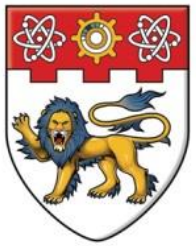


NANYANG TECHNOLOGICAL UNIVERSITY

Unsupervised Software-Specific Morphological Forms Inference from Informal Discussions

Chunyang Chen*, Zhenchang Xing+, Ximing Wang*

***Nanyang Technological University (Singapore) ,+Australian National University (Australia)**

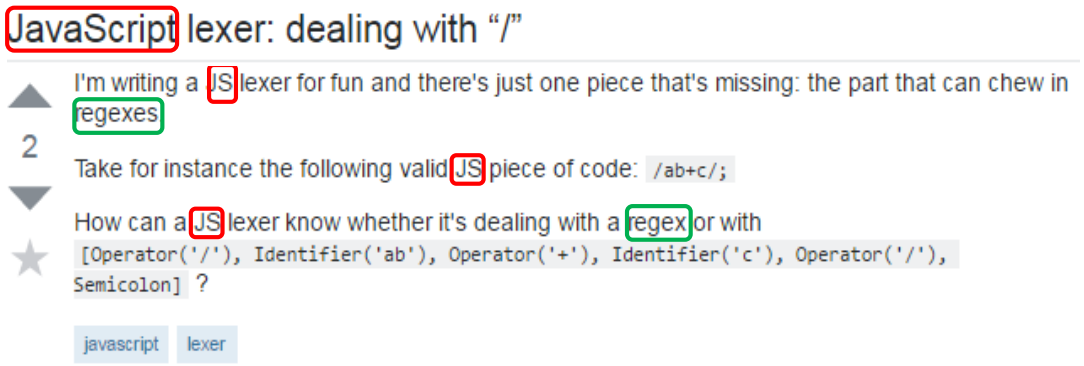
Background

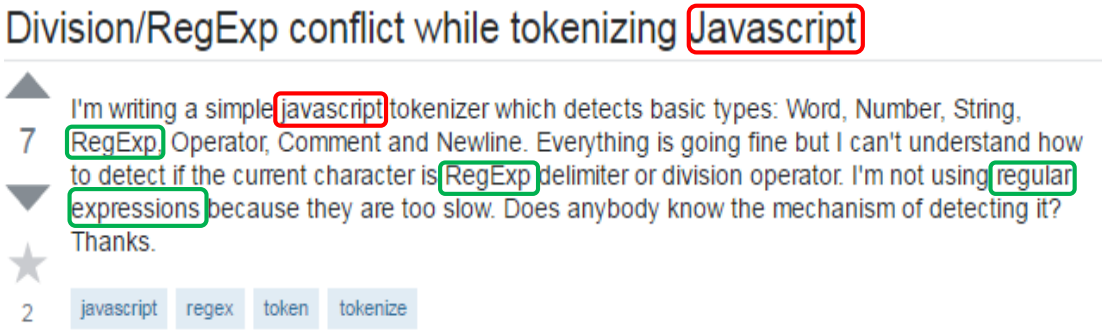
Informal discussions on social platforms are accumulated into a large body of programming knowledge in natural language text.

Background

The "**beauty**" of natural language is its dynamic:

- E.g., the same concept is often intentionally or accidentally mentioned in many different morphological forms in informal discussions.





Background

Morphological forms of one word:

- Abbreviations
- Synonyms
- Misspellings

TABLE I: Morphological forms of *visual c++*

Term	Frequency	Annotation
visual c++	10,294	Standard
msvc	8,477	abbreviation
vc++	7,154	abbreviation
microsoft visual c++	1,826	synonym
ms vc++	295	abbreviation
visual-c++	110	synonym

Motivation

The "beauty" can also be a nightmare for machine!

