BỘ GIÁO DỤC VÀ ĐÀO TẠO
VIỆN HÀN LÂM KHOA HỌC VÀ CÔNG NGHỆ VIỆT NAM
HỌC VIỆN KHOA HỌC VÀ CÔNG NGHỆ

NGUYỄN THỊ THU HIỀN

NGHIÊN CỨU PHƯƠNG PHÁP CHUẨN HOÁ VĂN BẢN VÀ NHẬN DẠNG THỰC THỂ ĐỊNH DANH TRONG NHẬN DẠNG TIẾNG NÓI TIẾNG VIỆT

LUẬN ÁN TIẾN SĨ NGÀNH MÁY TÍNH

HÀ NỘI - 2023

BỘ GIÁO DỤC VÀ ĐÀO TẠO **VIỆN HÀN LÂM KHOA HỌC VÀ CÔNG NGHỆ VIỆT NAM**

HỌC VIỆN KHOA HỌC VÀ CÔNG NGHỆ

NGUYỄN THỊ THU HIỀN

NGHIÊN CỨU PHƯƠNG PHÁP CHUẨN HOÁ VĂN BẢN VÀ NHẬN DẠNG THỰC THỂ ĐỊNH DANH TRONG NHẬN DẠNG TIẾNG NÓI TIẾNG VIỆT

LUẬN ÁN TIẾN SĨ NGÀNH MÁY TÍNH

Chuyên ngành: Hệ thống thông tin

Mã số: 9 48 01 04

Xác nhận của Học viện	**Người hướng dẫn 1**	**Người hướng dẫn 2**
Khoa học và Công nghệ	*(Ký, ghi rõ họ tên)*	*(Ký, ghi rõ họ tên)*

HÀ NỘI - 2023

LỜI CAM ĐOAN

Tôi xin cam đoan đây là công trình nghiên cứu của riêng tôi. Các kết quả được viết chung với các tác giả khác đều được sự đồng ý của các đồng tác giả trước khi đưa vào luận án. Các kết quả nêu trong luận án là trung thực và chưa từng được công bố trong các công trình nào khác.

Tác giả

Nguyễn Thị Thu Hiền

LỜI CẢM ƠN

Luận án của tác giả được thực hiện tại Học viện Khoa học và Công nghệ - Viện Hàn lâm Khoa học và Công nghệ Việt Nam, dưới sự hướng dẫn tận tình của PGS.TS. Lương Chi Mai và TS. Nguyễn Thị Minh Huyền. Tôi xin được bày tỏ lòng biết ơn sâu sắc đến hai Cô về những định hướng nghiên cứu, sự động viên và hướng dẫn tận tình giúp tôi vượt qua những khó khăn để hoàn thành luận án này.

Tôi cũng xin gửi lời cảm ơn chân thành đến các nhà khoa học, các đồng tác giả của các công trình nghiên cứu đã được trích dẫn trong luận án. Đây là những tư liệu quý báu có liên quan giúp tôi hoàn thành luận án.

Tôi xin chân thành cảm ơn đến Ban lãnh đạo Học viện Khoa học và Công nghệ, Viện Công nghệ Thông tin đã tạo điều kiện thuận lợi cho tôi trong quá trình học tập, nghiên cứu.

Tôi xin chân thành cảm ơn Ban giám hiệu trường Đại học Sư phạm - ĐH Thái Nguyên, Khoa Toán, Bộ môn Khoa học máy tính - Hệ thống thông tin và các đồng nghiệp đã giúp đỡ và tạo điều kiện thuận lợi để tôi có thể thực hiện kế hoạch nghiên cứu, hoàn thành luận án.

Tôi xin được bày tỏ tình cảm và lòng biết ơn vô hạn tới những người thân trong Gia đình, những người luôn dành cho tôi sự động viên, khích lệ, sẻ chia, giúp đỡ trong những lúc khó khăn.

Tác giả

Nguyễn Thị Thu Hiền

MỤC LỤC

Trang

DANH MỤC TỪ VIẾT TẮT

STT	Từ viết tắt	Từ tiếng Anh	Ý nghĩa tiếng Việt
1	ASR	Automatic Speech Recognition	Nhận dạng tiếng nói tự động
2	BERT	Bidirectional Encoder Representations from Transformers	Mã hóa biểu diễn hai chiều dựa trên Transformers
3	BiLSTM	Bidirectional Long Short Term Memory	Mô hình bộ nhớ ngắn-dài hạn hai chiều
4	BPE	Byte-Pair-Encoding	Mã hoá cặp byte
5	CaPu	Recovering Capitalization and Punctuation model	Mô hình khôi phục dấu câu và chữ hoa
6	CBOW	Continuous Bag of Words	Mô hình nhúng từ "Túi từ liên tục"
7	CNN	Convolutional Neural Network	Mạng nơ-ron tích chập
8	CRF	Conditional Random Fields	Trường ngẫu nhiên có điều kiện
9	DL	Deep Learning	Học sâu
10	DNN	Deep Neural Networks	Mạng nơ-ron sâu
11	ELMO	Embeddings from Language Model	Nhúng từ từ mô hình ngôn ngữ
12	E2E	End-to-End	Mô hình đầu - cuối
13	GloVe	Global Véc-tơs for Word Representation	Mô hình nhúng từ dựa trên biểu diễn từ
14	GRU	Gated Recurrent Unit	Mạng hồi tiếp có cổng

15	GPT	Generative pre-trained transformer	Mô hình biến đổi được huấn luyện trước
16	HMM	Hidden Markov Model	Mô hình Markov ẩn
17	LM	Language Model	Mô hình ngôn ngữ
18	LSTM	Long Short Term Memory	Mô hình bộ nhớ ngắn-dài hạn
19	ME	Maximum Entropy	Mô hình Entropy cực đại
20	MEMM	Maximum Entropy Markov Model	Mô hình Markov Entropy cực đại
21	MTL	Multi-Task Learning	Học đa tác vụ
22	NER	Named Entity Recognition	Nhận dạng thực thể định danh
23	OOV	Out-of-Vocabulary	Từ nằm ngoài từ điển
24	RNN	Recurrent Neural Network	Mạng nơ-ron hồi quy
25	Seq2seq	Sequence-to-Sequence	Mô hình ánh xạ từ chuỗi sang chuỗi
26	SLU	Spoken Language Understanding	Hiểu ngôn ngữ nói
27	SVM	Support Véc-tơ Machine	Máy véc-tơ hỗ trợ
28	VLSP	Vietnamese Language and Speech Processing	Hội thảo xử lý ngôn ngữ và tiếng nói tiếng Việt
29	XLNNTN		Xử lý ngôn ngữ tự nhiên
30	TTS	Text To Speech	Hệ thống chuyển văn bản sang tiếng nói
31	WER	Word Error Rate	Tỉ lệ lỗi từ

DANH MỤC BẢNG BIỂU

DANH MỤC HÌNH VẼ