Reflection-based Word Attribute Transfer

Yoichi Ishibashi, Katsuhito Sudoh, Koichiro Yoshino, Satoshi Nakamura

Nara Institute of Science and Technology

SNL-2019

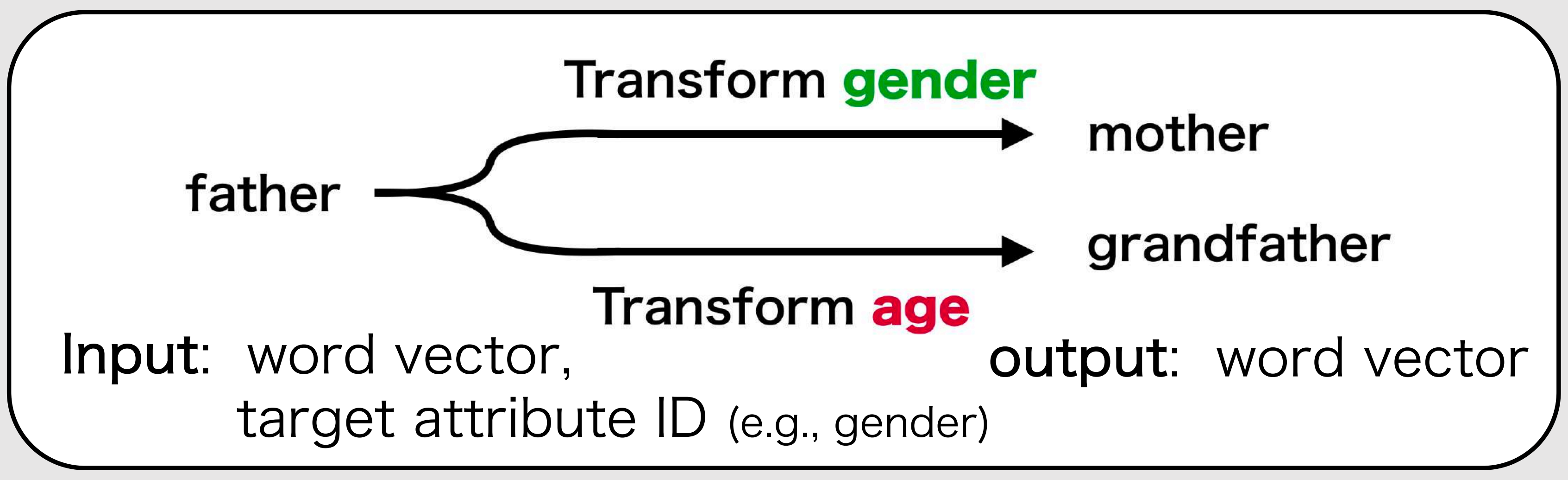
Motivation

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

Approach

What is this task?

