Reflection-based Word Attribute Transfer

Yoichi Ishibashi, Katsuhito Sudoh, Koichiro Yoshino, Satoshi Nakamura

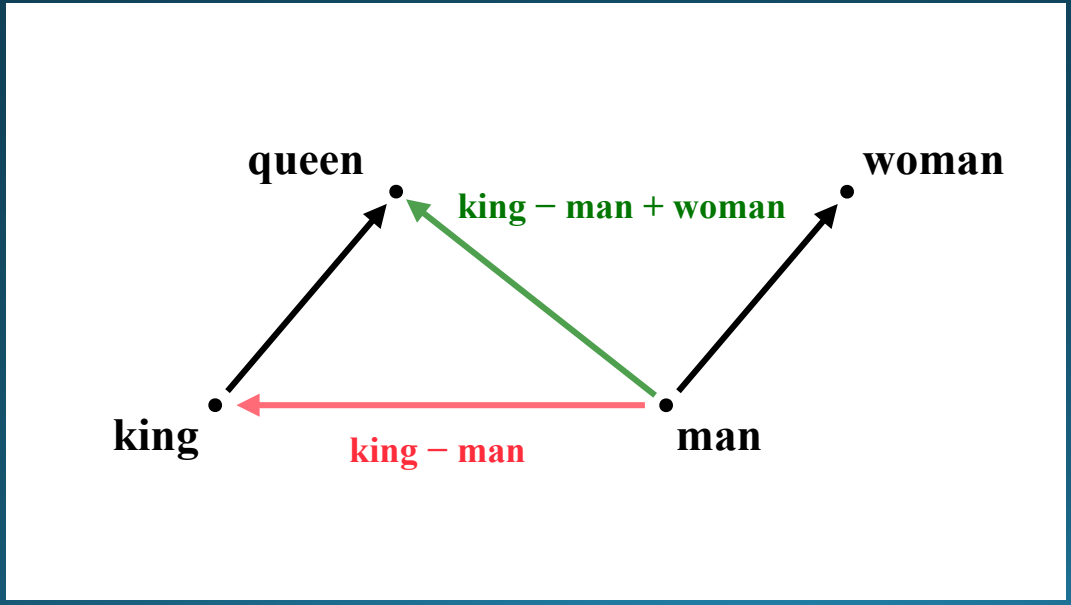
Nara Institute of Science and Technology (NAIST), Japan

Background： Analogy 2

Analogy in the embedding space

- is a operation that transfer word attributes
- Change word attributes (e.g. gender)

$$\overrightarrow{king} - \overrightarrow{man} + \overrightarrow{woman} \approx \overrightarrow{queen}$$

