Improve the Transliteration System from Nôm Script into Vietnamese National Script using Language Models

Thesis Advisor
Associate Professor Dinh Dien

Thesis Reviewer
Doctor Nguyen Truong Son

Student
Nguyen Thi Kim Phuong

Outline

- Problem
- Objective
- Background
- Proposed Approach
- Data
- Experiments and Results
- Conclusion and Future Work

Problem: inaccessible property

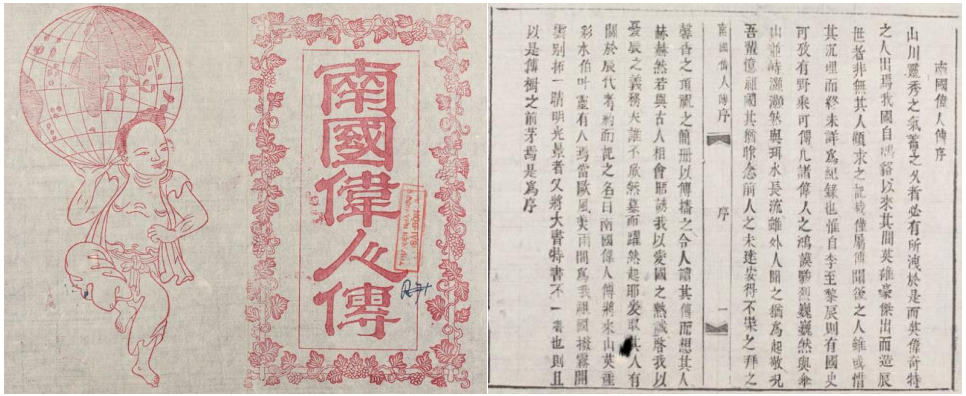


Figure 1. Biographies of prominent heroes in Vietnamese history
(Source: Vietnamese Nôm Preservation Foundation)

- Majority of Vietnamese cannot read Nôm
- Difficult, costly (time, effort) to learn
- Fewer than 100 people world-wide

Problem: barrier to research studies

- Ancient documents written in Nôm script
  - Historial, cultural, national values
  - Need to be understood and harnessed

- Nôm script 𡨸喃
  - Blocks access to ancient documents
  - Barrier to cultural and historical studies

