From Disfluency Detection to Intent Detection and Slot Filling for Vietnamese

Dat Quoc Nguyen

Head of NLP, VinAI

https://datquocnguyen.github.io

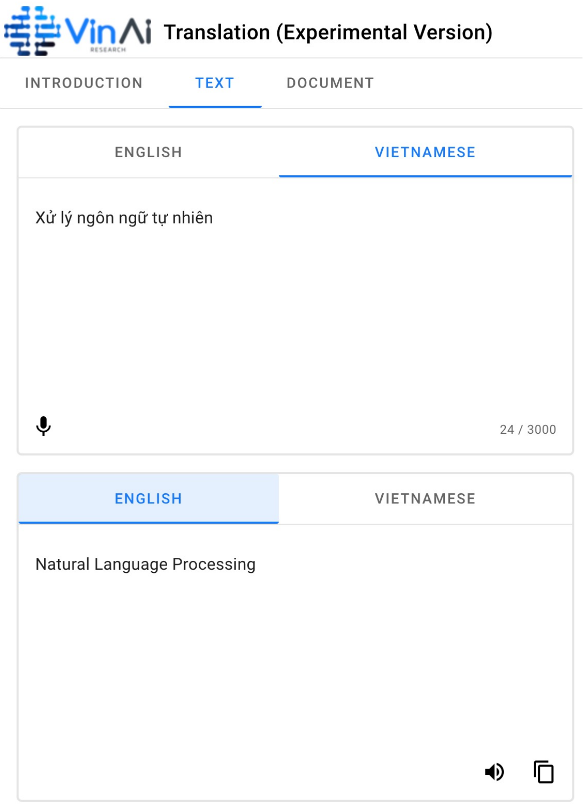
11/11/2022

Natural Language Processing

- Processing natural languages with computers, enabling computers to understand, generate and analyze natural languages
- Applications
  - Machine Translation
  - Information Retrieval
  - Question Answering
  - Dialogue Systems
  - Information Extraction
  - Summarization
  - Sentiment Analysis
  - …
- Core technologies
  - Language modeling
  - Part-of-speech tagging
  - Syntactic parsing
  - Named-entity recognition
  - Word sense disambiguation
  - Semantic role labeling
  - …
- **NLP lies at the intersection of computational linguistics and machine learning**

Slide content taken from https://www.dropbox.com/s/cz4zmb0I95p9r4y/NLP_1_intro.pdf

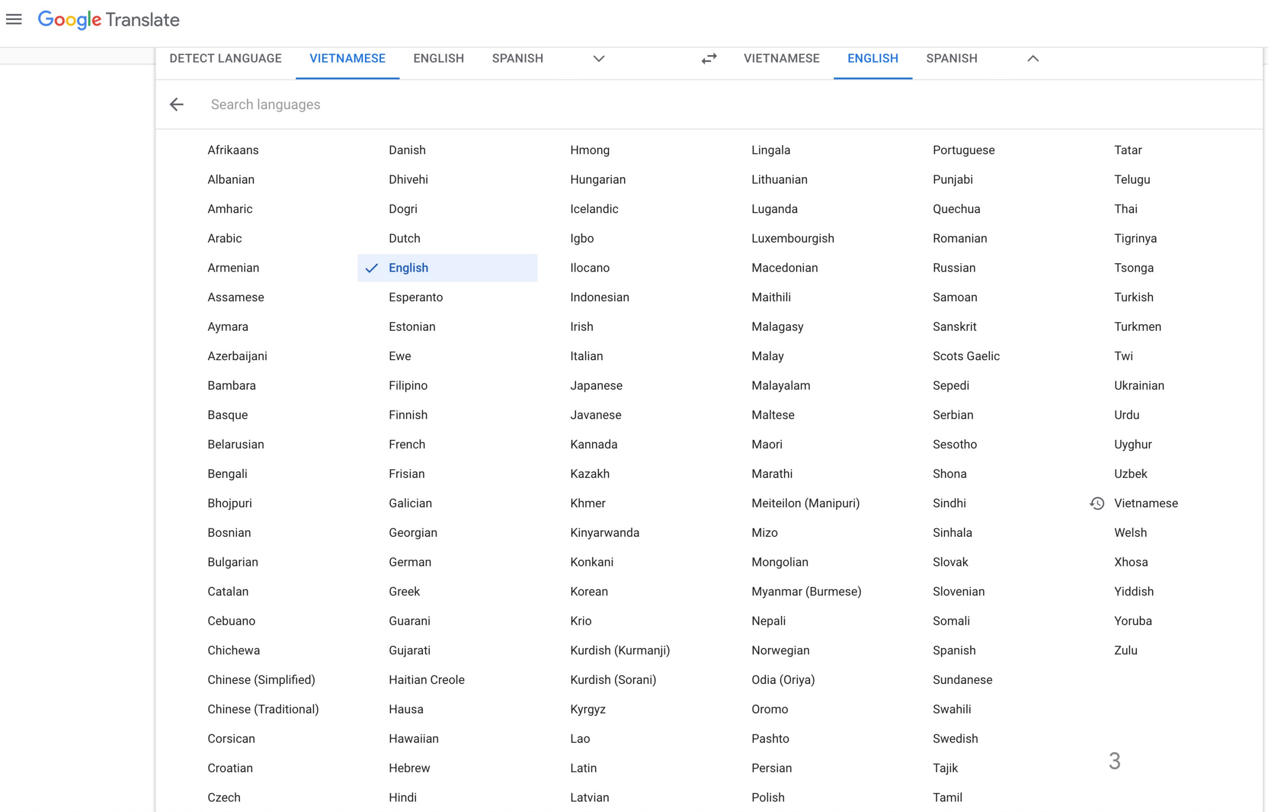
Machine Translation



Translation (Experimental Version)

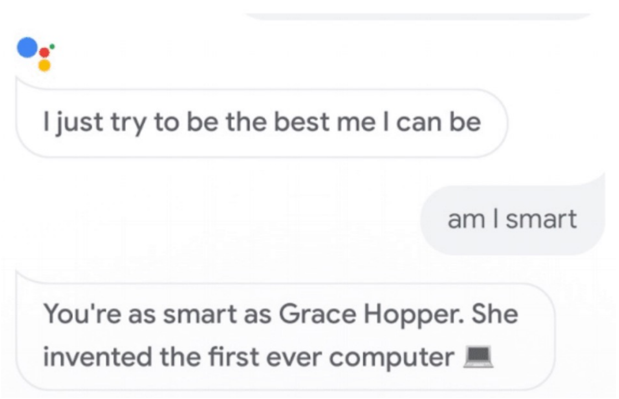
Xử lý ngôn ngữ tự nhiên

Natural Language Processing



Conversational Agents

- Conversational agents or dialogue systems contain:
  - Speech recognition
  - Language analysis
  - Dialogue processing
  - Information retrieval
  - Text to speech



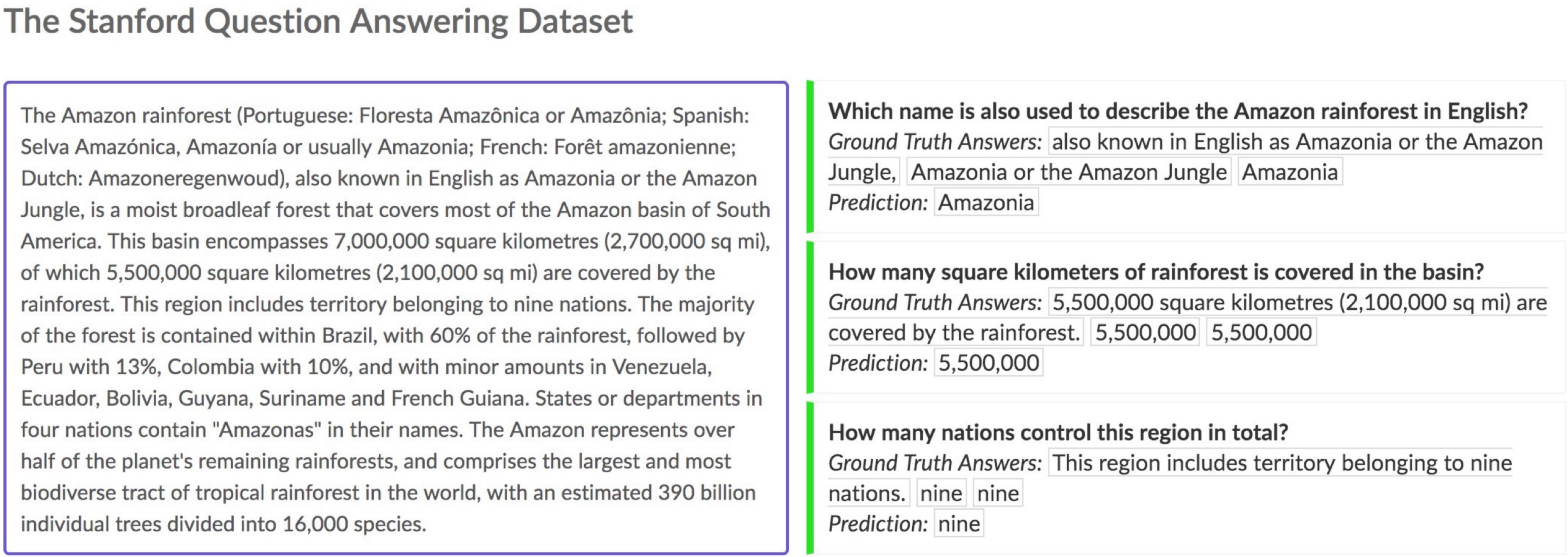
I just try to be the best me I can be

am I smart

You're as smart as Grace Hopper. She invented the first ever computer 💻

Slide content taken from https://www.dropbox.com/s/cz4zmb0l95p9r4y/NLP_1_intro.pdf

Machine Comprehension Question Answering



The Stanford Question Answering Dataset

The Amazon rainforest (Portuguese: Floresta Amazônica or Amazônia; Spanish: Selva Amazónica, Amazonía or usually Amazonia; French: Forêt amazonienne; Dutch: Amazoneregenwoud), also known in English as Amazonia or the Amazon Jungle, is a moist broadleaf forest that covers most of the Amazon basin of South America. This basin encompasses 7,000,000 square kilometres (2,700,000 sq mi), of which 5,500,000 square kilometres (2,100,000 sq mi) are covered by the rainforest. This region includes territory belonging to nine nations. The majority of the forest is contained within Brazil, with 60% of the rainforest, followed by Peru with 13%, Colombia with 10%, and with minor amounts in Venezuela, Ecuador, Bolivia, Guyana, Suriname and French Guiana. States or departments in four nations contain "Amazonas" in their names. The Amazon represents over half of the planet's remaining rainforests, and comprises the largest and most biodiverse tract of tropical rainforest in the world, with an estimated 390 billion individual trees divided into 16,000 species.