Problems brought by those morphological forms:

- Lexical gap in information retrieval
- Word sparsity in data analysis
- Inconsistent vocabulary for NLP related tasks

Motivation

Natural Language Processing:

- It groups English words into sets of synonyms called synsets.
- Problems:
  - big human efforts
  - The database is fixed, easy to be out of date.
  - few software-specific terms

Software-specific domain:

Domain-specific **Thesaurus**

- An (semi)automatic method without much manual efforts.
- Easy to update
- Consider domain-specific information

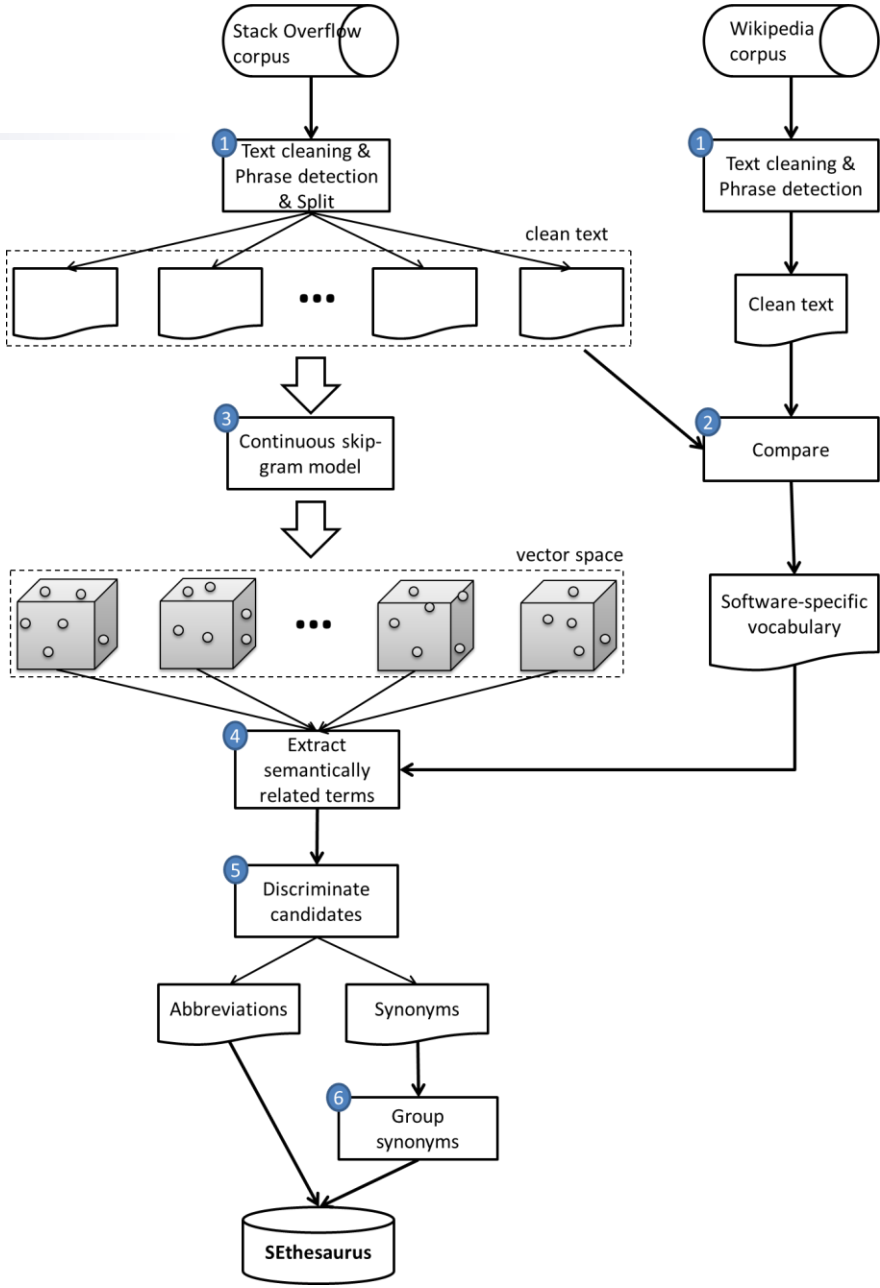
Challenge

- To spot morphological word forms, traditional methods heavily rely on the lexical similarity of words.

- However, they may misclassify (**opencv**, **opencsv**) as synonyms, while (**ie**, **view**) as abbreviations.

Overall approach

- Incorporate both semantic and lexical information;
- Large-scale unsupervised approach.