Problem: current approach

Input

Inaccessible

傳翹

咱越

Transliteration

Nôm Converter

Output

Accessible

Truyện Kiều

Tiếng Việt

Problem: limitation of current approach

- Unable to transliterate popular Nôm text

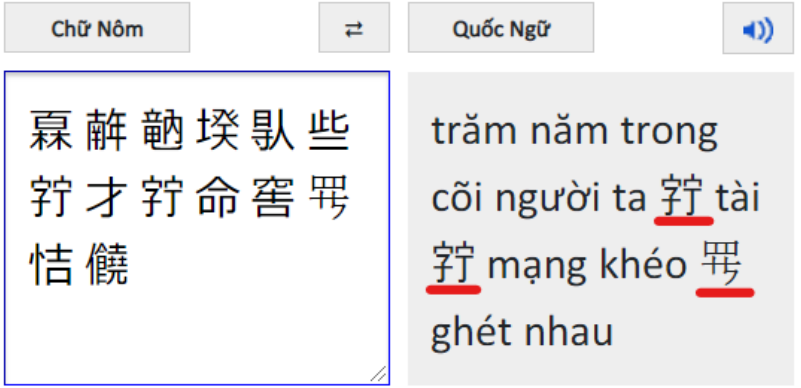


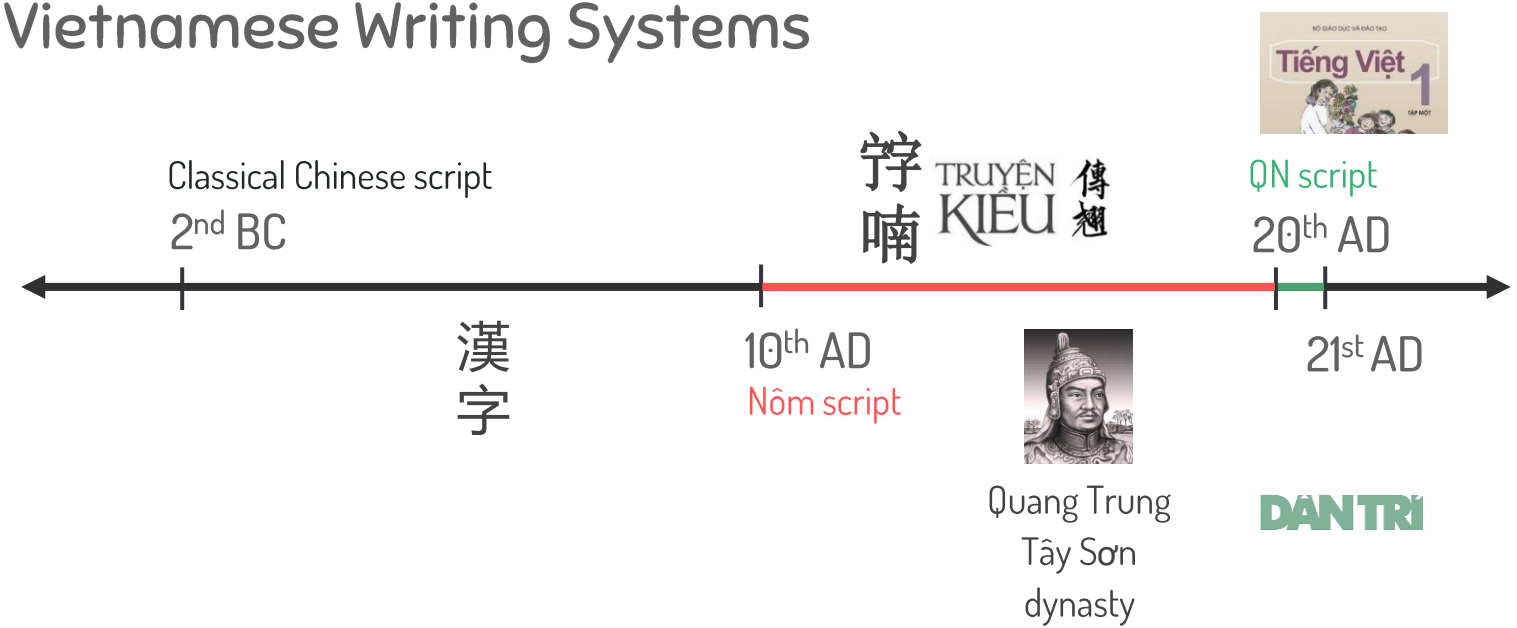
Figure 2. Two first sentences in *Tale of Kieu* cannot be transliterated by Nôm Converter

Objective

Input Inaccessible		Output Accessible
	Transliteration	
傳翹	Nôm Converter	Truyện Kiều
啫越	This Thesis	Tiếng Việt

Background

Vietnamese Writing Systems



Background

Nôm Script 字喃

- Based on Chinese script
- Record Vietnamese speech
- Chinese script alone is insufficient
- Long in history compared with QN script (National script)

Background

Transliteration

- Conversion from one script to another within **one language**
- Example:

Source script		Target script (Latin-based)
한글		hangul
碎艹馭越南		Tôi là người Việt Nam

Background

Statistical Machine Translation (SMT)

Input

Output

người việt

người vượt

ngài việt

...

Background

Statistical Machine Translation (SMT)

- Original problem: $\hat{q} = \underset{q}{argmax}\ P(q|n)$ (2.1)
- Reformulate with Bayes rule: $\hat{q} = \underset{q}{argmax}\ \frac{P(q)P(n|q)}{P(n)}$ (2.2)
- Final model: $\hat{q} = \underset{q}{argmax}\ P(q)P(n|q)$ (2.2)′

Background

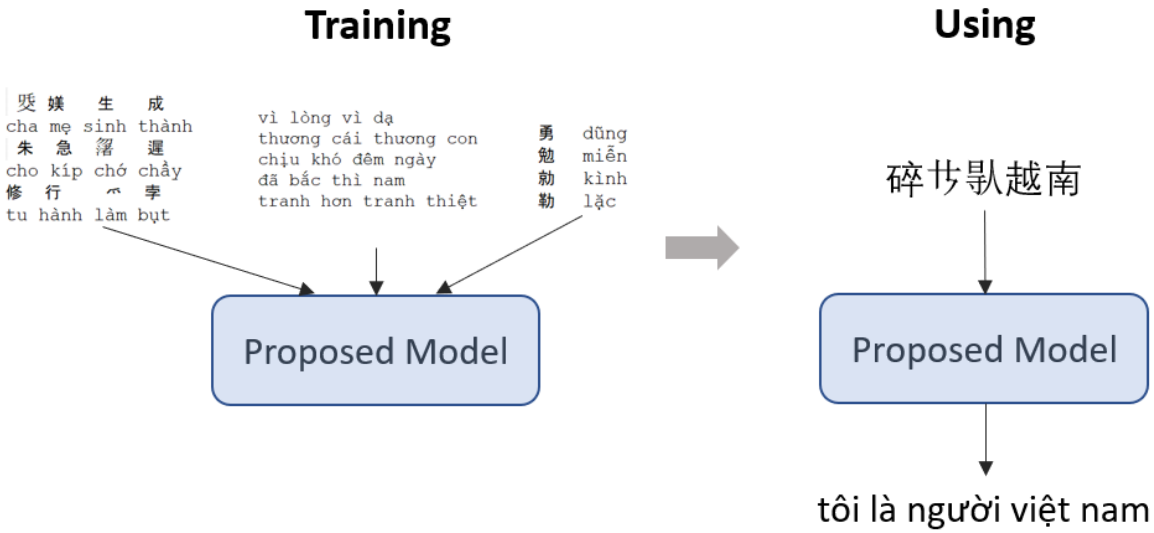
Statistical Machine Translation (SMT)