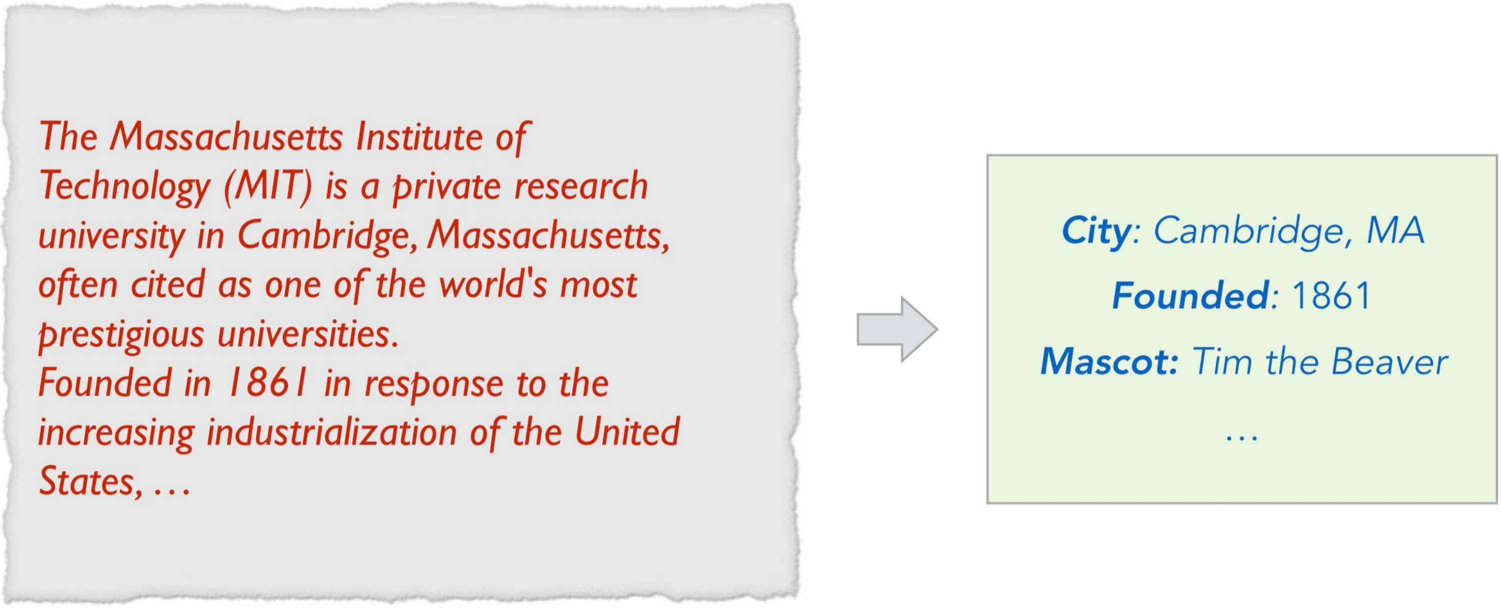
Which name is also used to describe the Amazon rainforest in English?
Ground Truth Answers: also known in English as Amazonia or the Amazon Jungle, Amazonia or the Amazon Jungle Amazonia
Prediction: Amazonia

How many square kilometers of rainforest is covered in the basin?
Ground Truth Answers: 5,500,000 square kilometres (2,100,000 sq mi) are covered by the rainforest. 5,500,000 5,500,000
Prediction: 5,500,000

How many nations control this region in total?
Ground Truth Answers: This region includes territory belonging to nine nations. nine nine
Prediction: nine

Slide content taken from https://princeton-nlp.github.io/cos484/lectures/lec1.pdf

Information Extraction



The Massachusetts Institute of Technology (MIT) is a private research university in Cambridge, Massachusetts, often cited as one of the world's most prestigious universities.
Founded in 1861 in response to the increasing industrialization of the United States, …

City: *Cambridge, MA*

Founded: 1861

Mascot: *Tim the Beaver*

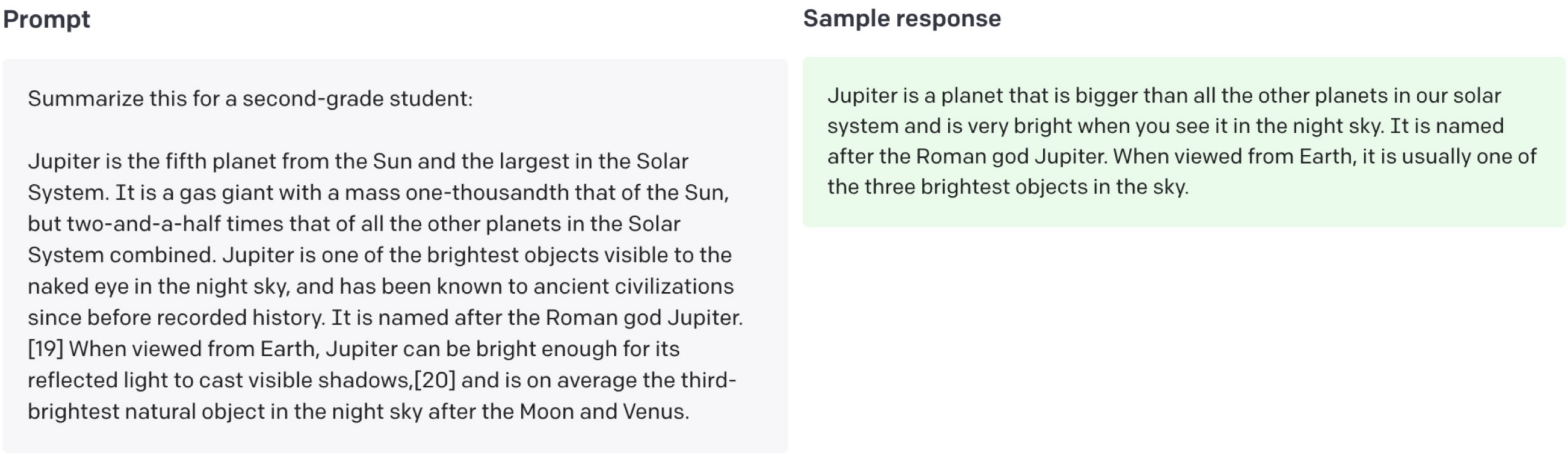
…

Article

Database

Slide content taken from https://princeton-nlp.github.io/cos484/lectures/lec1.pdf

Language Modeling/Generation Applications



Prompt

Summarize this for a second-grade student:

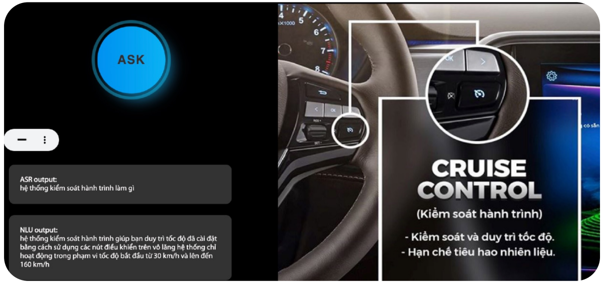
Jupiter is the fifth planet from the Sun and the largest in the Solar System. It is a gas giant with a mass one-thousandth that of the Sun, but two-and-a-half times that of all the other planets in the Solar System combined. Jupiter is one of the brightest objects visible to the naked eye in the night sky, and has been known to ancient civilizations since before recorded history. It is named after the Roman god Jupiter.[19] When viewed from Earth, Jupiter can be bright enough for its reflected light to cast visible shadows,[20] and is on average the third-brightest natural object in the night sky after the Moon and Venus.

Sample response

Jupiter is a planet that is bigger than all the other planets in our solar system and is very bright when you see it in the night sky. It is named after the Roman god Jupiter. When viewed from Earth, it is usually one of the three brightest objects in the sky.

Slide content taken from https://princeton-nlp.github.io/cos484/lectures/lec1.pdf

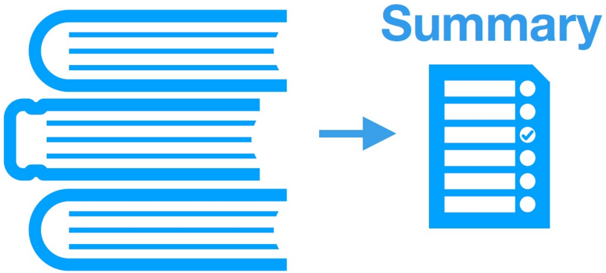
Natural Language Processing at VinAI


ASK
CRUISE CONTROL
(Kiểm soát hành trình)
- Kiểm soát và duy trì tốc độ.
- Hạn chế tiêu hao nhiên liệu.

