AN INCREMENTAL LEARNING BASED SPELL-CHECKER FOR LOCAL SEARCH USER QUERIES

HWAI-JIN PENG, PINZHU QIAN, TYAN TRINH, CHENG-YEN YANG

TELENAV®

20

INTRODUCTION

- According to Telenav, a company that provides personalized mobile navigation and location-based platform services, roughly 15% of web and local search queries have misspellings.
- Misspellings are classified as orthographical (the user guesses how to spell the word), typographical (the user types an adjacent letter on their keyboard of the desired letter), or transposition errors (the user mistakenly swaps two letters in the word).
- Telenav provided our team with a goal to develop a local search user query spell-checking API with minimal dependencies that can be packaged as one library or exposed as one restful service.

OBJECTIVES

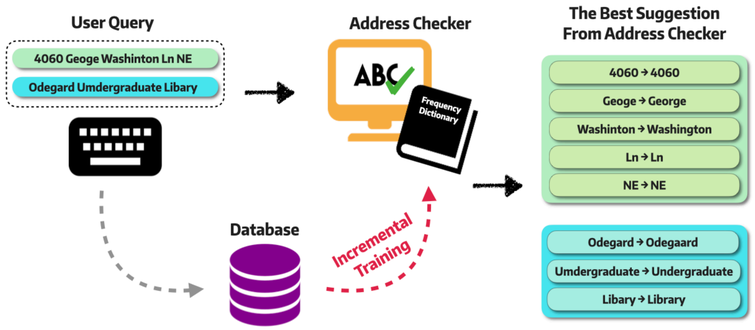
- Develop a spell-checking API that:
  - Includes semantic and contextual inference
  - Leverages existing open source code
  - Applies existing spell-checking research techniques
  - Performs under 200ms with an accuracy above 80%

DATASETS

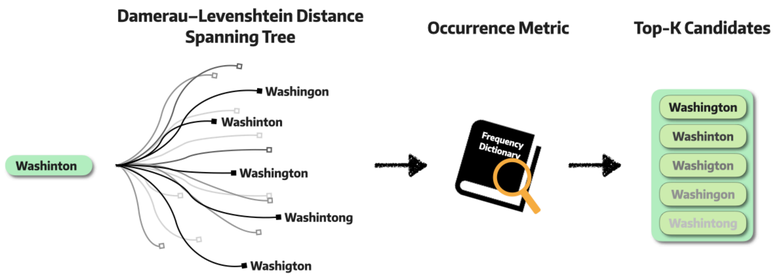
- **TELENAV USER QUERIES DATA**
  - 2M+ real user queries collected by Telenav (10M+ tokens)
  - Contains orthographic, typographic, and transposition type of misspelled words and noisy queries
  - Sample data:
    - *UPMC Herberman conference center*
    - *XXX Washington Rd., Augusta GA*
    - *breakfast restaurant nearme*
- **OPENADDRESS DATA**
  - 100M+ real address entries including street name and house number from across 52 states (300M+ tokens)
    - *XXX 11TH AVE NE*
    - *XXX ROOSEVELT WAY*
- **OTHER CORPUS**
  - OpenSubtitles: 3.2G+ tokens from English TV and movies subtitles
  - Wikitext: 100M+ tokens from English articles on Wikipedia

MODEL & APPROACH

- **SYSTEM CONFIGURATION**



- **ADDRESSCHECKER PYTHON PACKAGE**



- **TOP K CANDIDATES**
  - Spell-checking function can return top K candidates with its corresponding prediction scores for further application and usage.
- **INCREMENTAL LEARNING**
  - Allows model to incrementally adept new vocabulary from the new data without modifying original data.
- **WORD-LEVEL THRESHOLDING**
  - Helps eliminate words with infrequent appearances and reduce model's size.
- **CHAR-LEVEL PRUNING**
  - Helps eliminate invalid characters from the data and reduce time of the Damerau-Levenshtein distance candidates calculations.

RESULTS

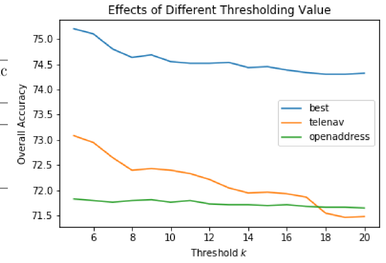
- **EVALUATION DATASET**
  - A dataset designed to evaluate the performance of spell-checking ability by simulating orthographical, typographical and transposition errors on 1000 different user queries.

INCREMENTAL LEARNING PERFORMANCE

Model	Vocabulary Size	Transition Accuracy	Typographic Accuracy	Orthographic Accuracy	Inference Time(ms)
Baseline	118292	59.95%	57.70%	35.55%	180
Ours + Telenav	182884	83.45%	81.85%	53.90%	55
Ours + Openaddress	278307	82.95%	81.30%	53.30%	70
Ours + ALL	158751	84.40%	84.25%	56.95%	40

- **WORD-LEVEL THRESHOLDING PERFORMANCE**

Model (Threshold)	Transition Accuracy	Typographic Accuracy	Orthographic Accuracy
Baseline (k=5)	59.95%	57.70%	35.55%
Ours + ALL (k=20)	83.00%	82.90%	57.05%
Ours + ALL (k=10)	83.50%	83.25%	56.85%
Ours + ALL (k=5)	84.40%	84.25%	56.95%



FUTURE WORK

- Support multi-lingual spell-checking in a single ensembled model.
- Introduce context-level information to eliminate ungrammatical structured candidates and incremental training data.
- Choose the candidates that fits the context best for text completion using pre-trained language model.

CONCLUSION & CONTRIBUTION

- Implemented a python address spell checker package for English spell-checking and correction.
- Designed an incremental learning-based technique to provide update to the exist model without further data preprocessing.
- Collected and created an evaluation dataset consists of three types of common misspellings.

ELECTRICAL & COMPUTER ENGINEERING
UNIVERSITY of WASHINGTON

ADVISOR: PAYMAN ARABSHAHI
INDUSTRY MENTOR: KUMAR MADDALI, CHANGZHENG JIANG
SPONSOR: TELENAV, INC.