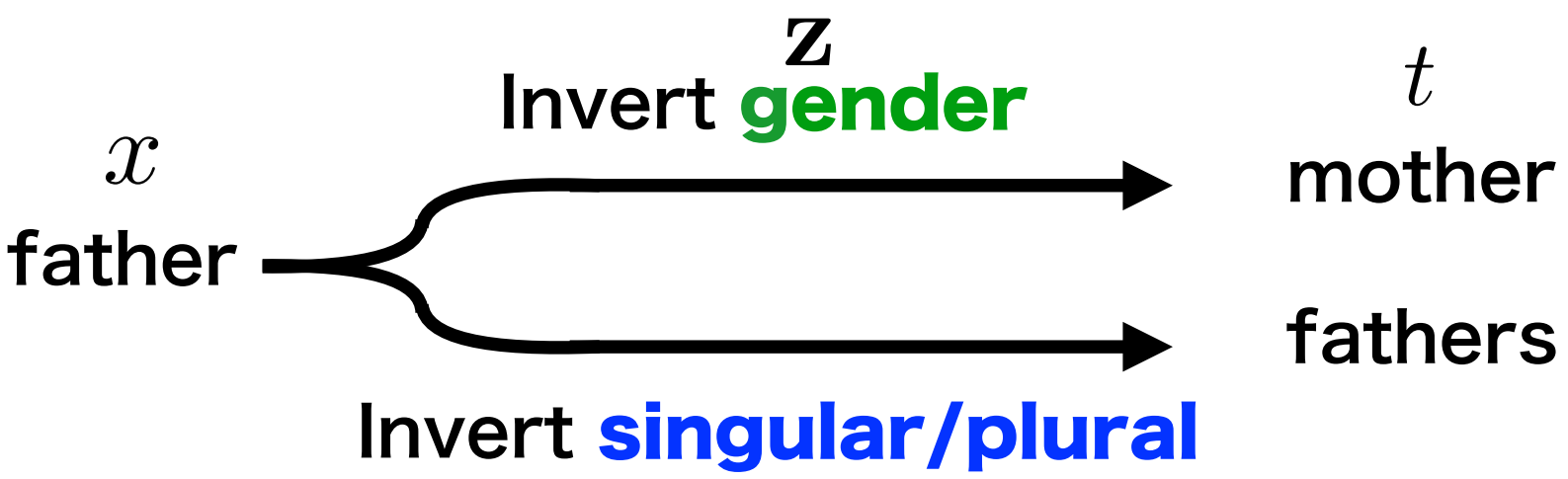


Background : Task overview 3

Word attribute transfer task

- Get a word vector that inverted attribute of an input word vector

$$\mathbf{v}_{\text{mother}} \approx f_{\text{gender}}(\mathbf{v}_{\text{father}})$$



Background： Application 4

What can word attribute transfer be used for?

- E.g. Data Augmentation

Target attribute	Input words	Output words
Gender	I am **his** **mother**.	I am **her** **father**.
Antonym	**Nobody** has a suit.	**Someone** has a suit.
Capital-Country	I live in **Japan**.	I live in **Tokyo**.

Transfer function

$f_{\mathbf{z}}$ transfer words if they have a target attribute $\mathbf{z}$

- E.g. man → woman (attribute: gender)

$$\mathbf{v}_{\text{woman}} \approx f_{\text{gender}}(\mathbf{v}_{\text{man}})$$

$f_{\mathbf{z}}$ does not transfer words if it does not has a target attribute $\mathbf{z}$

- E.g. person → person (attribute: gender)

$$\mathbf{v}_{\text{person}} \approx f_{\text{gender}}(\mathbf{v}_{\text{person}})$$

Analogy-based Word Attribute Transfer 6

Analogy-based word attribute transfer

king − (man − woman) = queen
queen + (man − woman) = king

- **Add** or **subtract** a difference vector

Problem

- **Need explicit knowledge** whether input word has the target attribute or not

Goal

- Transform word attributes **without the explicit knowledge**

Proposed method

Ideal function

What is an ideal transfer function?

- No explicit knowledge

  = Transfer any words with the same function

$$\mathbf{v_{man}} = f(\mathbf{v_{woman}})$$

$$\mathbf{v_{woman}} = f(\mathbf{v_{man}})$$

$$\mathbf{v_{person}} = f(\mathbf{v_{person}})$$

Combine above formulas

$$\mathbf{v}_x = f(f(\mathbf{v}_x))$$

← Nature of the ideal function

Reflection

Reflection is an ideal function

- Transfer any words with the same function

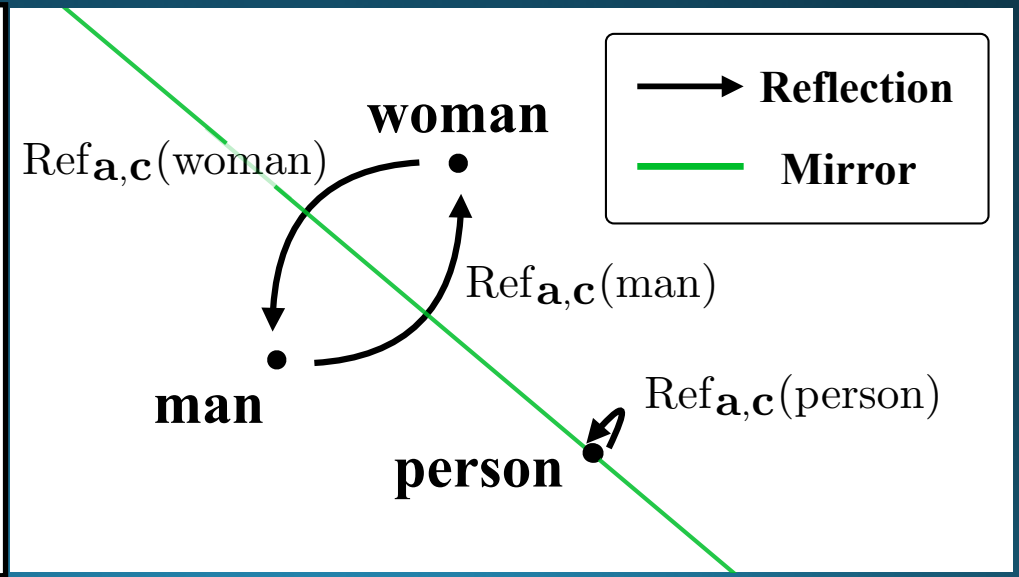
$$\mathbf{v} = Ref_{\mathbf{a},\mathbf{c}}(\, Ref_{\mathbf{a},\mathbf{c}}(\mathbf{v})\,) \qquad \mathrm{Ref}_{\mathbf{a},\mathbf{c}}(\mathbf{v}) = \mathbf{v} - 2\frac{(\mathbf{v}-\mathbf{c})\cdot\mathbf{a}}{\mathbf{a}\cdot\mathbf{a}}\mathbf{a}$$