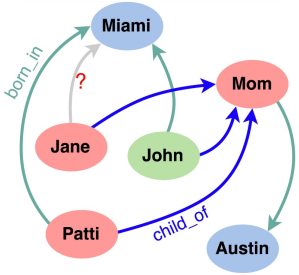
ASR output:
hệ thống kiểm soát hành trình làm gì

NLU output:
hệ thống kiểm soát hành trình giúp bạn duy trì tốc độ đã cài đặt bằng cách sử dụng các nút điều khiển trên vô lăng hệ thống chỉ hoạt động trong phạm vi tốc độ bắt đầu từ 30 km/h và lên đến 160 km/h

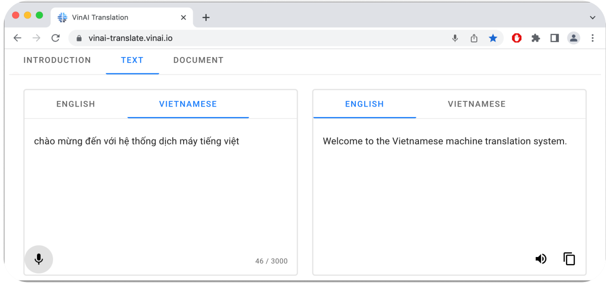
Question Answering and Dialogue System



Text Classification and Summarization



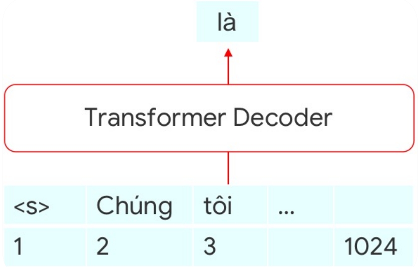
Relation and Event Extraction



chào mừng đến với hệ thống dịch máy tiếng việt

Welcome to the Vietnamese machine translation system.

Text and Speech Translation



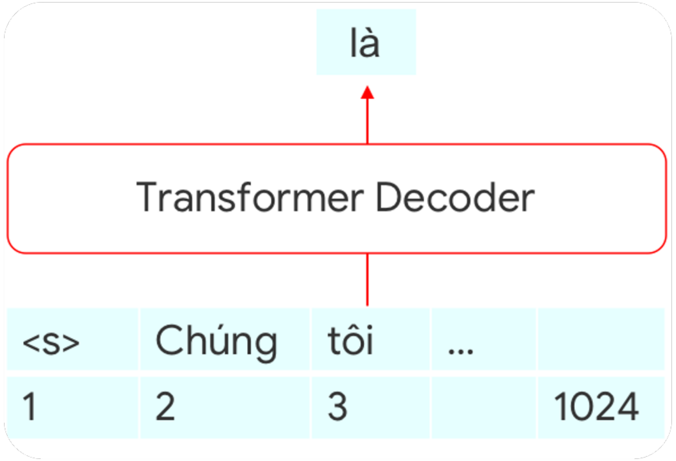
Language Modeling

32 research papers published

- EMNLP (11), ACL (3), NAACL (3)
- InterSpeech (6), AAAI (5), ICLR (2), NeurIPS (1), IJCAI (1)
- ICLR 2021 Outstanding Paper Award

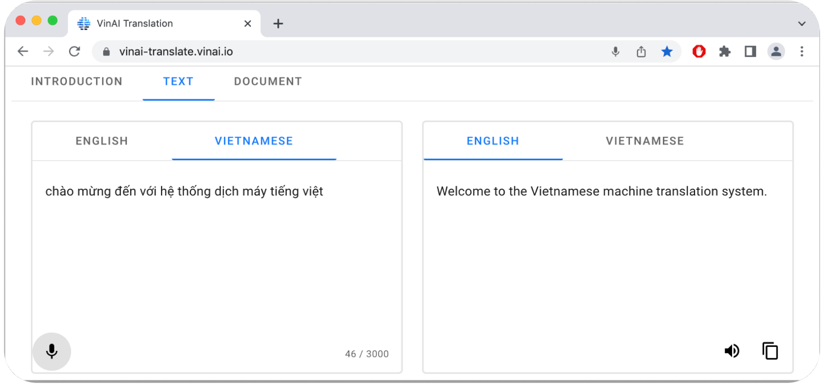
Research to enhance Vietnamese NLP capability

- PhoBERT and BARTpho: First large-scale pre-trained language models for Vietnamese
  - Improve state-of-the-art performance for various Vietnamese NLP tasks
  - 100K+ downloads
  - https://github.com/VinAIResearch/PhoBERT
  - https://github.com/VinAIResearch/BARTpho



Research to enhance Vietnamese NLP capability

- State-of-the-art translation models pre-trained for Vietnamese-to-English and English-to-Vietnamese (https://github.com/VinAIResearch/VinAI_Translate)
- PhoMT: A high-quality and large-scale Vietnamese-English parallel dataset (https://github.com/VinAIResearch/PhoMT)
- PhoST: A high-quality and large-scale dataset with 508 audio hours for English-Vietnamese speech translation (https://github.com/VinAIResearch/PhoST)



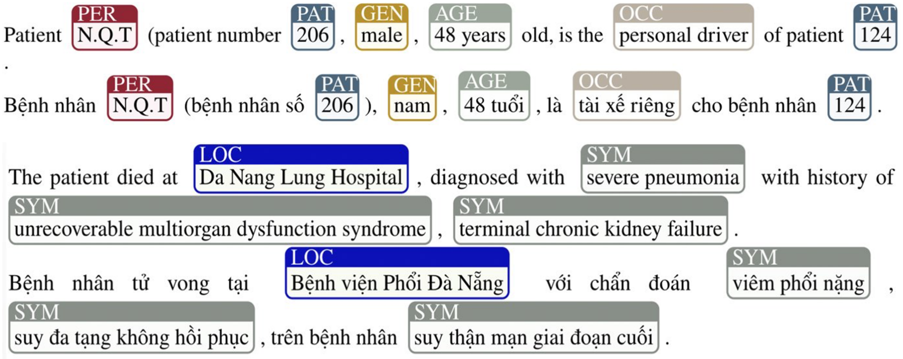
chào mừng đến với hệ thống dịch máy tiếng việt

46 / 3000

Welcome to the Vietnamese machine translation system.

Research to enhance Vietnamese NLP capability

- Develop named entity recognition task in COVID-19 specified domain: potentially impact research and applications
- Provide new dataset for recognizing COVID-19 related named entities in Vietnamese
  - 10 entity types with the aim of extracting key information related to COVID-19 patients
  - Entity types are used in the context of not only COVID-19 pandemic but also in other future epidemics
- https://github.com/VinAIResearch/PhoNER_COVID19



Patient N.Q.T (patient number 206 , male , 48 years old, is the personal driver of patient 124 .

Bệnh nhân N.Q.T (bệnh nhân số 206), nam , 48 tuổi , là tài xế riêng cho bệnh nhân 124 .

The patient died at Da Nang Lung Hospital , diagnosed with severe pneumonia with history of unrecoverable multiorgan dysfunction syndrome , terminal chronic kidney failure .

Bệnh nhân tử vong tại Bệnh viện Phổi Đà Nẵng với chẩn đoán viêm phổi nặng , suy đa tạng không hồi phục , trên bệnh nhân suy thận mạn giai đoạn cuối .

Research to enhance Vietnamese NLP capability

- **Intent detection and slot filling for Vietnamese (Interspeech 2021)**
- From disfluency detection to intent detection and slot filling (Interspeech 2022)

Motivation

- Intent detection and slot filling are important NLU tasks
  - Intent detection aims to identify speaker's intent from a given utterance
  - Slot filling is to extract from the utterance the correct argument value for the slots of the intent
- Example
  - **Query**: what flights are available from chicago to baltimore on thursday morning
  - **Intent**: flight info
  - **Slots**:

    - from_city: chicago

    - to_city: baltimore

    - depart_date: thursday

    - depart_time: morning

Motivation

- Vietnamese intent detection and slot filling
    - Only one Vietnamese dataset is relevant to intent detection, which is a dialog act corpus containing ISO-24617-2 based annotations over communication acts [1]
        - This corpus is not publicly available for the research community
    - To the best of our knowledge, *there is no public Vietnamese dataset available specifically for either intent detection or slot filling*

Motivation

- Recent research on intent detection and slot filling
  - Jointly learning helps improve performance results [2]
  - Attention mechanisms are employed to incorporate intent context information into slot filling via an utterance representation
  - We can incorporate more explicit intent context information via a “soft” intent label embedding

Motivation

- Contributions of our work
  - We introduce the *first public intent detection and slot filling dataset* - named **PhoATIS** - for Vietnamese
  - We present *a new joint model JointIDSF* for intent detection and slot filling, that extends JointBERT+CRF [2] with an intent-slot attention layer

Our PhoATIS dataset

- Manually translate each English utterance from the well-known intent detection and slot filling dataset ATIS [3] into Vietnamese
  - We require modifications to ensure that our Vietnamese utterances are natural, fitting in real-world scenarios in Vietnam and of high-quality
    - e.g. replacing American-popular slot values, such as locations, airline names and the like, with their counterparts in Vietnam
- Manually project intent and slot annotations from each ATIS English utterance to its Vietnamese-translated version
- Manually fix inconsistencies among projected annotations in our Vietnamese dataset