1. Preprocessing

- Dataset
  - Stack Overflow: 10M questions & 16.5M answers
  - Wikipedia: 5M articles
- Text cleaning
  - Remove HTML tags, lowercase and tokenize words
- Phrase Detection
  - E.g., visual studio, sql server, quick sort
  - Find bigram phrases that appear frequently enough in the text compared with the frequency of each unigram. Repeat that process to find longer phrases.

2. Building Software-Specific Vocabulary

- Dataset:
  - Stack Overflow: software-specific
  - Wikipedia: general (almost including all-domain knowledge)

- Identify software-specific terms by contrasting the term frequency of a term in the software specific corpus compared with its frequency in the general corpus:

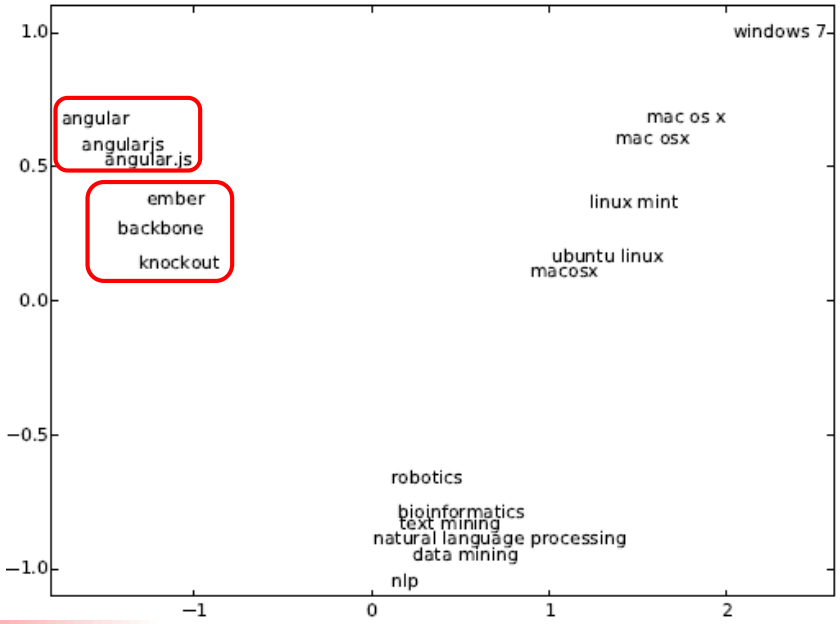
$$\text{domainSpecificity(t)} = \frac{p_d(t)}{p_g(t)} = \frac{\frac{c_d(t)}{N_d}}{\frac{c_g(t)}{N_g}}$$

$p_x(t)$is the probability of the term t in corpus x and $c_x(t)$ is the count of t in corpus x.

3 & 4. Extracting Semantically Related Terms

- Split the whole Stack Overflow into 11 small bulks;
- Train one word2vec model on one bulk;
- For each domain-specific term, get its top 20 semantic related words in each model;
- Merge and rerank candidates from different bucks into one list.

