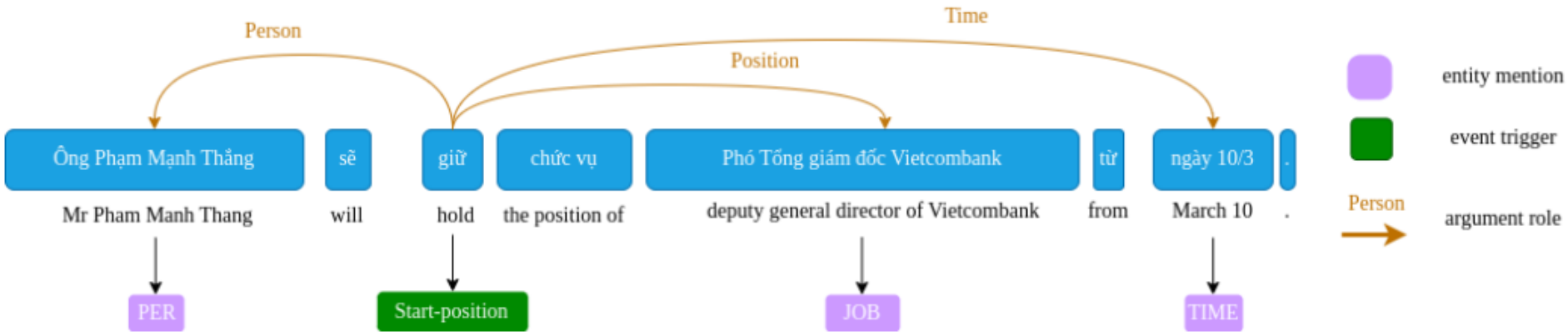
[1]VinAI Research

[2]School of Information and Communication Technology, Hanoi University of Science and Technology

[3]Department of Computer Science, University of Oregon

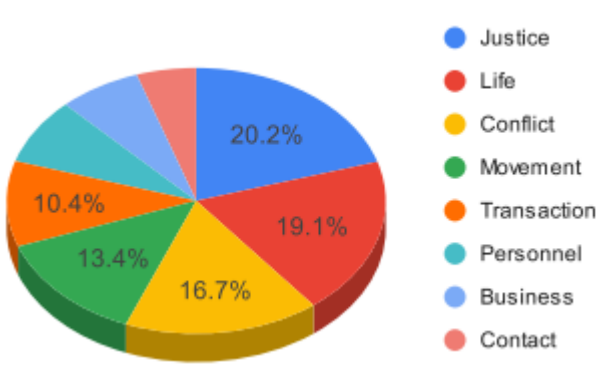
Event Extraction

- EE main tasks:
  - Entity Mention Detection (EMD): to find words that refer to real-world entities and their types
  - Event Detection (ED): to find the words (event trigger) that refer to the occurrence of the event and their types
  - Event Argument Extraction (EAE): to find entities that are involved in the event and their roles
- Example

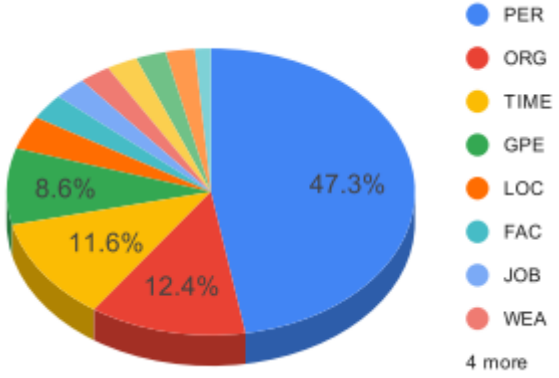


BKEE Dataset

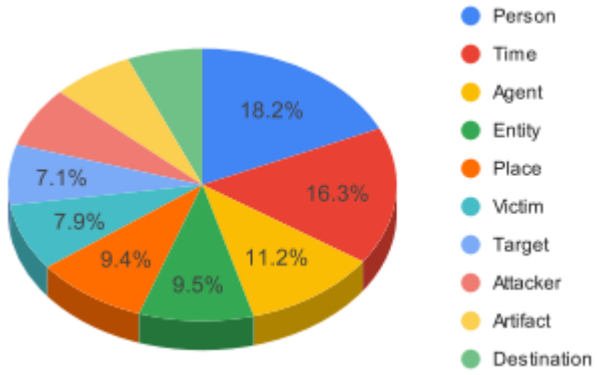
- Dataset properties:
  - News media baomoi.com: 2018-2020
  - Included 11 domains: entertainment, transportation, business, …
  - 1066 documents, 21318 sentences, ~ 17 word/sentence
  - ~ 9k events, 16k entities and arguments:



Event types



Entity types



Key arguments

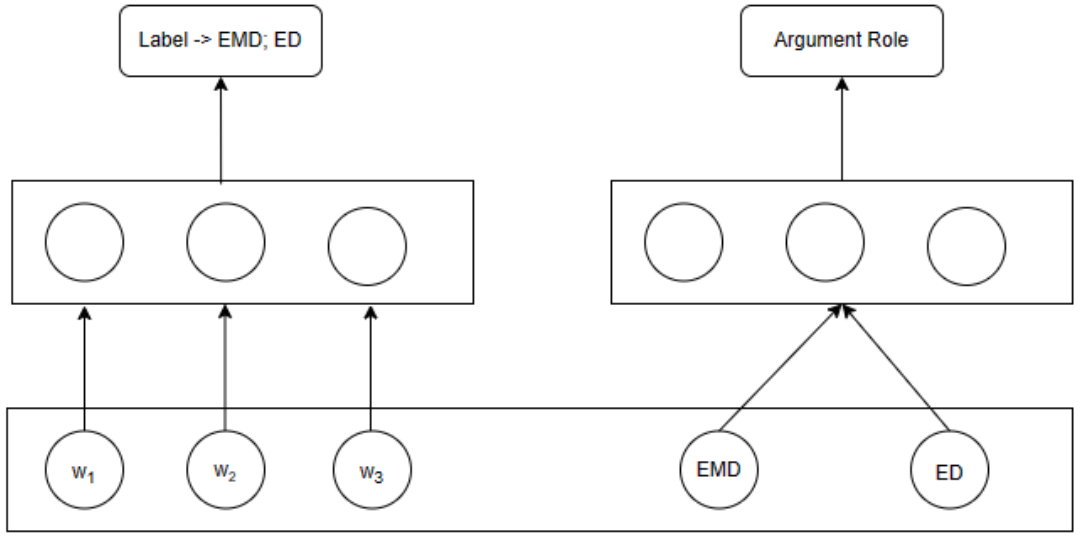
Data Anotation

- Same rules with **ACE 2005 dataset** (Walker et al., 2006)
- 8 event types, 33 event subtypes, 12 entity types, 28 argument roles
- 3 experienced labelers involved in annotation tasks
- Two-stage process:
  - 10% are co-annotated to assess agreement scores (using **Krippendorff's alpha** (Krippendorff, 2011) + **MASI score** (Passonneau, 2006))
  - Rest 90% are distributed for separate annotation

Task	Count	IAA(%)	Challenges
EMD	8,717	83.0	JOB
ED	16,010	83.0	DEMONSTRATE
EAE	16,010	85.1	DESTINATION

Experiments

- Pipeline: use **XLM-Roberta/ PhoBERT** to handle each task seperately.
- Joint learning models: simultaneously infers all three tasks
  - **OneIE** (Lin et al., 2020)
  - **FourIE** (Nguyen et al., 2021)



Input	Mr	Pham	Manh	Thang	will	hold	the	position	of	deputy	general	director	of	VietcomBank	from	March	10/3	.
ED	O	O	O	O	O	B-start-position	O	O	O	O	O	O	O	O	O	O	O	O
EMD	B-PER	I-PER	I-PER	I-PER	O	O	O	O	O	B-JOB	I-JOB	I-JOB	I-JOB	I-JOB	O	B-TIME	I-TIME	O
EAE	(hold-Mr Pham Manh Thang):Person; (hold-deputy general director of Vietcombank):JOB; (hold-March 10/3):TIME																	

Performances on BKEE

Task	Pipeline	OneIE	FourIE
Entity	54.4	55.8	**57.6**
Event	61.8	**62.8**	61.9
Argument	44.4	53.0	**53.4**

Joint learing models

The performance (F1-score) of baselines using PhoBERT on BKEE

Task	Pipeline	OneIE	FourIE
Entity	55.0	56.3	**56.4**
The Event	60.3	60.0	**61.5**
Argument	44.9	**51.7**	51.6

Joint learing models

The performance (F1-score) of baselines using XLM-RoBERTa on BKEE

Error Analysis

- Overlapping context (35%): Sentences containing events are often long and have many overlapping contextual elements.
  - Ex: “After the crime, Hung visited Tran Van Chien’s house to discuss it, and Chien purchased a SIM card to stay in touch with Hung during his escape”
- Span errors (28%): These errors occur when the model captures part of a mention but does not overlap completely with the gold one.
  - Ex: “Vietnam Bank for Agriculture and Rural Development Agribank” -> “Agribank”

Error Analysis

- Potentially relevant (12%): Entities, triggers, and arguments are identified that can be considered valid based on manual review.
  - Ex: "Iraq and Syria“ -> 2 entity “Irag”; “Syria”

- Abbreviations (6%): Abbreviations in the text are sometimes misunderstood.
  - Ex: “CEO” -> “JOB” instead of “PER”

Conclusion

- The first Vietnamese-language EE dataset that achieves three main goals:
  - Reducing the gap between rich-resource and low-resource languages
  - Pioneering EE development for the Vietnamese language
  - Establishing strong baselines to support future works and analyzing the challenges faced
- Vietnamese EE encounters cases of overlapping context, complex event and entity mentions

Limitations & Future works

- Limitations:
  - Currently intra-sentence due to annotation resource.
  - Lower performance compared to SOTA models for high-resource languages.
- Future works:
  - Expand to document level to understand global semantics and complex information.
  - Improve word boundary detection or increase the ability to understand document structure