Our PhoATIS dataset

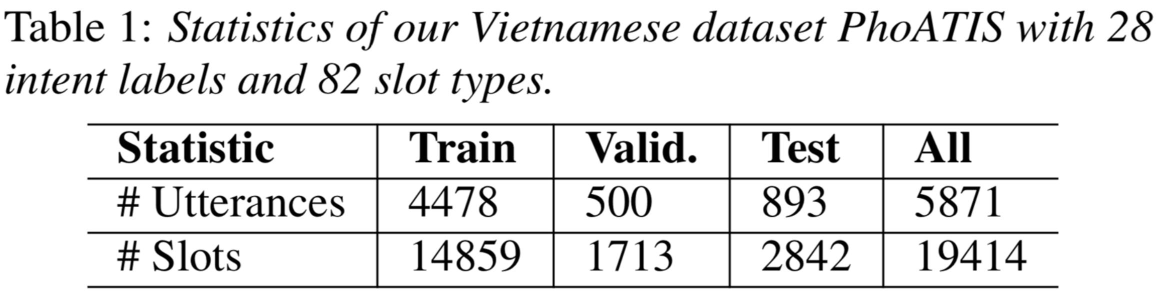


Table 1: *Statistics of our Vietnamese dataset PhoATIS with 28 intent labels and 82 slot types.*

Statistic	**Train**	**Valid.**	**Test**	**All**
# Utterances	4478	500	893	5871
# Slots	14859	1713	2842	19414

- Automatic Vietnamese word segmentation is performed by employing VnCoreNLP [4] to obtain a word-level variant of the dataset
    - The outputs of automatic Vietnamese word segmentation do not affect the span boundaries of slot annotations

Syllable level: tôi cần đến phú quốc vào tối thứ tư từ đà lạt

Word level: tôi cần đến phú_quốc vào tối thứ_tư từ đà_lạt

Our new model JointIDSF

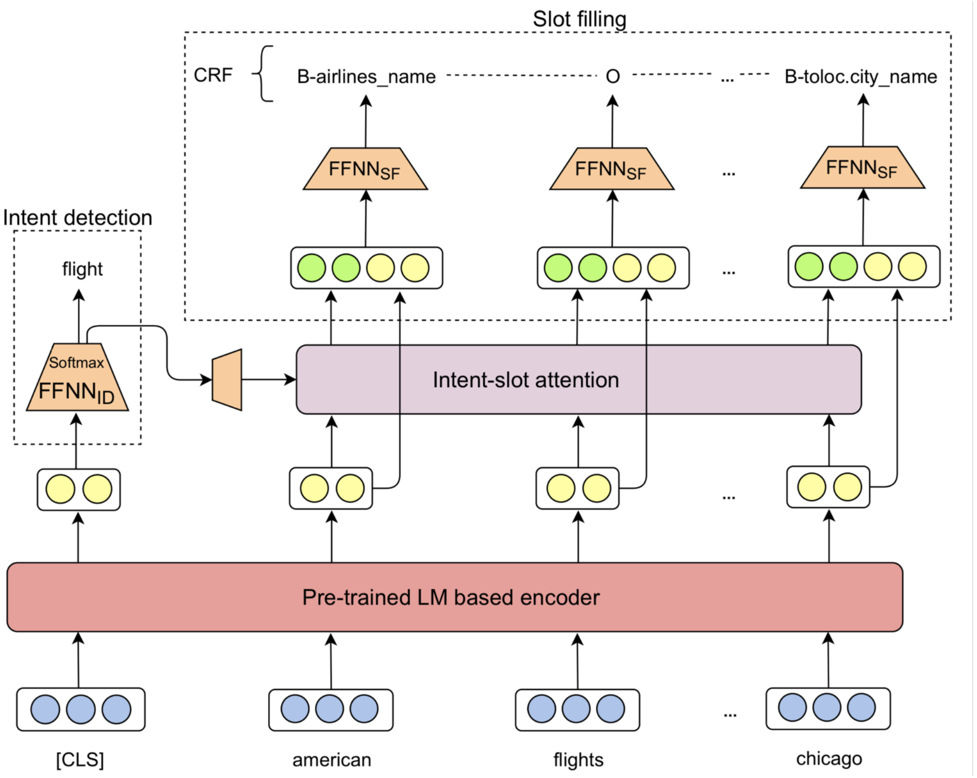
- Encoding layer

$$\mathbf{c}_i = \text{PretrainedLM}(w_{0:n}, i) \qquad (1)$$

- Intent detection layer

$$\mathbf{p} = \mathsf{softmax}(\text{FFNN}_{\text{ID}}(\mathbf{c}_0)) \qquad (2)$$

A cross-entropy objective loss $\mathbf{L}_{\text{ID}}$ is calculated for intent classification during training



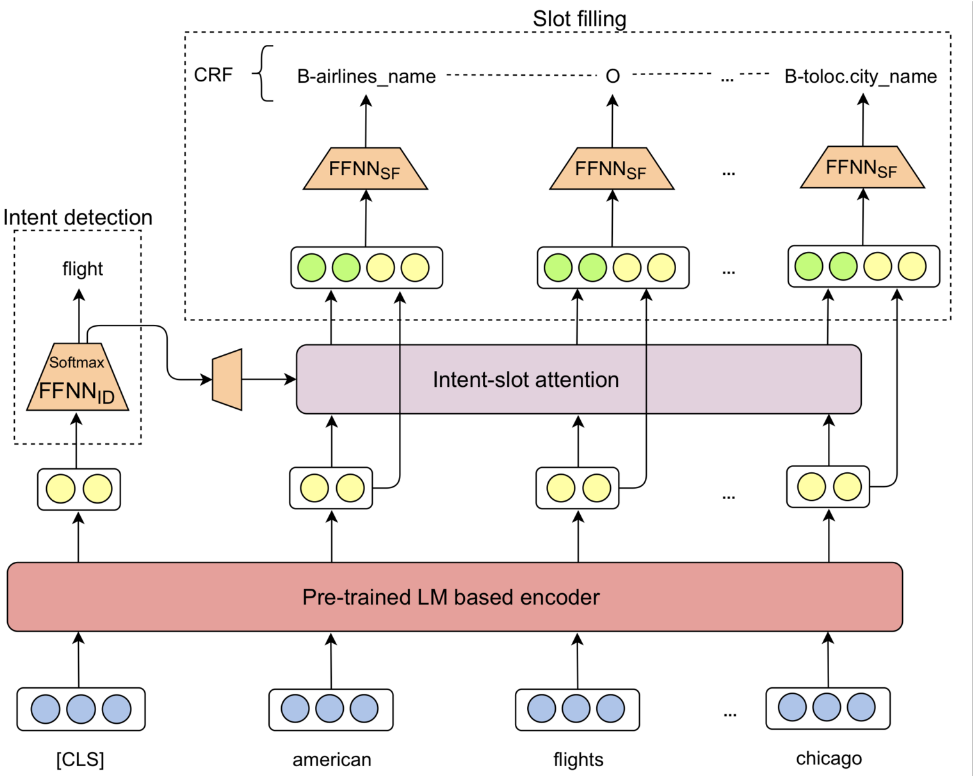
Our new model JointIDSF

- Intent-slot attention layer

$$\mathbf{w} = \mathbf{W}\mathbf{p} \tag{3}$$

$$\alpha_i = \frac{\exp(\mathbf{w}^\top \mathbf{c}_i)}{\sum_{j=1}^{n} \exp(\mathbf{w}^\top \mathbf{c}_j)} \tag{4}$$

$$\mathbf{s}_i = \alpha_i \mathbf{w} \tag{5}$$



Our new model JointIDSF

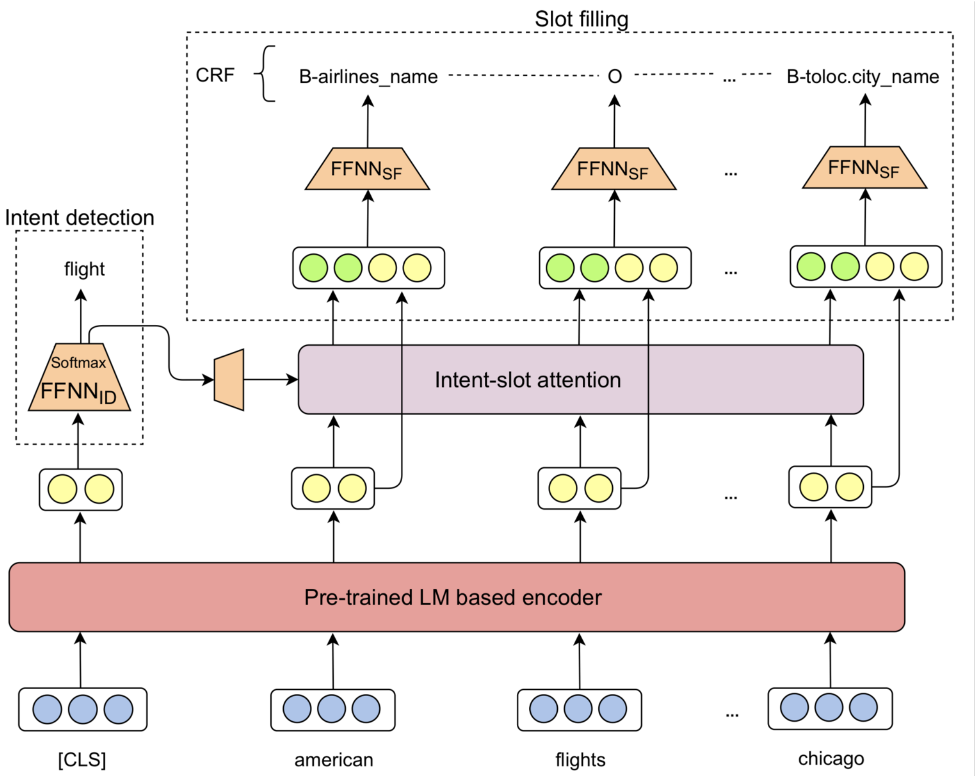
- Slot filling layer

$$\mathbf{v}_i = \mathbf{s}_i \circ \mathbf{c}_i \qquad (6)$$

$$\mathbf{h}_i = \mathrm{FFNN}_{\mathrm{SF}}(\mathbf{v}_i) \qquad (7)$$

h$_i$ vectors are fed into a linear-chain CRF predictor for slot type prediction

A cross-entropy objective loss **L**$_{SF}$ is calculated for intent classification during training

