Lexical Normalization of Spanish Tweets

Jhon Adrián Cerón-Guzmán Elizabeth León-Guzmán

Departamento de Ingeniería de Sistemas e Industrial
Universidad Nacional de Colombia, Bogotá D.C., Colombia
{jacerong,eleonguz}@unal.edu.co

April, 2016

Introduction

- Social media platforms have led to deep changes in the paradigm of information generation and consumption.
- Twitter is nowadays a popular microblogging site where users receive and exchange information instantaneously.
- 'Tweeting', therefore, has become an activity *par excellence* to say what one thinks or feels.

Briefly

What people say on Twitter has turned into a rich source of information to understand social behavior.

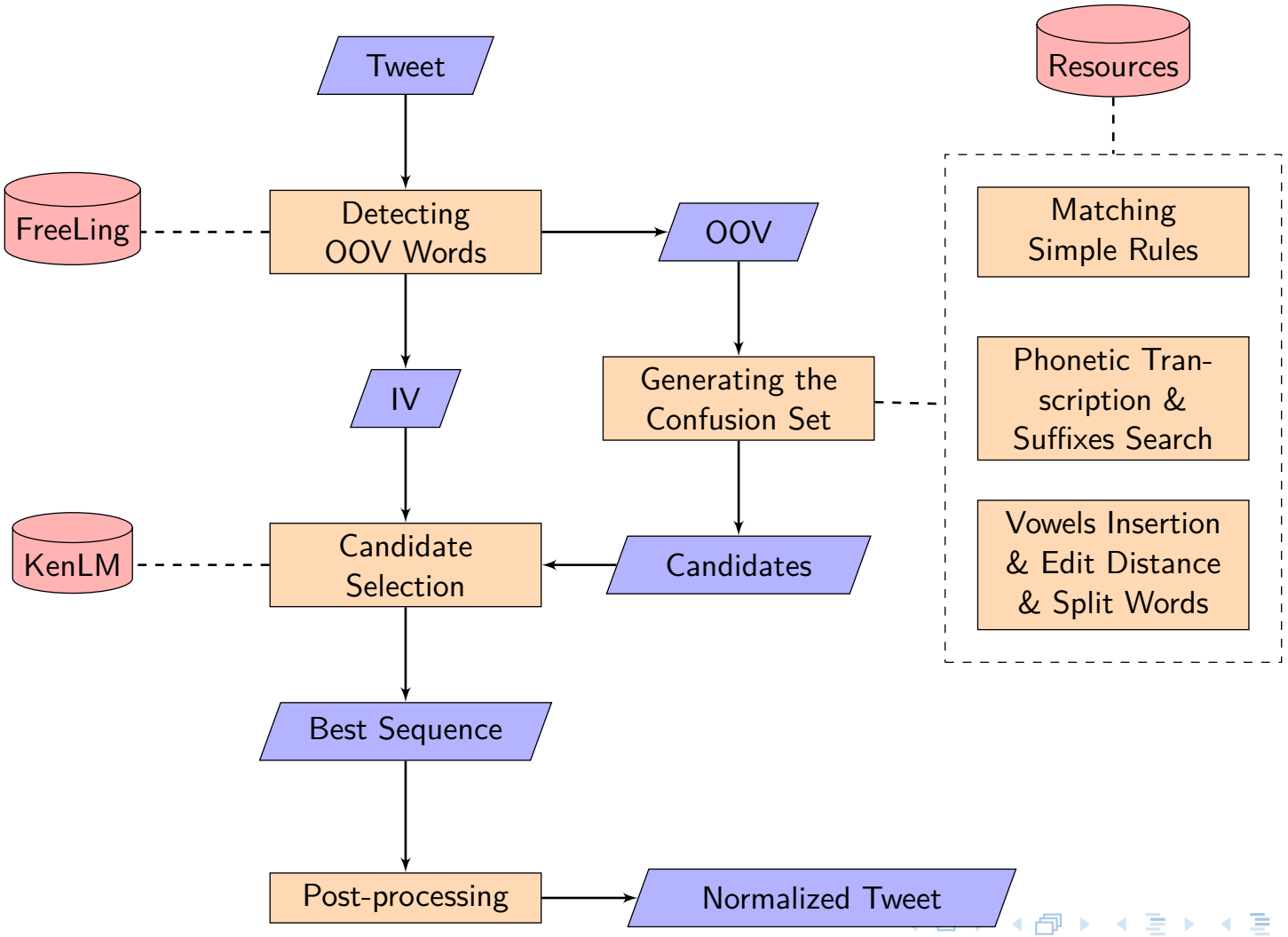
The Problem

- However, in the same way that Twitter data constitute an useful source, a vast amount of noise can be found in them.
- Analyzing this large amount of user-generated content presents several challenges, including:
  1. To distinguish noisy, useless, and irrelevant information from valuable data.
  2. To develop text analysis approaches based on NLP techniques, which properly adapt to the informal genre and the free writing style of Twitter.

Scope of the Work

- To deal with several lexical variation phenomena that occur on the content generation in order to improve the quality of natural language analysis.
- A lexical normalization approach of Spanish tweets, which normalizes non-standard word forms (i.e., out-of-vocabulary words) to their standard lexical forms, is proposed.
- This work proposes a one-to-many normalization approach to also deal with word segmentation problems such as lack of spacing between words.