- Original problem: $\hat{q} = \underset{q}{argmax}\ P(q|n)$ (2.1)

- Final model: $\hat{q} = \underset{q}{argmax}\ P(q)P(n|q)$ $(2.2)'$

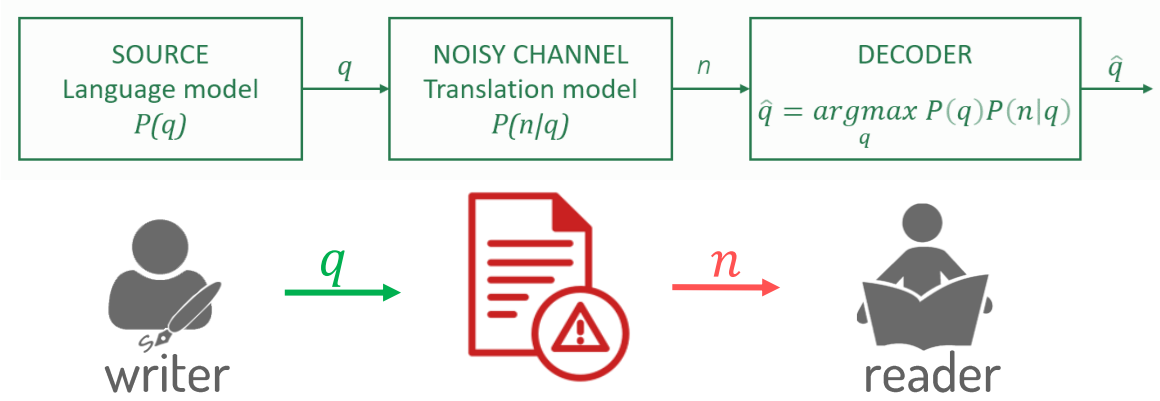
Language model Translation model

Proposed Approach: overview



Proposed Approach: details

- SMT: Statistical Machine Translation
- Moses[*]: a collection of tools in SMT for decoding
- Noisy-channel model:



[*] Moses: *https://github.com/moses-smt/mosesdecoder*

Proposed Approach: algorithms

- Decoding: search problem - heuristic Beam Search
- Word alignment: Expectation Maximization (EM)

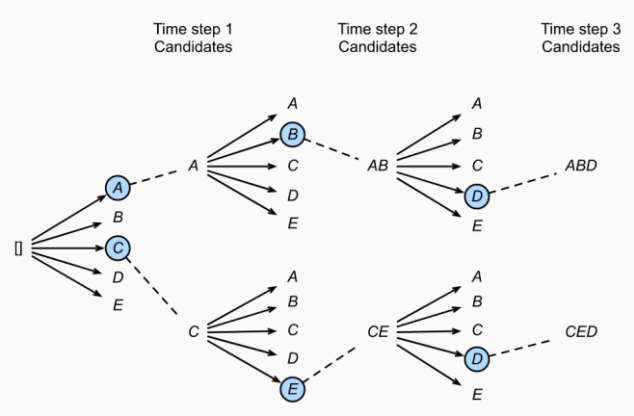
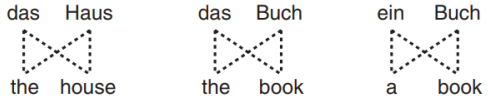


Figure 3. Beam Search
(Source: Dive into Deep Learning)



e	*f*	Initial	1st it.	2nd it.	3rd it.	...	Final
the	*das*	0.25	0.5	0.6364	0.7479	...	1
book	*das*	0.25	0.25	0.1818	0.1208	...	0
house	*das*	0.25	0.25	0.1818	0.1313	...	0
the	*buch*	0.25	0.25	0.1818	0.1208	...	0
book	*buch*	0.25	0.5	0.6364	0.7479	...	1
a	*buch*	0.25	0.25	0.1818	0.1313	...	0
book	*ein*	0.25	0.5	0.4286	0.3466	...	0
a	*ein*	0.25	0.5	0.5714	0.6534	...	1
the	*haus*	0.25	0.5	0.4286	0.3466	...	0
house	*haus*	0.25	0.5	0.5714	0.6534	...	1