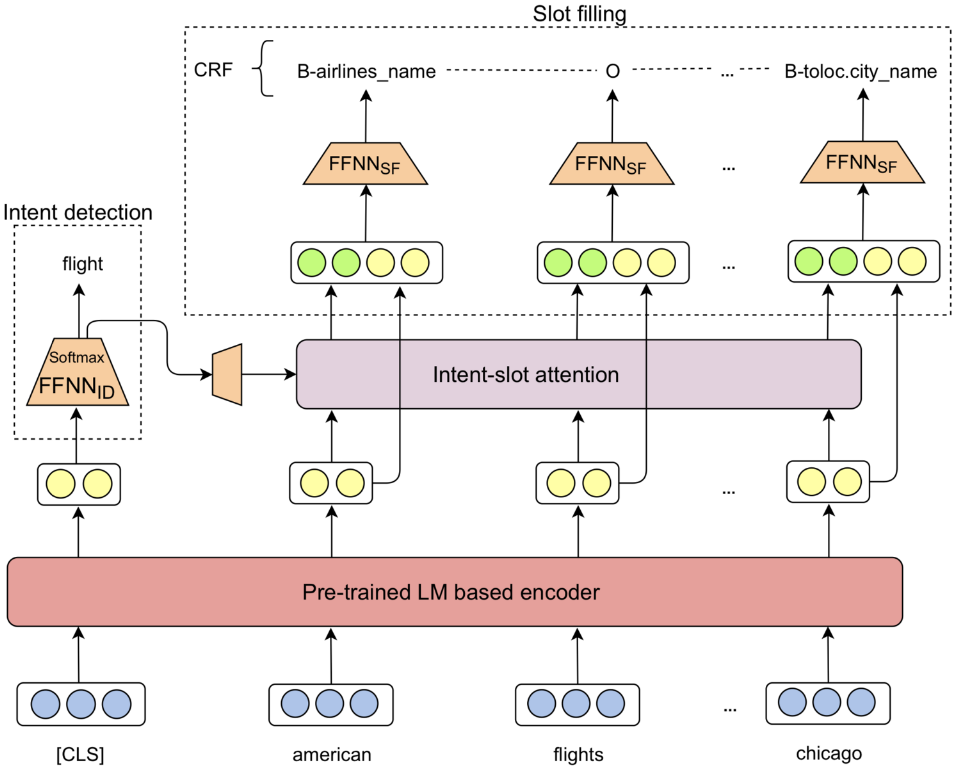


Our new model JointIDSF

- Joint training

$$\mathcal{L} = \lambda\mathcal{L}_{\text{ID}} + (1-\lambda)\mathcal{L}_{\text{SF}} \quad (8)$$

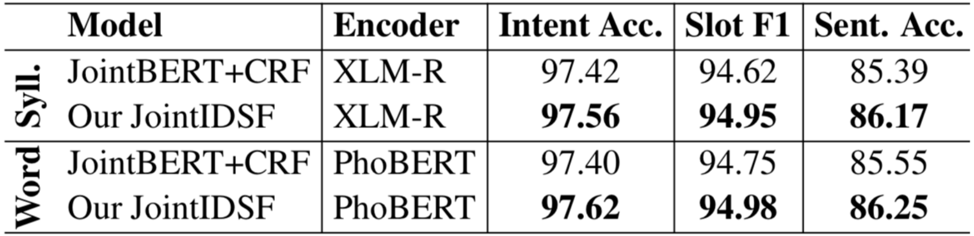
- Compared to JointBERT+CRF
  - We introduce the intent-slot attention layer to explicitly incorporate intent context information into slot filling



Experiments

- We conduct experiments on our dataset to study:
  - A quantitative comparison between our model JointIDSF and the baseline JointBERT+CRF
  - The influence of Vietnamese word segmentation (here, input utterances can be formed in either syllable or word level)
  - The usefulness of pre-trained language model-based encoders XLM-R [5] and PhoBERT [6]
    - XLM-R is pre-trained on a 2.5TB multilingual dataset that contains 137GB of syllable-level Vietnamese texts
    - PhoBERT is pre-trained on 20GB of word-level Vietnamese texts

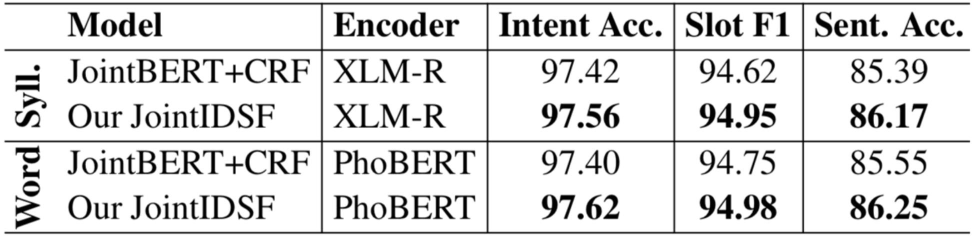
Experiments



	Model	Encoder	Intent Acc.	Slot F1	Sent. Acc.
Syll.	JointBERT+CRF	XLM-R	97.42	94.62	85.39
	Our JointIDSF	XLM-R	**97.56**	**94.95**	**86.17**
Word	JointBERT+CRF	PhoBERT	97.40	94.75	85.55
	Our JointIDSF	PhoBERT	**97.62**	**94.98**	**86.25**

- JointIDSF significantly outperforms JointBERT+CRF
- The highest improvements are accounted for the sentence accuracy
  - Our intent-slot attention layer helps better capture correlations between intent labels and slots in the same utterances

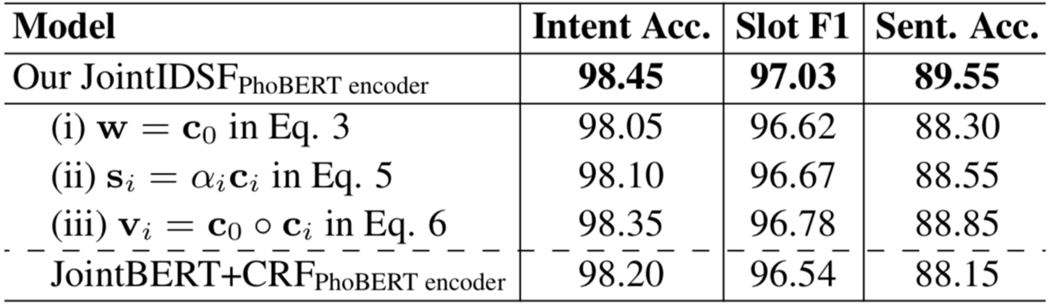
Experiments



	Model	Encoder	Intent Acc.	Slot F1	Sent. Acc.
Syll.	JointBERT+CRF	XLM-R	97.42	94.62	85.39
	Our JointIDSF	XLM-R	**97.56**	**94.95**	**86.17**
Word	JointBERT+CRF	PhoBERT	97.40	94.75	85.55
	Our JointIDSF	PhoBERT	**97.62**	**94.98**	**86.25**

- The performances of word-level models are higher, but not significantly, than their syllable-level counterparts
  - Automatic Vietnamese word segmentation and the pre-trained monolingual language model PhoBERT are less effective for these Vietnamese intent detection and slot filling tasks than for other Vietnamese NLP tasks
  - Possible reason: the utterances in our dataset are domain-specific and medium-length ones with an average length of 15 word tokens

Experiments

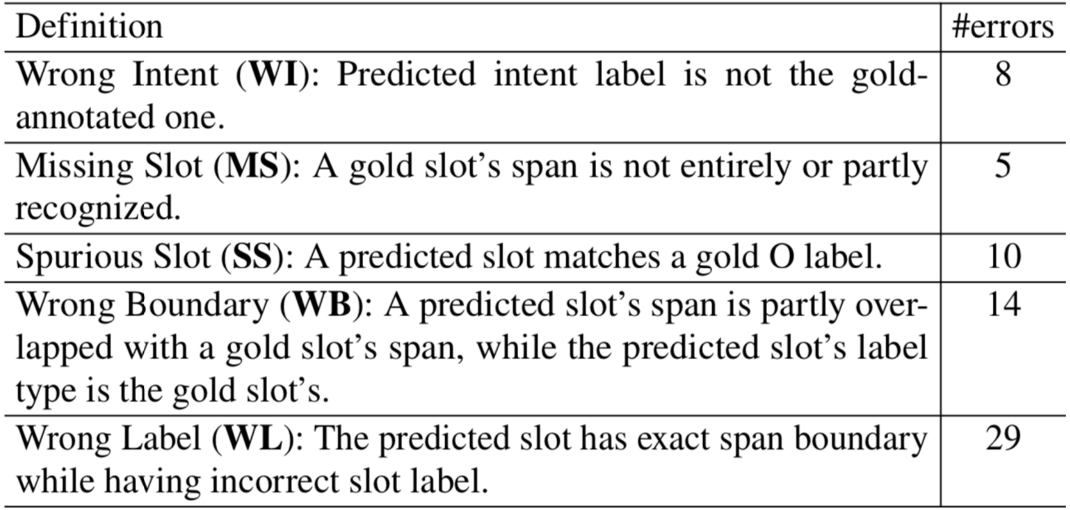


Model	Intent Acc.	Slot F1	Sent. Acc.
Our JointIDSF$_{\text{PhoBERT encoder}}$	**98.45**	**97.03**	**89.55**
(i) $\mathbf{w} = \mathbf{c}_0$ in Eq. 3	98.05	96.62	88.30
(ii) $\mathbf{s}_i = \alpha_i \mathbf{c}_i$ in Eq. 5	98.10	96.67	88.55
(iii) $\mathbf{v}_i = \mathbf{c}_0 \circ \mathbf{c}_i$ in Eq. 6	98.35	96.78	88.85
JointBERT+CRF$_{\text{PhoBERT encoder}}$	98.20	96.54	88.15

