Reflection-based Word Attribute Transfer

Yoichi Ishibashi, Katsuhito Sudoh, Koichiro Yoshino, Satoshi Nakamura
Nara Institute of Science and Technology

SNL-2019

Motivation
- Word attribute transfer can be used for data argumentation
- Analogy-based word attribute transfer requires the explicit knowledge whether the input word is male or female
- We propose Reflection-based word attribute transfer, a method without such explicit knowledge

Conclusion
- Reflection-based word attribute transfer can transform word attributes without explicit knowledge
  - E.g., girl ⇒ boy, boy ⇒ girl
- Reflection has high stability (99.9% of non-attribute words were not changed)
  - E.g., apple ⇒ apple, human ⇒ human
- Reflection has a property similar to logical negation

Relationship with Symbolic Logic
- Reflection is similar to logical negation
  - \( \neg \text{man} = \text{woman} \)
  - \( \neg \neg \text{man} = \text{man} \)

Approach

What is this task?
Transform word attributes on a word embedding space
- father
- mother
- Transform gender
- grandfather
- Transform age
- output: word vector

Input: word vector, target attribute ID (e.g., gender)

Analogy-based Word Attribute Transfer
We can transform an attribute by using analogy, but...
- Problem: Analogy method requires the explicit knowledge whether the input word is for male or female
- Goal: No knowledge = Transform with same function

Reflection-based Word Attribute Transfer
What is Reflection?
- A mapping that transfers vector x to y with a hyperplane (Mirror)
- An identity mapping is obtained when Ref(x) is applied twice

How to apply to word attribute transfer?
- Learn a mirror to transform an attribute (e.g., gender)

Experiment
- Dataset: 106 pairs of gender words (train/val/test = 58/24/24)
  - \(|A| = 58, |V| = 4\) (in the training)
  - Add random noise to \(x\) because the train data size is small
- Accuracy: Transformation accuracy of words with gender attribute
  - E.g. 1 if the nearest neighbor of f (boy) is "girl", otherwise 0
- Stability: Stability of words without gender attribute
  - E.g. 1 if the nearest neighbor of f (apple) is "apple", otherwise 0

Results
- Reflection can transform a word attribute without explicit knowledge (Transformation accuracy is 55.55%)
- Reflection is very stable (99.9% of non-attribute words were not changed)

Method | know ledge | Accuracy (%) | Stability (%) | @1 | @3 | @1 | @3
--- | --- | --- | --- | --- | --- | --- | ---
Ref | Mean@3 | 50.00 | 99.60 | 99.50 | 99.50 | 99.40 | 99.40
Ref + PM | 50.00 | 99.60 | 99.50 | 99.40 | 99.40 | 99.40 | 99.40
MLP | 50.00 | 99.60 | 99.50 | 99.40 | 99.40 | 99.40 | 99.40
Diff(−) | 50.00 | 99.60 | 99.50 | 99.40 | 99.40 | 99.40 | 99.40
AvgDiff(−) | 50.00 | 99.60 | 99.50 | 99.40 | 99.40 | 99.40 | 99.40
AvgDiff(+) | 50.00 | 99.60 | 99.50 | 99.40 | 99.40 | 99.40 | 99.40

How many non-attribute words \(|V|\) do we need when training?
- Reflection has high stability even only \(|V| = 10\)

Method | Accuracy @1 (%) | Stability @1 (%) | @1 | @3 | @1 | @3
--- | --- | --- | --- | --- | --- | ---
Ref | 54.83 | 99.60 | 99.50 | 99.40 | 99.40 | 99.40
Ref + PM | 54.83 | 99.60 | 99.50 | 99.40 | 99.40 | 99.40
MLP | 54.83 | 99.60 | 99.50 | 99.40 | 99.40 | 99.40

Reflection-based word attribute transfer examples
- x when my father was a boy, he had liked the lady who is an actress
- y when he is my father

Apply to other attributes
- Original | she is my mother
- Gender | +
- Age | he is my grandfather
- Tense | he was my grandfather

Reflection-based Word Attribute Transfer