- Move two vectors through a hyperplane (**mirror**)

$$\mathrm{man} = \mathrm{Ref}_{\mathbf{a},\mathbf{c}}(\mathrm{woman})$$

$$\mathrm{woman} = \mathrm{Ref}_{\mathbf{a},\mathbf{c}}(\mathrm{man})$$

$$\mathrm{person} = \mathrm{Ref}_{\mathbf{a},\mathbf{c}}(\mathrm{person})$$



Reflection-based Word Attribute Transfer 10

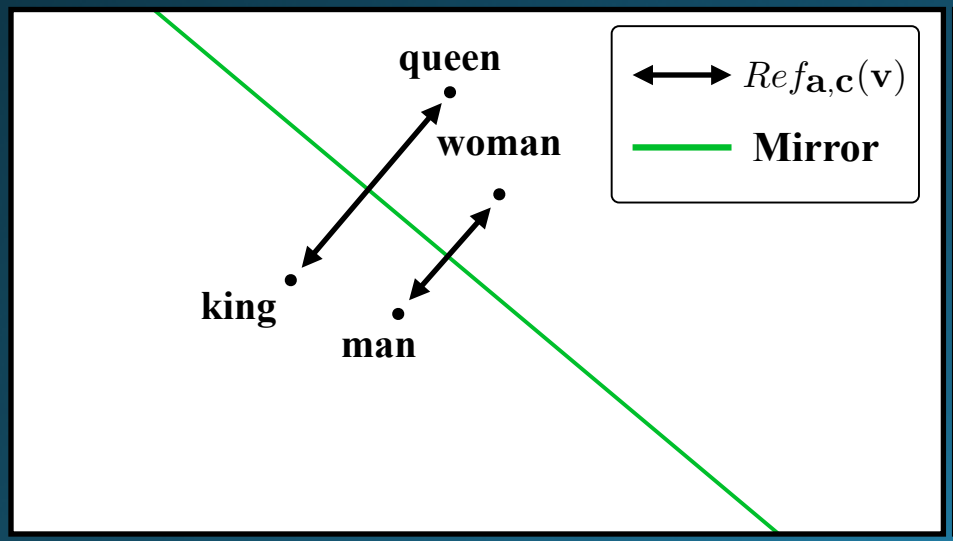
How to apply to word attribute transfer?

- Transfer an input word vector $\mathbf{v}_x$
  to a target word vector $\mathbf{v}_t$