Transform word attributes on a word embedding space



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

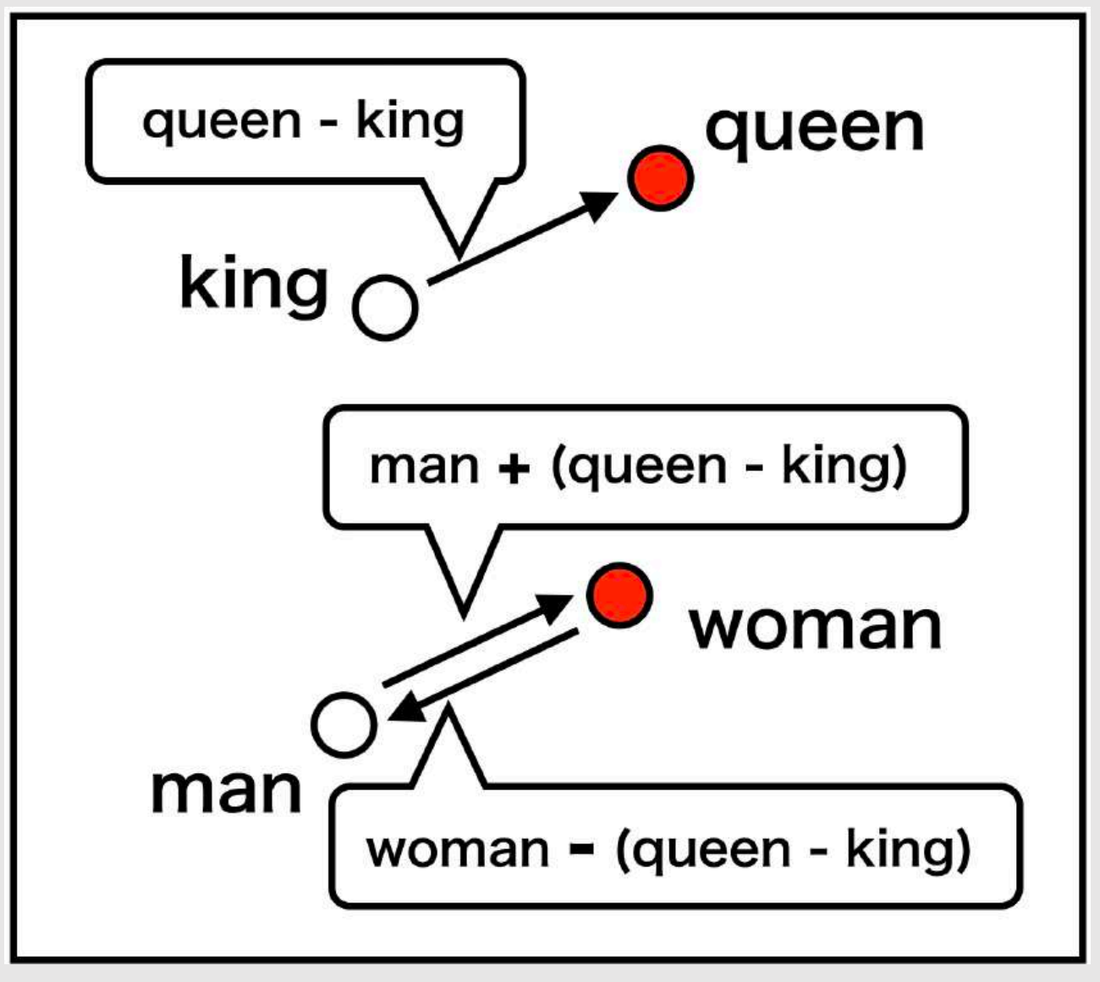
output: word vector

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 x is for male or female

$$f_{gender}(x) = \begin{cases} x + (\text{queen} - \text{king}) & (x \in \text{Male}) \\ x - (\text{queen} - \text{king}) & (x \in \text{Female}) \end{cases}$$



- □ Goal : No knowledge = Transform with same function

$$\left.\begin{aligned} f_{gen}(\text{man}) &= \text{woman} \\ f_{gen}(\text{woman}) &= \text{man} \end{aligned}\right\} f_{gen}(f_{gen}(\text{man})) = \text{man}$$

We need this function

Reflection-based Word Attribute Transfer

What is Reflection?

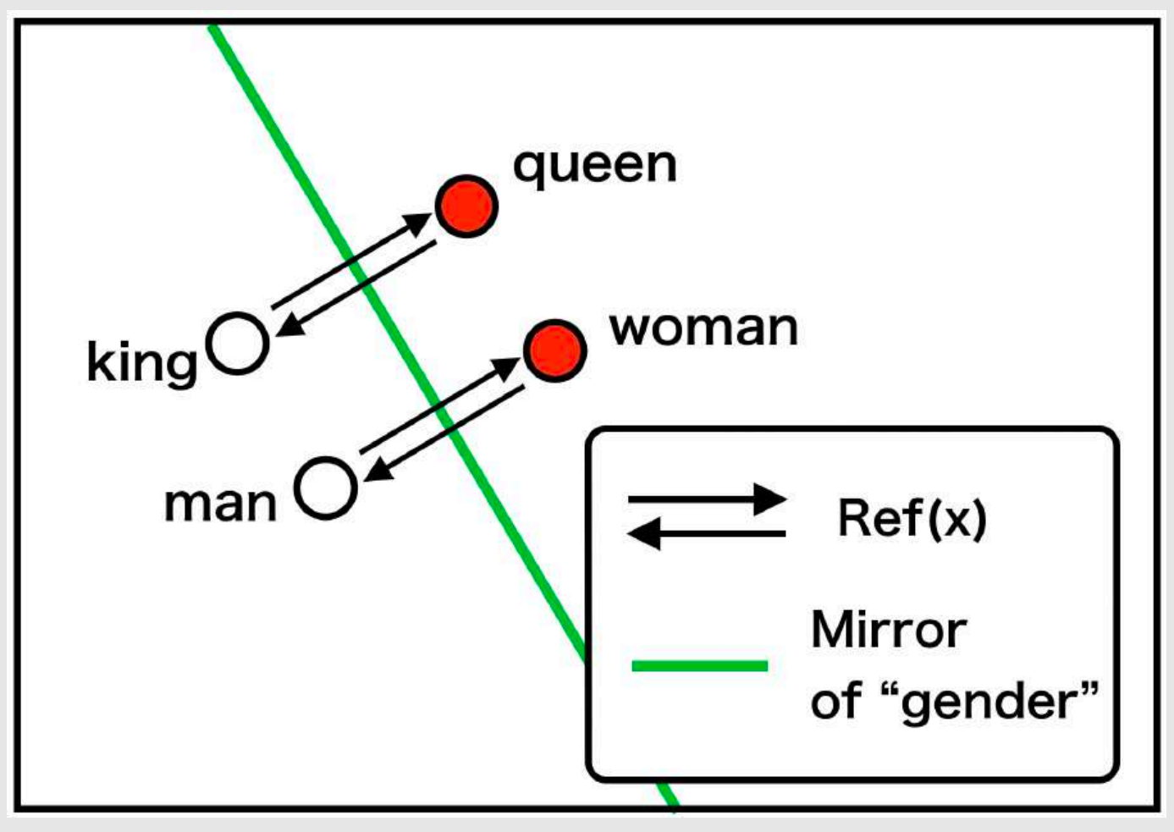
- □ A mapping that transfers vector x to y with a hyperplane (**Mirror**)
- □ An identity mapping is obtained when Ref(x) is applied twice