- (i) Using the "[CLS]"-based utterance context representation instead of the intent label representation
- (ii) Using the scalar multiplication between the attention weights and the slot vector representations $\mathbf{c}_i$ instead of the intent label representation
- (iii) Concatenating the utterance context representation instead of the attention layer's output vectors to all the slot vector representations

Experiments

- Error analysis



Definition	#errors
Wrong Intent (**WI**): Predicted intent label is not the gold-annotated one.	8
Missing Slot (**MS**): A gold slot's span is not entirely or partly recognized.	5
Spurious Slot (**SS**): A predicted slot matches a gold O label.	10
Wrong Boundary (**WB**): A predicted slot's span is partly overlapped with a gold slot's span, while the predicted slot's label type is the gold slot's.	14
Wrong Label (**WL**): The predicted slot has exact span boundary while having incorrect slot label.	29

Experiments

- The most frequently appeared error is 29 cases accounted for the Wrong Label error category
    - These cases often are induced by ambiguities between which the "*departure*" part is and which the "*arrival*" part is in an utterance since many utterances do not have an explicit context
    - Given the utterance "*tôi cần đến phú_quốc vào tối thứ_tư từ đà_lạt*" (I need to go to Phu Quoc on Wednesday's night from Da Lat)
        - It is relatively ambiguous to determine whether the phrase "*tối thứ_tư*" (Wednesday's night) refers to as an arrival time or a departure time without a clearer context

Experiments

- There are 8 errors counted for the Wrong Intent category
  - Most of them are induced by the multi-intent labels since the model is likely to predict the most clearly manifested or first appeared intent
  - The model predicts an intent label of “*airfare*” instead of the gold one “*airfare#flight_time*” for the utterance “*cho tôi biết chi_phí và thời_gian của các chuyến bay từ phú_quốc đến cam_ranh*” (show me the cost and time for flights from Phu Quoc to Cam Ranh)

Experiments

- There are 5 and 10 errors counted for the error categories Missing Slot and Spurious Slot, respectively
    - The model is often ambiguous about the slot types that rarely appear in the training set such as “*connect*”, “*airport_code*” and the like
- The Wrong Boundary error category has 14 error cases that are related to multi-word spanned slots

Main takeaways

- We have presented the first public dataset for Vietnamese intent detection and slot filling
- We also have proposed an effective model, namely JointIDSF, for jointly learning intent detection and slot filling
- We empirically conduct experiments and perform a detailed error analysis on our dataset, and show that: JointIDSF significantly outperforms JointBERT+CRF
- We publicly release our dataset and the implementation of our model at: https://github.com/VinAIResearch/JointIDSF

Research to enhance Vietnamese NLP capability

- Intent detection and slot filling for Vietnamese (Interspeech 2021)
- **From disfluency detection to intent detection and slot filling (Interspeech 2022)**

Motivation

- Disfluency
  - In natural conversation, humans inevitably produce interruptions in their speech, which is formally referred to as *disfluency*
    - E.g. : is there any *train um no* flight from Ha Noi to Da Lat?
  - Modern Spoken Language Understanding (SLU) models are primarily trained on curated and cleaned input without disfluencies

⇨ Disfluencies might have negative effects on the performances of downstream SLU tasks

⇨ Disfluency detection that detects and removes disfluencies to produce fluent versions of disfluent inputs is crucial in real-world applications

Motivation

- Disfluency detection’s effects on downstream SLU task investigation
  - Most previous works study the disfluency detection task isolatedly and evaluate the task using gold disfluency annotations
  - Investigation of disfluency detection’s influence on downstream tasks is relatively limited
    - Including punctuation restoration [7], machine translation [8], syntactic parsing [9] and question answering [10]

Motivation

- Disfluency detection’s effects on intent detection and slot filling
  - Intent detection and slot filling: important SLU tasks
  - To the best of our knowledge, no study has investigated the effect of disfluencies on the intent detection and slot filling tasks
  - There is no available dataset containing linguistic annotations over both disfluencies, intents, and the slots of the intents

⇨ Research question: “*How disfluencies affect two important downstream SLU tasks intent detection and slot filling*?”

Motivation

- In this work
  - We present the first study that investigates the influence of disfluency detection on the downstream intent detection and slot filling tasks by:
    - Creating a dataset with disfluency annotations by manually adding contextual disfluencies as distractors into the Vietnamese fluent dataset PhoATIS
    - Conducting experiments using strong baselines for disfluency detection and joint intent detection and slot filling, which are based on pre-trained language models

Our dataset

- Manually add contextual disfluencies as distractors into the intent detection and slot filling dataset PhoATIS (5871 fluent utterances in total)
- The disfluent version should satisfy the following requirements:
  - semantically equivalent to the original one
  - natural in terms of human usage, grammatical errors, and meaningful distractors
  - containing disfluent words that are corrected by following intent or slot value keywords in the original utterance
  - containing both disfluent Reparandum- and Interragnum-type words where possible

    Reparandum: word or words that the speaker intends to be abandoned or corrected by the following words

    Interregnum: filled pauses, discourse cue words and the like

is there any train um no flight from Ha Noi to Da Lat?

Our dataset

- Example

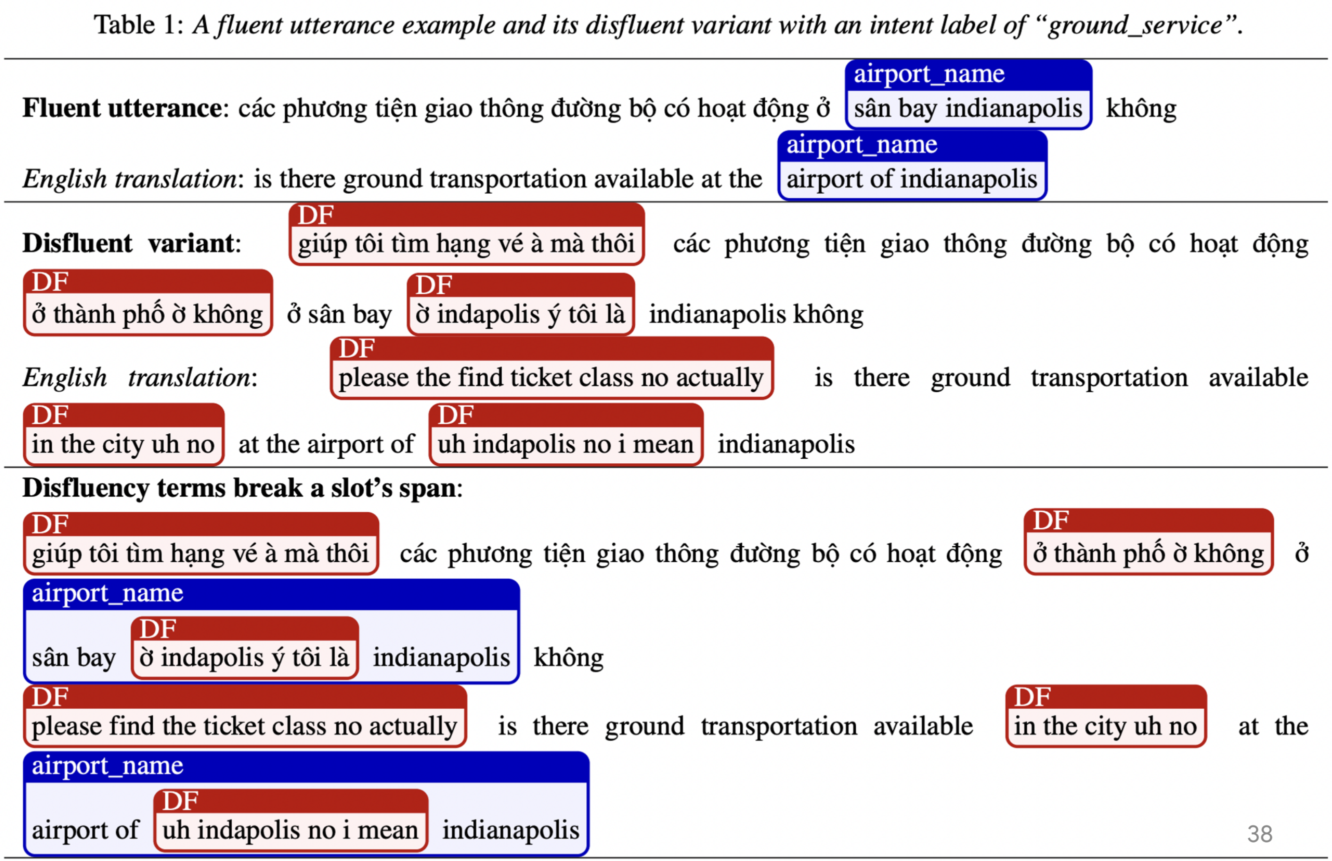


Table 1: *A fluent utterance example and its disfluent variant with an intent label of "ground_service".*

Fluent utterance: các phương tiện giao thông đường bộ có hoạt động ở airport_name sân bay indianapolis không

English translation: is there ground transportation available at the airport_name airport of indianapolis

Disfluent variant: DF giúp tôi tìm hạng vé à mà thôi các phương tiện giao thông đường bộ có hoạt động DF ở thành phố ở không ở sân bay DF ở indapolis ý tôi là indianapolis không

English translation: DF please the find ticket class no actually is there ground transportation available DF in the city uh no at the airport of DF uh indapolis no i mean indianapolis

Disfluency terms break a slot's span:

DF giúp tôi tìm hạng vé à mà thôi các phương tiện giao thông đường bộ có hoạt động DF ở thành phố ở không ở airport_name sân bay DF ở indapolis ý tôi là indianapolis không

DF please find the ticket class no actually is there ground transportation available DF in the city uh no at the airport_name airport of DF uh indapolis no i mean indianapolis

Our dataset

- Automatic Vietnamese word segmentation is performed by employing VnCoreNLP to obtain a word-level variant of the dataset

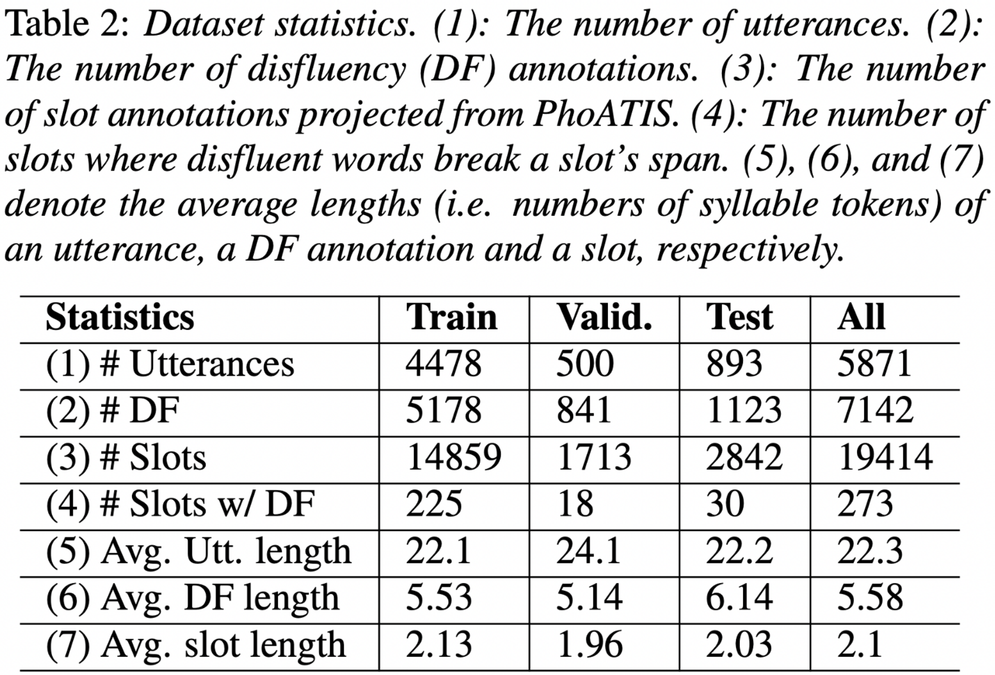


Table 2: *Dataset statistics. (1): The number of utterances. (2): The number of disfluency (DF) annotations. (3): The number of slot annotations projected from PhoATIS. (4): The number of slots where disfluent words break a slot's span. (5), (6), and (7) denote the average lengths (i.e. numbers of syllable tokens) of an utterance, a DF annotation and a slot, respectively.*

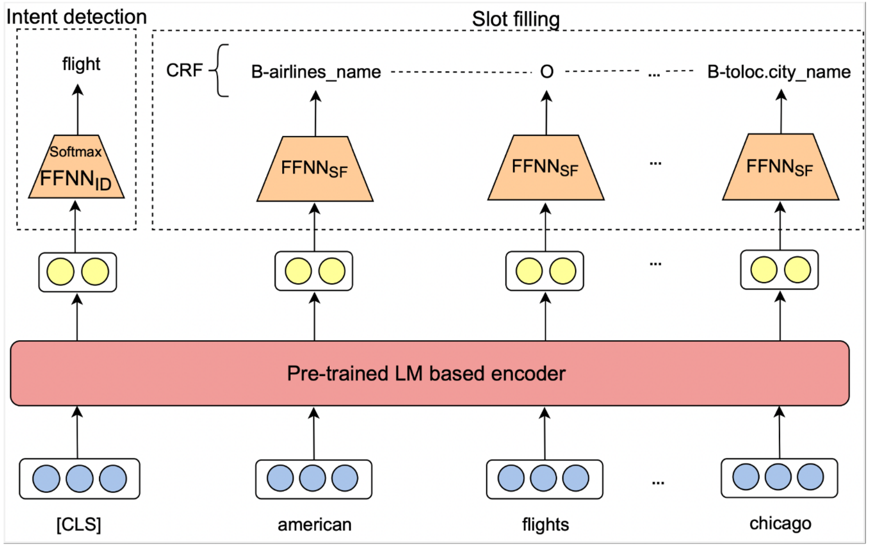
Statistics	**Train**	**Valid.**	**Test**	**All**
(1) # Utterances	4478	500	893	5871
(2) # DF	5178	841	1123	7142
(3) # Slots	14859	1713	2842	19414
(4) # Slots w/ DF	225	18	30	273
(5) Avg. Utt. length	22.1	24.1	22.2	22.3
(6) Avg. DF length	5.53	5.14	6.14	5.58
(7) Avg. slot length	2.13	1.96	2.03	2.1

Our cascaded approach

- Two separate models: disfluency detection and joint intent detection and slot filling
- Given an input utterance:
  - First, the disfluency detection model to automatically identify disfluent terms and then remove these identified terms to generate a “fluent” variant
  - Second, the “fluent” variant is fed into the joint intent detection and slot filling model to predict intent and slot types

Our cascaded approach

- Disfluency detection model
  - The Vietnamese disfluency detection task is formulated as a sequence labeling problem with BIO tagging scheme (B-DF, I-DF, and O)
  - A pre-trained LM-based encoder is employed to generate contextualized latent feature embeddings for the input tokens
  - Each latent feature embedding is then fed into a linear-chain CRF layer for disfluency label prediction (Similar to slot filling)
- Joint intent detection and slot filling

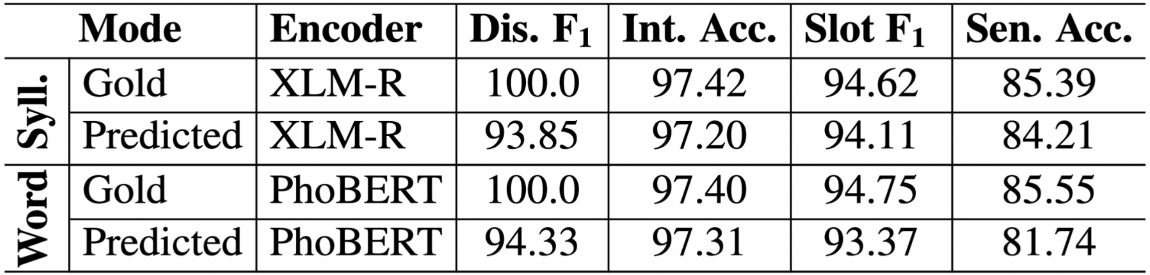


Our cascaded approach

- Implementation details
    - The disfluency detection model is trained using the training set of disfluent utterances with disfluency annotations only
    - The JointBERT+CRF model for joint intent detection and slot filling is trained using the gold fluent PhoATIS training set
    - We use XLM-R for the syllable-level input and PhoBERT for the word-level input

Experimental results

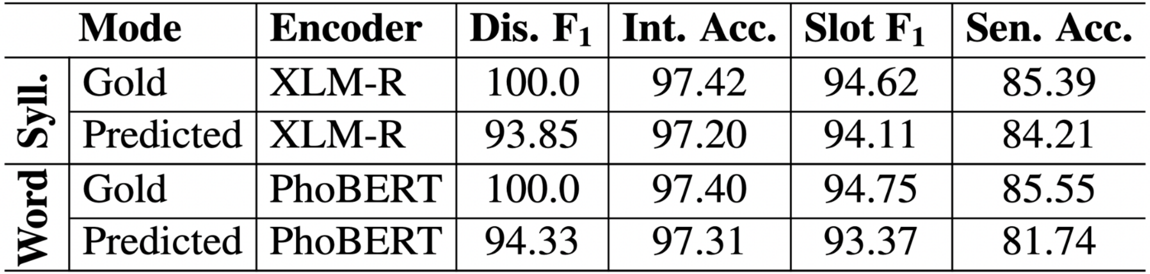
- Main results
  - For disfluency detection, the model that employs PhoBERT encoder obtains a higher F1 score than the one employing XLM-R encoder
  - As syllables constitute words, resulting in disfluent phrases at the syllable level which are “longer” ⇒ the model thus likely finds it more difficult to predict “longer” disfluent phrases at the syllable level



	Mode	Encoder	Dis. F_1	Int. Acc.	Slot F_1	Sen. Acc.
Syll.	Gold	XLM-R	100.0	97.42	94.62	85.39
	Predicted	XLM-R	93.85	97.20	94.11	84.21
Word	Gold	PhoBERT	100.0	97.40	94.75	85.55
	Predicted	PhoBERT	94.33	97.31	93.37	81.74

Experimental results

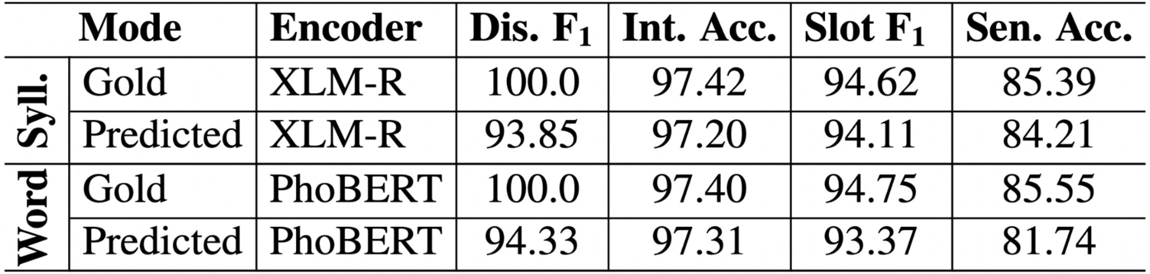
- Main results
  - Effect of disfluency detection on the two downstream tasks: all downstream performance scores are decreased
  - It can be explained by errors propagation from the disfluency detection phase which generates utterances with missing fluent tokens and left-over disfluent tokens