- Candidates:
  - Synonyms & abbreviations
  - Similar terms

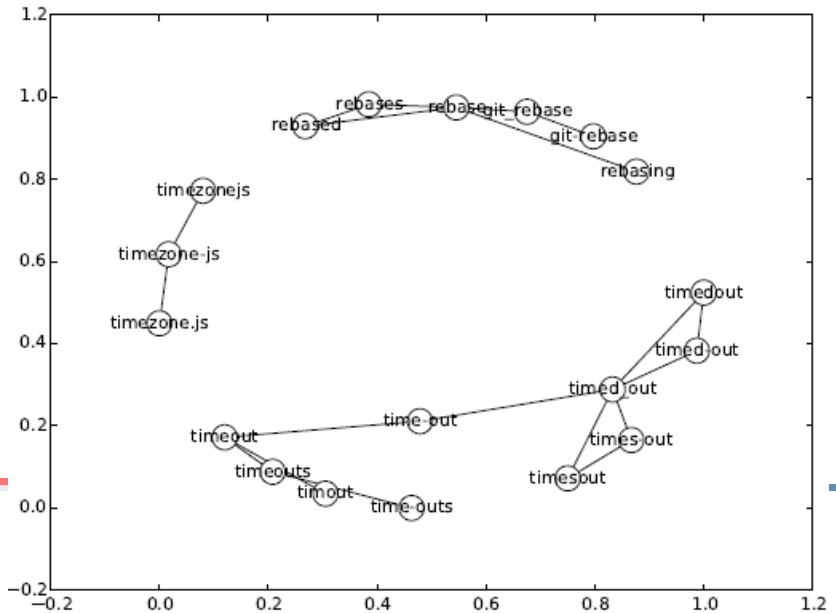


5. Discriminating Synonyms & Abbreviations

- Discriminating Morphological Synonyms
  - Damerau-Levenshtein distance
  - $\text{similarity}_{\text{morph}}(\text{t}, \text{w}) = 1 - \frac{DLdistance(t,w)}{max(len(t); len(w))}$
- Discriminating Abbreviations
  - The characters of the abbreviation must be in the same order as they appear in the term;
  - The length of the abbreviation must be shorter than that of the term;
  - If there are digits in the abbreviation, there must be the same digits in the term;
  - …

6. Grouping Morphological Synonyms

- Existing synonyms are separated and overlapped.
  - timeout: timeouts, timout, time out;
  - timed out: timed-out, times out, time out
- Build a graph of morphological synonyms
  - All existing pairs of synonyms are regarded edges for the graph
- Take all terms in a connected component as mutual synonyms



SEthesaurus

- 52,645 software-specific terms,
- 4,773 abbreviations for 4,234 terms,
- 14,006 synonym groups containing 38,104 morphological terms.

Evaluation

- The coverage of software-specific vocabulary
- Abbreviation coverage
- Synonym coverage
- Human evaluation of the accuracy

The Coverage of Software-Specific Vocabulary

- Ground truth
  - A tag (in Stack Overflow and Code Project) is a word or phrase that describes the topic of the question.
  - All tags are software-specific terms.

- Results
  - Our thesaurus contains
    - **70.1**% tags in Stack Overflow
    - **79.2**% tags in Code Project

Abbreviation & Synonym Coverage

- Abbreviation coverage
  - Ground truth: 1,292 abbreviations of computing and IT in Wikipedia
  - Result: **86%** of them are covered in our thesaurus.

- Synonym coverage
  - Ground truth: 3,231 synonym pairs of tags in Stack Overflow are community created and approved.
  - Result:

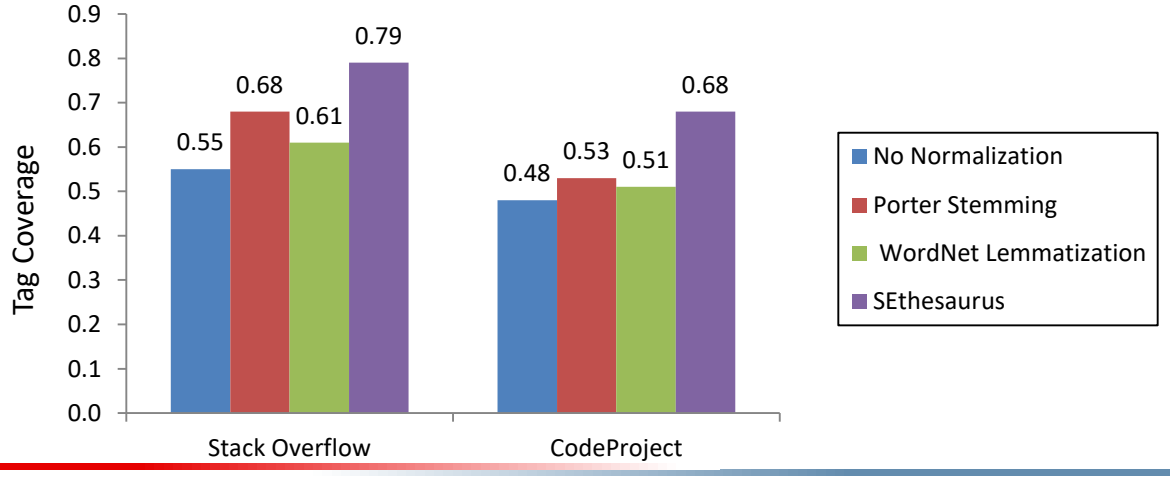
Method	#CoveredSynonym	#CoveredMaster
SEthesaurus	2,316	1,439
WordNet	725	218
SEWordSim	941	86