$$Ref_{a,c}(x) = x - 2\frac{(x-c)^{\mathrm{T}}a}{a^{\mathrm{T}}a}a$$

- ● $Ref_{a,c}(x)$: Reflection
- ● x : Input vector
- ● a, c : Parameters of mirror
- ● $Ref(Ref(\text{man})) = \text{man}$

How to apply to word attribute transfer?

- □ Learn a mirror to transform an attribute (e.g., gender)



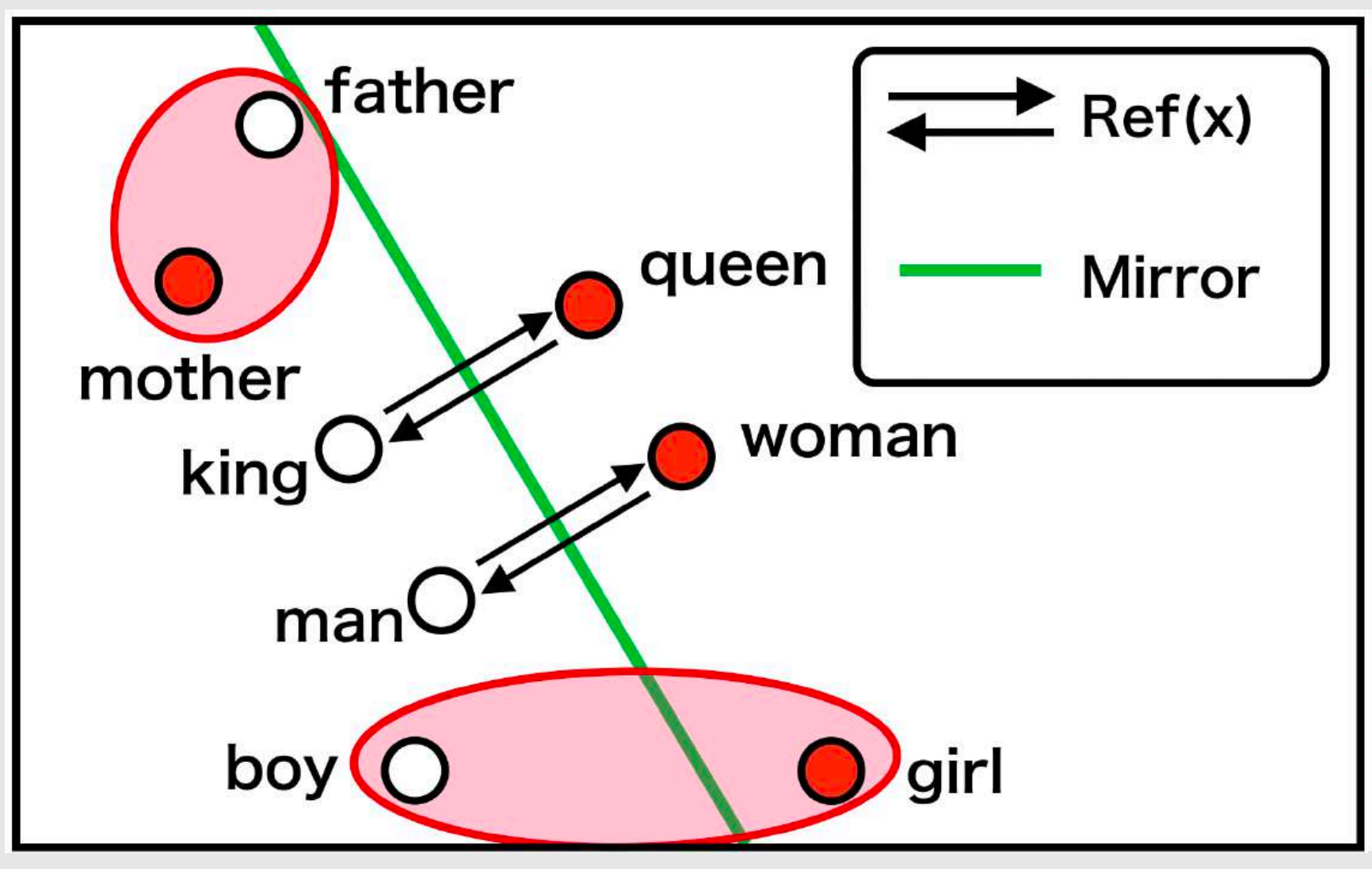
$$y = Ref_{a,c}(x) \qquad a = MLP(\text{attr\_id}) \qquad c = MLP(\text{attr\_id})$$

