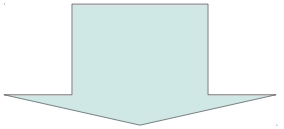
A Pointwise Approach for Vietnamese Automatic Diacritics Restoration

Tuan Anh Luu
Kazuhide Yamamoto（山本和英）
Nagaoka University of Technology, JAPAN

Note that I don't speak Vietnamese.

Diacritics Restoration?

xu ly ngon ngu tu nhien



xử lý ngôn ngữ tự nhiên

No work reported so far for Vietnamese!

Vietnamese diacritics

original	dropped
a, à, ả, ã, á, ạ, ă, ằ, ẳ, ẵ, ắ, ặ, â, ầ, ẩ, ẫ, ấ, ậ	a
e, è, ẻ, ẽ, é, ẹ, ê, ề, ể, ễ, ế, ệ	e
i, ì, ỉ, ĩ, í, ị	i
o, ò, ỏ, õ, ó, ọ, ơ, ờ, ở, ỡ, ớ, ợ, ô, ồ, ổ, ỗ, ố, ộ	o
u, ù, ủ, ũ, ú, ụ, ư, ừ, ử, ữ, ứ, ự	u
y, ỳ, ỷ, ỹ, ý, ỵ	y
đ, d	d

There are 67 (out of 89) characters that contain diacritics.

Important?

- Yes.
- >30 languages have diacritics.
  - French, Romanian, Croatian, Sindhi, Vietnamese, ...
- Many texts are missing diacritics (on the Web).

Difficult?

- Yes, as for Vietnamese.
- So many diacritics; 95% words in Vietnamese, whereas 15% in French and 35% in Romanian.
- So ambiguous; 80% of missing diacritics are ambiguous in Vietnamese, whereas 50% in French and 25% in Romanian.

Two standard approaches:

Word-based

- Language-dependent
- Large lexical resources, language models, additional processing tasks required.
- Accuracy high

Character-based

- Language-independent
- Statistical information on n-grams
- Easy to implement, very fast

They can't be applied for Vietnamese.

- Word-based
  - NO word segmenter w/o diacritics
  - NOT enough text corpus and dictionary
- Character-based
  - Diacritics used more extensively than other languages

Proposal: Pointwise Approach

- assumes that the restoration is done independently.
- uses machine learning (SVM)
- given context of surrounding information of the target word (missing diacritics)

Diacritics depend on context

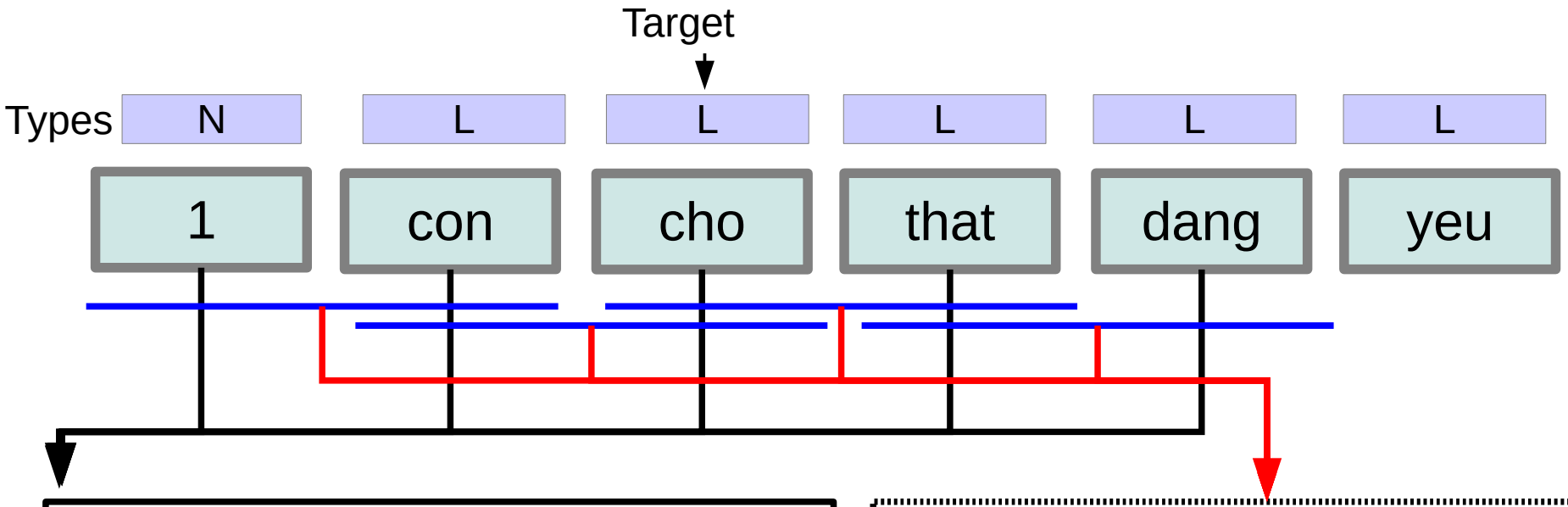
cho mot muc tieu	cho (to give)
On nhu cai cho	chợ (market)
Hay cho den dung thoi diem	chờ (to wait)
1 con cho ngoi ngoai cong	chó (dog)

- Missing diacritics depend on the context.
- Thus, if the context is given, missing diacritics can be restored independently.