The System Architecture



Detecting OOV Words

Input: A tweet written in Spanish

1. Tokenize the tweet text using the FreeLing tool.
2. Each resulting token is passed through a set of basic modules to identify standard word forms and other valid constructions.
3. If a token is not recognized by any of the modules, it is marked as OOV.

Output: A list of OOV tokens and a list of (token, POS tag) pairs

False Positives of Named Entities

(Real) Tweet

"Lo mejor es que me da igual todo SOI FELIZ."

`(The best is that I do not care anything, I AM HAPPY)`

The tokens "SOI FELIZ" are wrongly recognized as an entity:

- "SOI" is a typo of the standard word form "soy" (`I am`).
- "Feliz" (`happy`) is a standard word form.

Treatment of Named Entities

Input: A filtered list of (token, POS tag) pairs whose POS tag is *named entity*

1. Tokenize each named entity by white-space.
2. Look up each token in the dictionary of standard word forms.
3. If there is not an entry matching the token, it is marked as OOV.

Output: An expanded list of OOV words and a reduced list of (token, POS tag) pairs

Lexical Variation Phenomena

- Character repetition: "*claseeeeesss*" → *clases* (`classes`)
- Alteration of valid onomatopoeia: "*ajajajjaja*" → *ja* (`ha`)
- Language-dependent orthographic errors:
  - Missing of diacritical marks: "*tendre*" → *tendré* (`I will have`)
  - Uppercase/lowercase confussion: "*francia*" → *Francia* (`France`)
  - Letter confusion: *v* → *b*, *ll* → *y*, *h* → *0*
- Initialisms: "*xk*" → *porque* (`because`)