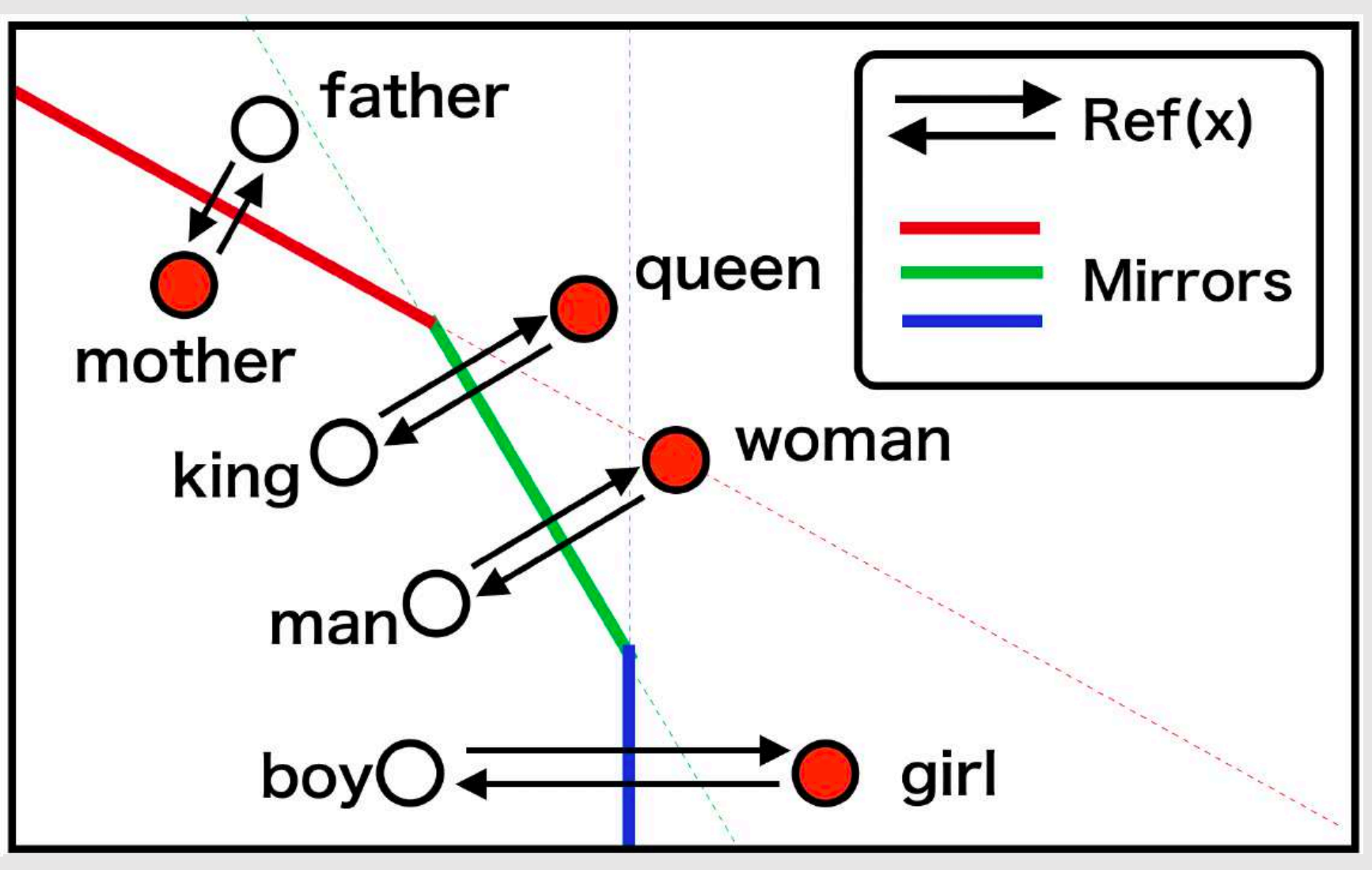
$$L = \frac{1}{|\mathcal{A}|}\sum_{(x_i,t_i)\in\mathcal{A}}(y_i - t_i)^2 + \frac{1}{|\mathcal{N}|}\sum_{x_j\in\mathcal{N}}(y_j - x_j)^2$$