	Mode	Encoder	Dis. F_1	Int. Acc.	Slot F_1	Sen. Acc.
Syll.	Gold	XLM-R	100.0	97.42	94.62	85.39
	Predicted	XLM-R	93.85	97.20	94.11	84.21
Word	Gold	PhoBERT	100.0	97.40	94.75	85.55
	Predicted	PhoBERT	94.33	97.31	93.37	81.74

Experimental results

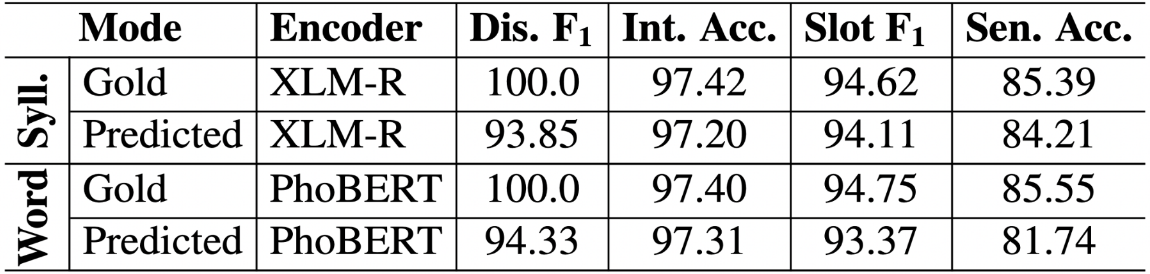
- Main results
  - Final sentence-level accuracies illustrate the strong negative impact of disfluency detection on intent detection and slot filling



	Mode	Encoder	Dis. F_1	Int. Acc.	Slot F_1	Sen. Acc.
Syll.	Gold	XLM-R	100.0	97.42	94.62	85.39
	Predicted	XLM-R	93.85	97.20	94.11	84.21
Word	Gold	PhoBERT	100.0	97.40	94.75	85.55
	Predicted	PhoBERT	94.33	97.31	93.37	81.74

Experimental results

- Main results
  - In disfluent context, the word-level model produces lower intent detection and slot filling performances than its syllable-level counterpart
  - This is opposite to the obtained results with gold fluent input utterances in the table as well as to what is generally found with other Vietnamese NLP tasks in the fluent context



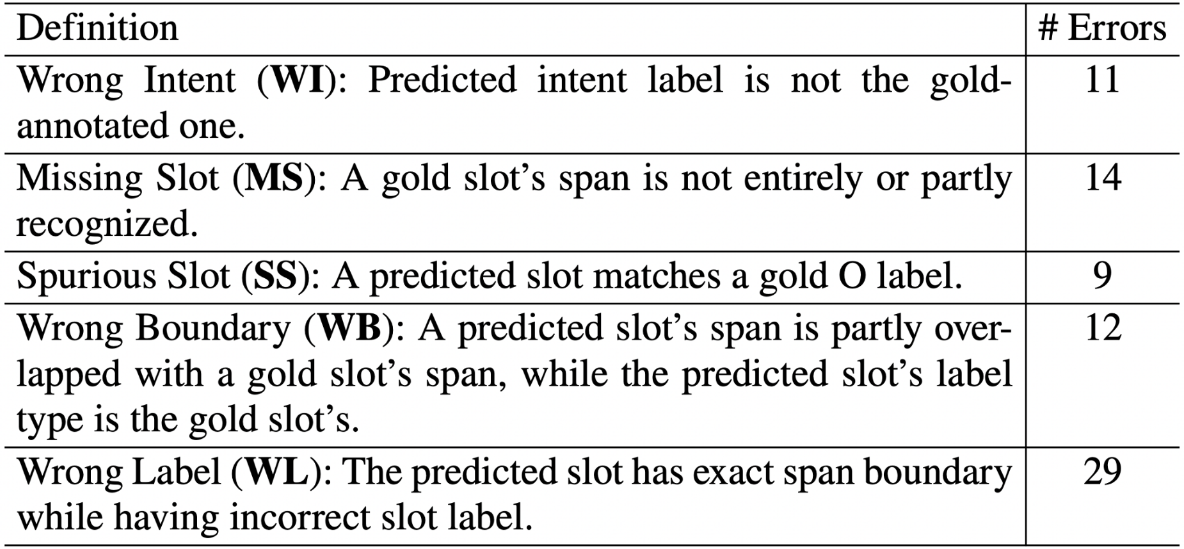
	Mode	Encoder	Dis. F_1	Int. Acc.	Slot F_1	Sen. Acc.
Syll.	Gold	XLM-R	100.0	97.42	94.62	85.39
	Predicted	XLM-R	93.85	97.20	94.11	84.21
Word	Gold	PhoBERT	100.0	97.40	94.75	85.55
	Predicted	PhoBERT	94.33	97.31	93.37	81.74

Experimental results

- Error analysis
  - Disfluency detection errors
    - 53 false positive instances: phrases are fluent but being predicted as disfluent phrases
    - 44 false negative instances: real disfluent phrases are mis-detected or partly recognized
    - Disfluency detection errors are more likely to happen in relatively long sentences with multi-token disfluent phrases and ambiguous contexts

Experimental results

- Error analysis
  - Intent detection and slot filling errors



Definition	# Errors
Wrong Intent (**WI**): Predicted intent label is not the gold-annotated one.	11
Missing Slot (**MS**): A gold slot's span is not entirely or partly recognized.	14
Spurious Slot (**SS**): A predicted slot matches a gold O label.	9
Wrong Boundary (**WB**): A predicted slot's span is partly overlapped with a gold slot's span, while the predicted slot's label type is the gold slot's.	12
Wrong Label (**WL**): The predicted slot has exact span boundary while having incorrect slot label.	29

Experimental results

- Error analysis
  - Intent detection and slot filling errors
    - There are 11 error cases for the Wrong Intent (WI) category
    - Most of them are caused by the multi-intent labels (e.g. "*airfare#flight*")
    - There are also WI cases induced by disfluency detection errors
    - E.g. "*could you please show me the fare of the buses uh no i mean the flights between Hue and Ca Mau*", the disfluency detection model identifies fluent terms "*the fare*" as disfluent terms (and then remove these terms), leading to a wrong prediction of intent "*flight*" (instead of "*airfare*")

Examples are word-orderly translated from Vietnamese for illustration

Experimental results

- Error analysis
  - Intent detection and slot filling errors
    - There are 14 and 9 error cases counted for Missing Slot (MS) and Spurious Slot (SS) categories, respectively
    - These two types of errors are generally caused by the ambiguities over slot types that rarely occur in the training set such as “*connect*” (36/14859 training slot values) or “*economy*” (34/14859)

Experimental results

- Error analysis
  - Intent detection and slot filling errors
    - The Wrong Boundary (WB) category has 12 error cases that are mostly induced by multi-syllable slot values, especially the incorrectly removed fluent tokens, and the incorrectly preserved disfluent ones

      Disfluent version: “*show me the transportation uh no airlines for flights to or from the airport of Tan Son Nhat no actually Doncaster Sheffield*”
      Automatic disfluency removal version: “*show me the airlines for flights to or from the Tan Son Nhat Sheffield airport*”

Examples are word-orderly translated from Vietnamese for illustration

Experimental results

- Error analysis
  - Intent detection and slot filling errors
    - The most common error type is Wrong Label (WL) which contains 29 error cases
    - These errors exist mostly because of the ambiguities between the "*departure*" part and the "*arrival*" part of an utterance
    - Some of the cases are also caused by disfluency detection errors

      Disfluent version: "*i need a flight from Phu Quoc to Ha Noi no actually Ho Chi Minh City and then Ho Chi Minh city to Singapore uhm no Jakarta and from Jakarta to Ha Noi*"
      Automatic disfluency removal version: "*i need a flight from Phu Quoc to actually Ho Chi Minh City and then Ho Chi Minh city Jakarta and from Jakarta to Ha Noi*"

Examples are word-orderly translated from Vietnamese for illustration

Main takeaways

- We present the first empirical study investigating the influence of disfluency detection on two downstream SLU tasks of intent detection and slot filling
- We have created the first dataset with disfluency, intent detection and slot filling annotations for Vietnamese
- We have conducted experiments under the “Cascaded” manner with strong baselines, performed detailed error analysis and showed that:
    - disfluencies cause substantial performance degradation in the intent detection and slot filling tasks
    - the pre-trained monolingual LM PhoBERT is less effective than the pre-trained multilingual LM XLM-R for intent detection and slot filling under the disfluency context
- We publicly release our dataset at: https://github.com/VinAIResearch/PhoATIS_Disfluency

Thanks for your attention!

VinAI
RESEARCH

References

1. A Vietnamese Dialog Act Corpus Based on ISO 24617-2 standard
2. BERT for Joint Intent Classification and Slot Filling
3. Evaluation of spoken language systems: the ATIS domain
4. VnCoreNLP: A Vietnamese Natural Language Processing Toolkit
5. Unsupervised Cross-lingual Representation Learning at Scale
6. PhoBERT: Pre-trained language models for Vietnamese
7. Combining Punctuation and Disfluency Prediction: An Empirical Study
8. Fluent Translations from Disfluent Speech in End-to-End Speech Translation
9. Joint Transition-based Dependency Parsing and Disfluency Detection for Automatic Speech Recognition Texts
10. DisflQA: A Benchmark Dataset for Understanding Disfluencies in Question Answering