Lexical Variation Phenomena

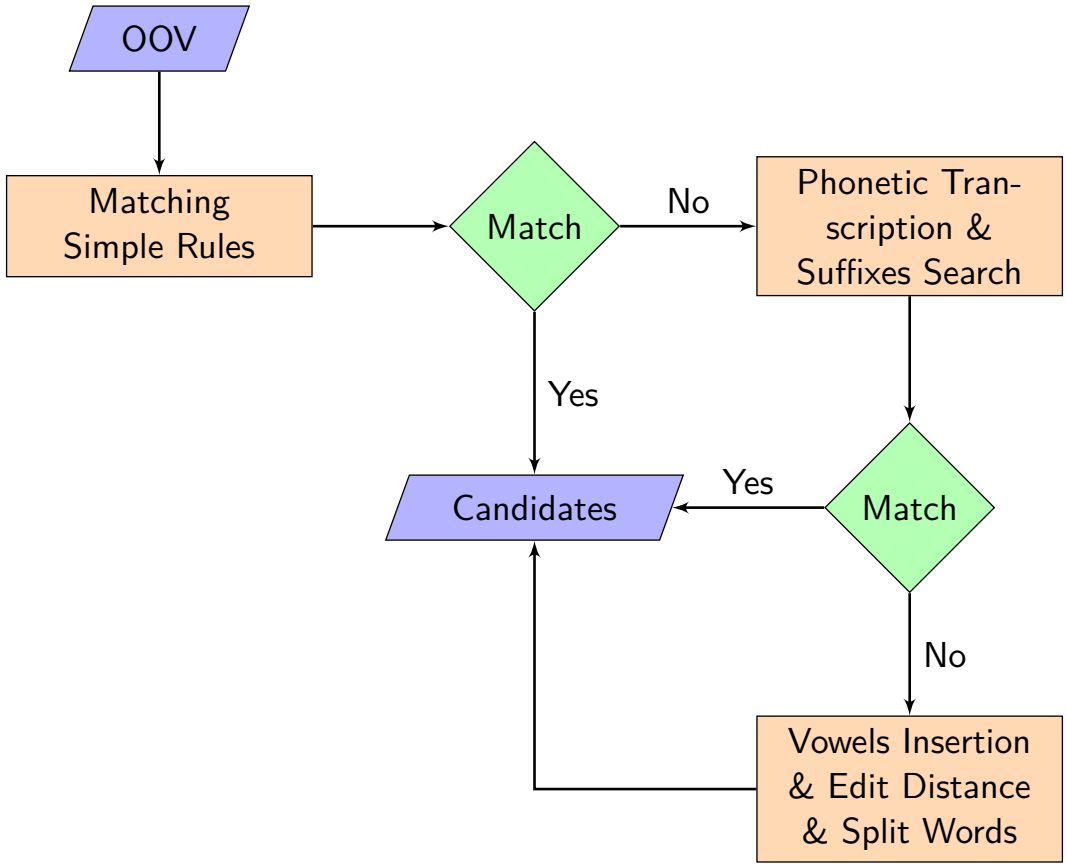
- Shortenings: "*pa*" → *para* (`for`)
- Homophonic confusion: "*pokitin*" → *poquitín* (`little bit`)
- Standard non-correct endings: "*mercao*" → *mercado* (`market`)
- Word segmentation problems: "*alomejor*" → *a lo mejor* (`at best`)

Correct OOV Words and Ill-formed OOV Words

To determine if an OOV token is a correct word:

1. The OOV token is included in its confusion set.
2. If the OOV token is which best fits the language model, it is considered as correct.

Generating the Confusion Set



Matching Simple Rules

- Simple regular expressions.
- A normalization dictionary.

Phenomena
Alteration of valid onomatopoeia
Missing of diacritical marks
Initialisms
Shortenings

Phonetic Transcription & Suffixes Search

- Normalization candidates are suggested by their phonetic similarity to the OOV word.
- A suffixes search is performed to recognize inflected word forms:
  - Enclitic pronouns.
  - Adverbs ending in *-mente*.
  - Diminutive forms.

Phenomena
Homophonic confusion
Standard non-correct endings
Uppercase/lowercase confusion
Letter confusion

Vowels Insertion & Edit Distance & Split Words

- Vowels insertion.
- Language $L(\_L)^+$, where L is the language of standard word forms.
- Levenshtein distance of one.
- The Longest Common Subsequence filters out candidates whose ratio is below a threshold.

Phenomena
Letter omission
Word segmentation problems

Candidate Selection

Input: OOV words and their confusion set

1. Candidate combinations are evaluated against the 3-grams language model.
2. The one that obtains the highest log probability of sequence of words is selected.

Output: Normalization candidate selected for each OOV word

Post-processing

A selected candidate is uppercased if one of the following conditions is satisfied:

1. The OOV word is in initial position of tweet.
2. The OOV word is preceded by one of the following punctuation marks: ". ! ?".
3. The previous token is an ellipsis mark, and the OOV word begins with a capital letter.