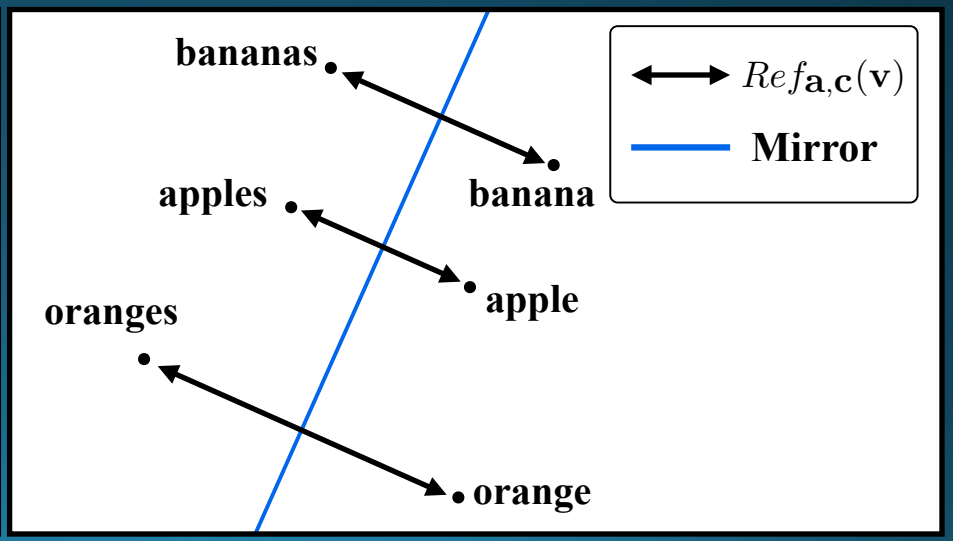
$$\mathbf{v}_t \approx \mathbf{v}_y = \mathrm{Ref}_{\mathbf{a},\mathbf{c}}(\mathbf{v}_x)$$

- **Learn a mirror for each attributes**

Male⇔Female



Singular⇔Plural



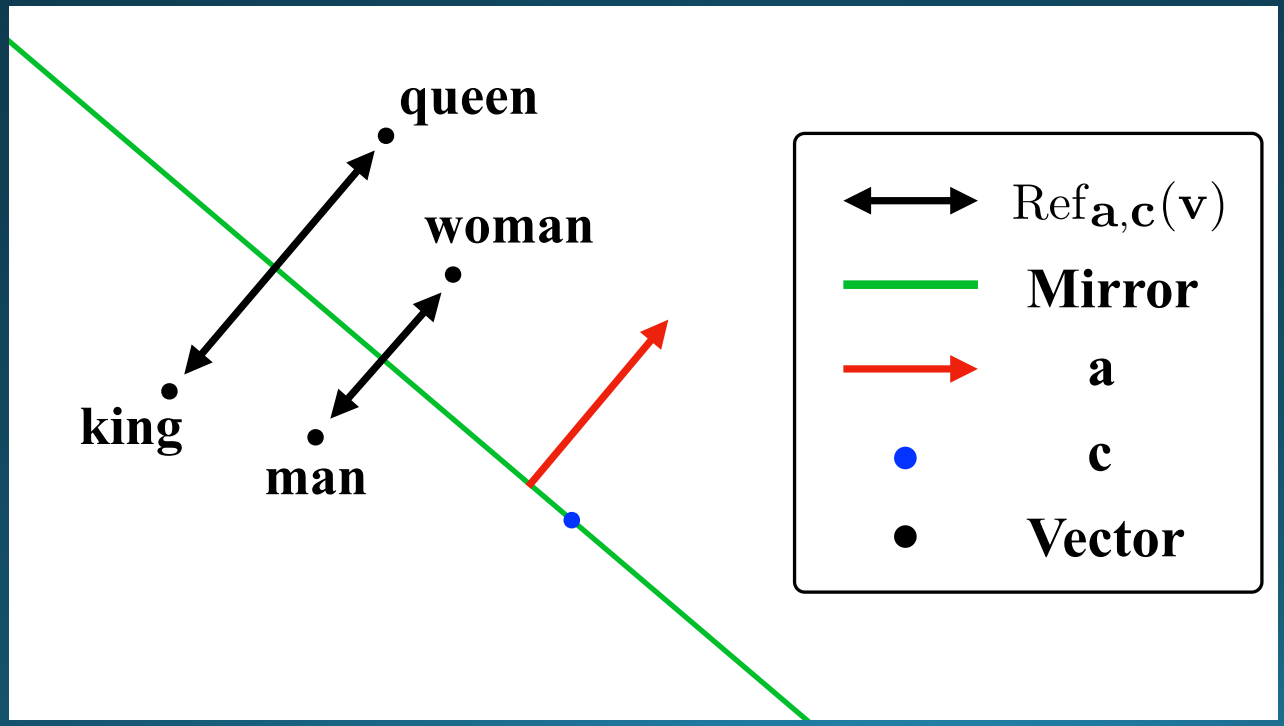
Reflection-based Word Attribute Transfer 11

How to learn the mirror?