Human Evaluation of Accuracy

- Experiment
  - 3 final-year undergraduate and 1 RA with master degree
  - Randomly sample 400 synonym pairs and 200 abbreviation pairs for evaluation
- Result
  - **74.3**% abbreviation pairs are correct
  - **85.8**% synonym pairs are correct

Usefulness Evaluation

- Experiment
  - Normalize software-specific questions and corresponding tags with our thesaurus.
  - Investigate how much the text normalization can make question content more consistent with its metadata (i.e., tags).
  - Randomly sample 100K questions from Stack Overflow and 50K questions from CodeProject
- Result



Tool

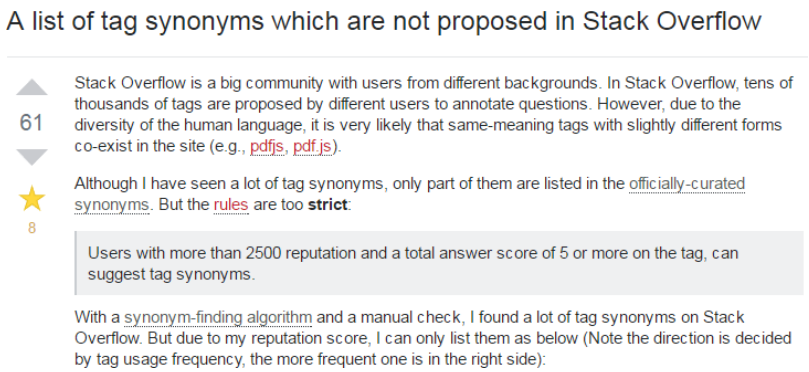
- Website
  - https://se-thesaurus.appspot.com/
- API
  - https://se-thesaurus.appspot.com/api

Ongoing Application

- Spell checking
  - General spell-checker is not suitable for software-specific text

Term	Misspellings
ubuntu	ubunutu
jquery	jqeury, jquey
eclipse	eclispe, eclise, eclips, eclipe
android	anroid, andoid, andriod, adroid, andorid
bootstrap	bootstarp, bootstap, boostrap, bootsrap
postgresql	postgressql, postresql, posgresql, postgesql

- Find tag synonyms
  - Propose 917 tag synonym pairs in Stack Overflow.
  - Get 61 upvotes and 8 favorites in two days.
  - https://meta.stackoverflow.com/questions/342097


A list of tag synonyms which are not proposed in Stack Overflow

61

8

Stack Overflow is a big community with users from different backgrounds. In Stack Overflow, tens of thousands of tags are proposed by different users to annotate questions. However, due to the diversity of the human language, it is very likely that same-meaning tags with slightly different forms co-exist in the site (e.g., pdfjs, pdf.js).

Although I have seen a lot of tag synonyms, only part of them are listed in the officially-curated synonyms. But the rules are too **strict**:

> Users with more than 2500 reputation and a total answer score of 5 or more on the tag, can suggest tag synonyms.

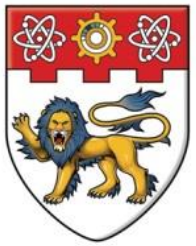
With a synonym-finding algorithm and a manual check, I found a lot of tag synonyms on Stack Overflow. But due to my reputation score, I can only list them as below (Note the direction is decided by tag usage frequency, the more frequent one is in the right side):

Ongoing Application

- IR & text preprocessing
  - Manually check the accurate synonyms & abbreviation, more than 3K groups so far.

    https://se-thesaurus.appspot.com/synonymAbbreviation_manualCheck.txt

  - Used to normalize software-specific text



NANYANG TECHNOLOGICAL UNIVERSITY

- Chen, Chunyang, Zhenchang Xing, and Ximing Wang. "Unsupervised software-specific morphological forms inference from informal discussions." In *Proceedings of the 39th International Conference on Software Engineering*, pp. 450-461. IEEE Press, 2017.
- Chen, Xiang, Chunyang Chen, Dun Zhang, and Zhenchang Xing. "SEthesaurus: WordNet in Software Engineering." *IEEE Transactions on Software Engineering* (2019).

Thanks for listening, questions?

Chunyang Chen*, Zhenchang Xing[+], Ximing Wang*

***Nanyang Technological University (Singapore) ,[+]Australian National University (Australia)**