Lexical Resources

1. A dictionary of 619,550 Spanish standard word forms.
2. A normalization dictionary of 529 (non-standard word form, canonical form) pairs.
3. A gazetteer of 53,531 unigrams of proper nouns.

Dataset

Resources provided by the TweetNorm 2013 shared task:

- **Development set:** 475 tweets with 653 OOV tokens.
- **Test set:** 462 tweets with 572 OOV tokens.

Annotation Process

(Real) Tweet

"@bykikomatamoros Qué te pasa a ti con Iker? Diego y valdés lo estarán haciendo bien, pero que rápido olvidamos. A Mou le falta humildad."

(@bykikomatamoros What's wrong with Iker? Diego and valdés are doing it well, but We forget fast. Mou lacks humility.)

OOV tokens:

- Iker (ill-formed OOV) → Íker
- valdés (ill-formed OOV) → Valdés
- Mou (correct OOV)

Metrics

1. $Detection\ rate = \frac{\sum_{t \epsilon T} \sum_{oov \epsilon OOV'_t} [oov \epsilon OOV_t]}{\sum_{t \epsilon T} |OOV_t|}$
2. $Candidate\ coverage = \frac{\sum_{t \epsilon T} \sum_{oov \epsilon OOV'_t} [corr^t_{oov} \epsilon C^t_{oov}]}{\sum_{t \epsilon T} |OOV'_t|}$
3. $Precision\ (P) = \frac{\sum_{t \epsilon T} \sum_{oov \epsilon OOV'_t} [sel^t_{oov} = corr^t_{oov}]}{\sum_{t \epsilon T} |OOV'_t|}$
4. $Recall\ (R) = \frac{\sum_{t \epsilon T} \sum_{oov \epsilon OOV'_t} [sel^t_{oov} = corr^t_{oov}]}{\sum_{t \epsilon T} |OOV_t|}$
5. $F1\text{-}score\ (F) = \frac{2 \times P \times R}{P + R}$

Where,

- T is the collection of tweets,
- OOV_t the set of OOV tokens in tweet $t \epsilon T$,
- OOV'_t the set of detected OOV tokens,
- C^t_{oov} the confusion set of an OOV token,
- sel^t_{oov} the normalization candidate selected,
- $corr^t_{oov}$ the proper correction of the OOV token.

Detecting OOV Words

Approach	**Detection rate**
Tokens without analysis	75%
Tokens without analysis + Named Entities	98.77%

Table: Detection approaches

Estimating the Language Model

N-grams	Precision
2-grams	71.32%
3-grams	71.78%
4-grams	71.32%

Table: Orders of the language model

Results and Evaluation

Active components	**Candidate coverage**	**P**	**R**	**F1**
All	79.65	69.65	69.41	69.53
All – Matching simple rules	68.95	55.96	55.77	55.86
All – Confusion set generation	63.68	61.40	61.19	61.29
All – Phonetic transcription – Suffixes search	80.35	64.39	64.16	64.27
All – Vowels insertion – Edit distance – Split words	74.21	69.30	69.06	69.18
All – Post-processing	72.46	62.11	61.89	62.00

Table: Performance on the test set with different isolated components. Values are given in percentages

Performance Comparison

Rank	System	R
1	RAE	78.32%
2	**ours**	**69.41%**
3	Citius-Imaxin	66.43%
4	UPC	65.56%
5	Elhuyar	63.81%

Table: Performance comparison with participating systems in the TweetNorm 2013 shared task

Final Remarks

- The Lexical Normalization system correctly detected OOV tokens in Spanish tweets.
- Most of the cases the proper normalization of an OOV token was suggested.
- There is a great room for improvement in the candidate selection, which was not properly adapted to the informal genre and the free writing style of Twitter.

Future Work

To build a large corpus of tweets from users who, in theory, write correctly, in order to improve the performance of the candidate selection.

Thank you!