$$(\text{man}, \text{woman}) \in \mathcal{A} \qquad \text{apple} \in \mathcal{N}$$

- □ **Problem** : Linear inseparable
- □ **Idea** : Parameterized mirror
  - ➢ Estimate a mirror from each input word x

$$a = MLP(\text{attr\_id}, \ x\ ) \qquad c = MLP(\text{attr\_id}, \ x\ )$$





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** ¬

$$\neg\,\text{man} = \text{woman} \qquad \neg\neg\,\text{man} = \text{man}$$

$$Ref(\text{man}) = \text{woman} \qquad Ref(Ref(\text{man})) = \text{man}$$

Experiment

- □ **Dataset** : 106 pairs of gender words (train/val/test = 58/24/24)
  - □ $|\mathcal{A}| = 58, |\mathcal{N}| = 4$ (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 (24 words for evaluation)
- □ **Stability** : Stability of words without gender attribute
  - □ E.g. 1 if the nearest neighbor of f (apple) is "apple", otherwise 0 (1000 words for evaluation)

Ref	Reflection-based word attribute transfer
Ref + PM	Reflection-based transfer with parameterized mirror
Diff	Analogy-based transfer with one differential vector
AvgDiff	Analogy-based transfer with average of differential vectors

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	knowledge	Accuracy (%)			Stability (%)		
		Mean@3	@1	@3	Mean@3	@1	@3
Ref		40.27	25.00	54.16	**99.53**	**99.50**	**99.60**
Ref + PM		**55.55**	**45.83**	62.50	96.90	96.50	97.30
MLP		19.44	8.33	33.00	0.00	0.00	0.00
Diff (-)		21.31	7.61	30.74	83.29	79.36	85.87
AvgDiff (-)		23.61	4.16	33.33	98.13	98.10	98.20
Diff	✓	40.65	15.94	57.67	-	-	-
AvgDiff	✓	47.20	12.50	**66.66**	-	-	-

How many non-attribute words $|\mathcal{N}|$ do we need when training?

- □ Reflection has high stability even only $|\mathcal{N}| = 10$

Method	Accuracy @1 (%)				Stability @1 (%)			
	$\|\mathcal{N}\| = 0$	4	10	50	0	4	10	50
Ref	20.83	25.00	25.00	25.00	**97.10**	**99.50**	98.40	95.60
Ref + PM	**45.83**	**45.83**	**37.50**	**29.16**	35.80	96.90	**99.90**	**99.30**
MLP	4.16	8.33	0.00	0.00	0.00	0.00	0.00	0.00

Reflection-based transfer examples

x	**when my father was a boy, he had liked the lady who is an actress**
Ref (x)	**when my mother was a girl, she had liked the gentleman who is an actor**
Ref (Ref (x))	**when my father was a boy, he had liked the lady who is an actress**

Apply to other attributes

Original	she is my mother
+ **Gender**	**he** is my **father**
+ **Age**	he is my **grandfather**
+ **Tense**	he **was** my grandfather