Figure 4. Expectation Maximization
(Source: Philipp Koehn, SMT)

Data

- Two component models: Translation and Language models
- Two datasets: parallel and monolingual
- Sources:
  - The Internet
  - Manually typed from printed books

Data

- Parallel:

Type	Domain	Size (entries)
Mono-syllabic dictionary	General	38,897
Poly-syllabic dictionary	General	6,205
Sentence pairs	Literature	11,636
	Religion	1,056

- Monolingual:

Type	Domain	Size (sentences)
Sentence	Literature	39,675
Sentence	Religion	84,381

Experiments

- Data splitting
- Parallel data: 8-1-1 for train-tune-test
- QN script from test set: measure perplexity of LM
- Monolingual data: train 2 domain-specific LMs
- Evaluation metrics: perplexity and BLEU score
- Verify 3 research questions

Results: research question 1

Impact of LM: verified with a note

ID	Training Data		BLEU Score
	Parallel	LM	
1	38897 entries of mono-syllabic dictionary	No	14.56
2	38897 entries of mono-syllabic dictionary	Yes	65.94
Auxiliary	• 38897 entries mono-syllabic dictionary • 6205 entries poly-syllabic dictionary • 6348 sentence-pairs	Yes	85.38

Results: research question 2

Performance effective: verified; Cost-effective: not verified

ID	Domain of Test Set	Domain of LM	Perplexity	BLEU Score
3	Literature	Literature	226.3	82.80
4		History	948.1	81.56
5		Religion	1290.8	79.54
6	Religion	Literature	1095.3	85.90
7		History	1006.5	86.91
8		Religion	341.0	89.72

Results: research question 3

Improves Nôm Converter: verified

Domain	Nôm Converter	Proposed System	
		LM: No	LM: Yes
Literature	56.84	79.85	82.80
Religion	50.95	87.11	89.72

Conclusion

Contributions:

- Studies Vietnamese writing systems in relation to MT
- Prepares data for experiments: parallel & monolingual
- Builds an improved MT system compared with Nôm Converter

Hypotheses:

- Impact of LM: verified with a note
- Performance effective: verified; Cost-effective: not verified
- Improves Nôm Converter: verified

Future Work

- Name entity recognition (not output upper-case for proper names)
- Linguistic knowledge integration (not integrate any)

References

- https://icon-library.com/icon/read-book-icon-24.html
- https://thenounproject.com/icon/document-error-729356/
- https://github.com/FortAwesome/Font-Awesome/issues/2543
- https://lib.nomfoundation.org/collection/1/volume/301/
- https://www.iconfinder.com/icons/5938213/finance_proposition_unique_value_icon
- https://thenounproject.com/icon/speech-recognition-3022008/

Publications

[1] Dien Dinh, **Phuong Nguyen** and Long H. B. Nguyen, "Transliterating Nôm Scripts into Vietnamese National Scripts using Statistical Machine Translation", International Journal of Advanced Computer Science and Applications(IJACSA), 12(2), 2021. http://dx.doi.org/10.14569/IJACSA.2021.0120205 (ESCI-indexed journal).

[2] Dinh Dien, Trang Minh Chien, **Nguyen Thi Kim Phuong**, and Nguyen Hong Buu Long, "Automatic Nôm-script Transliteration using Neural Machine Translation", in Proceedings Of The 13Th National Conference On Fundamental And Applied Information Technology Research, December, 2020, Nha Trang, Vietnam, doi: 10.15625/vap.2020.00149.

[3] **Nguyen Thi Kim Phuong**, Nguyen Hong Buu Long, Dinh Dien, and Luong An Vinh, "Improving Transliteration System from Nôm Scripts into Vietnamese Scripts using Language Model", in Proceedings Of The 14Th National Conference On Fundamental And Applied Information Technology Research, December, 2021, Ho Chi Minh City, Vietnam, doi: 10.15625/vap.2021.0092.

[4] Dinh Dien, **Nguyen Thi Kim Phuong**, Diep Gia Han, and Tran Nguyen Son Thanh, "Chuyển tự tự động từ chữ Nôm sang chữ Quốc ngữ", Hội thảo 100 năm chữ Quốc ngữ, December, 2019.


thank
you