Features for machine learning

- Window: W words around the target as context.
  - W = 2 and 3 tried this time.
- syllable n-gram
  - 1-gram & 2-gram
- syllable type n-gram
  - either of uppercase (U), lowercase (L), number (N) or other (O)

Feature (1): syllable n-gram & syllable type n-gram



1-gram :

“1”, “con”, “cho”, “that”, “dang”,

Type 1-gram :

“N”, “L”

2-gram :

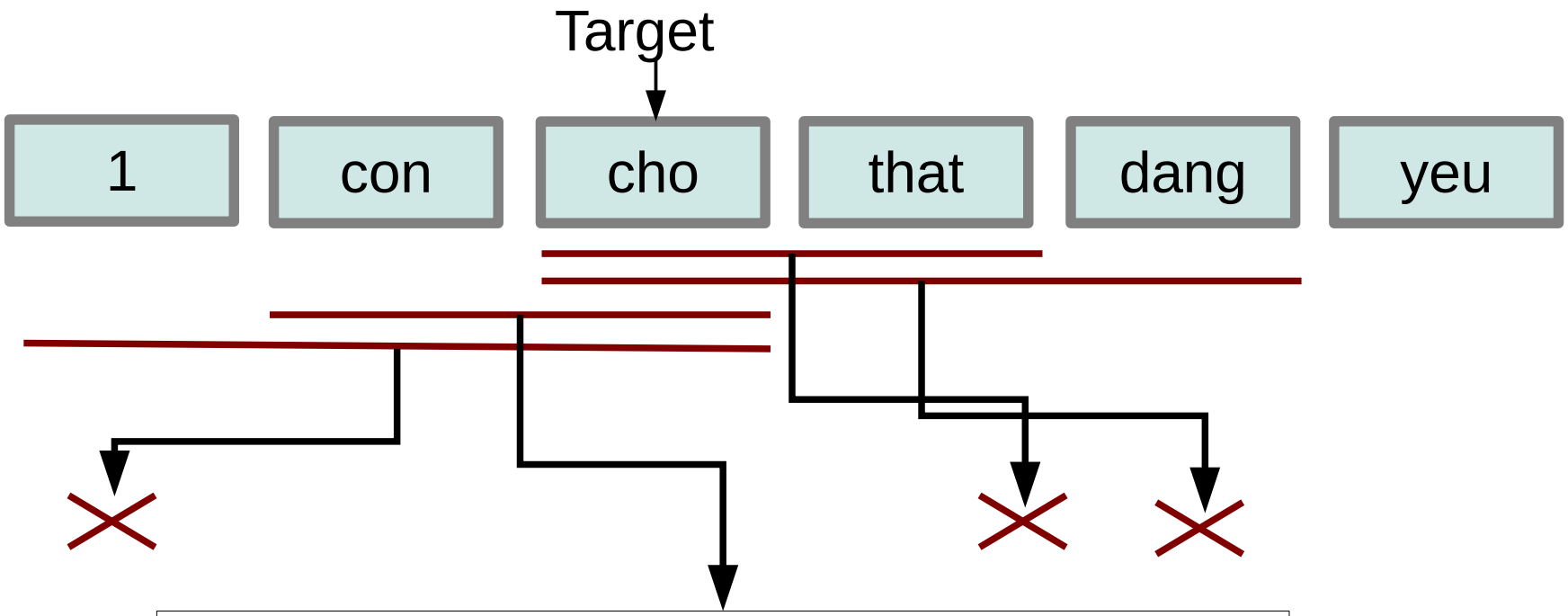
“1 con”, “con cho”, “cho that”, “that dang”

Type 2-gram :

“NL”, “LL”

Feature(2): dictionary word

- Dictionary words that contain the given syllable.



Dictionary word features :
“con cho” (dog)

Experimental setting

- Uppercase words (= proper nouns) are out of target for restoration.
- Linear SVM, LIBLINEAR software package
- Text: journalism Web pages, crawled
  - difficult due to many unknown words and errors.
- 320 Mbytes for training, different 15 Mbytes for test.
- A classifier build for each non-diacritical strings (1525 strings).

Result

- 94.7% accuracy attained

  when W=3 and max training
- Outperforms baselines:
  - 15.9% for random selection
  - 71.8% for most-frequent approach

Result: different window size and training data size

Conclusion

- Method of Vietnamese diacritics restoration proposed.
  - pros: simple, language-independent
  - cons: computationally expensive
- 94.7% of Vietnamese diacritics correctly restored.
  - First attempt for Vietnamese

Cảm ơn sự quan tâm của các bạn!