Idea : Estimate $\mathbf{a}$ and $\mathbf{c}$ by MLP

$\mathbf{a}$ … A vector orthogonal to the mirror

$\mathbf{c}$ … A point through which the mirror passes



Two types of mirror estimation

① **Single mirror**

Estimate from an attribute $\mathbf{z}$

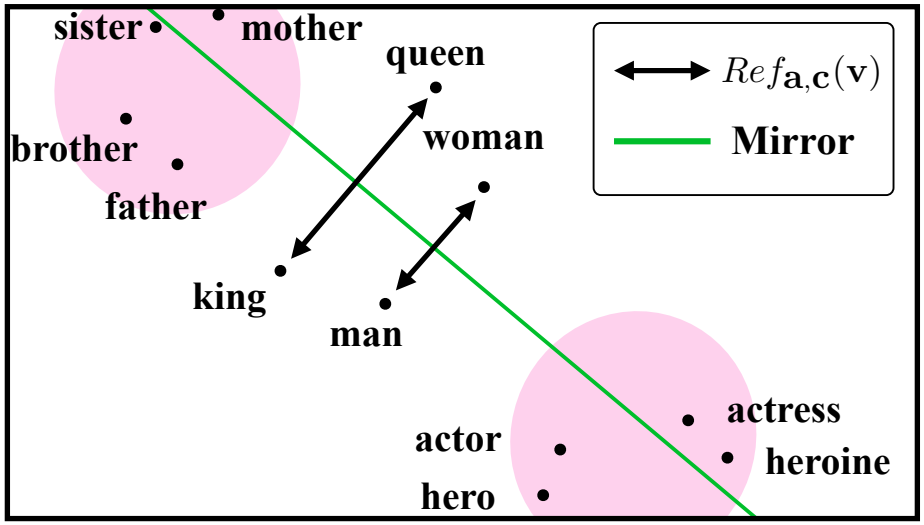
⇒ Some pairs are non-transferable

② **Parameterized mirrors**

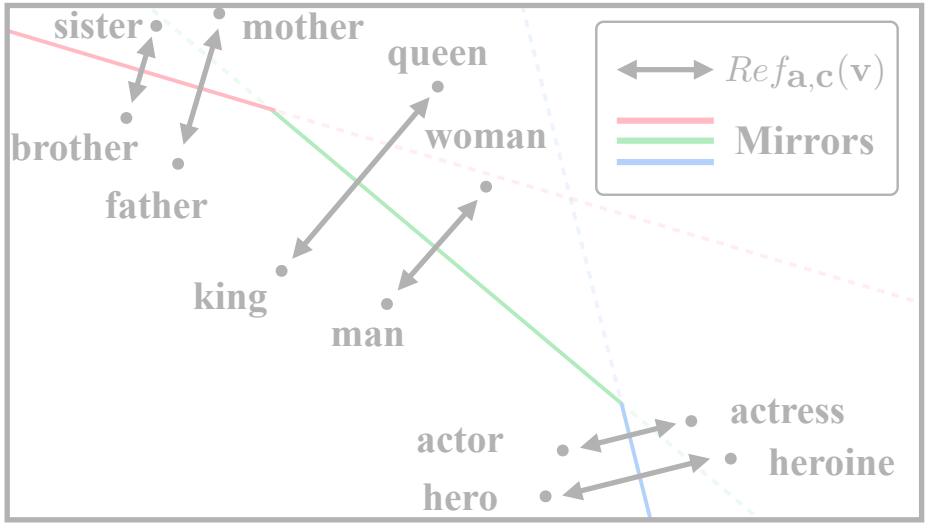
Estimate from $\mathbf{z}$ and an input word vector $\mathbf{v}_x$

⇒ Work more flexibly

$$\mathbf{a} = \mathrm{MLP}_{\theta_1}(\mathbf{z})$$
$$\mathbf{c} = \mathrm{MLP}_{\theta_2}(\mathbf{z})$$



$$\mathbf{a} = \mathrm{MLP}_{\theta_1}([\mathbf{z}; \mathbf{v}_x])$$
$$\mathbf{c} = \mathrm{MLP}_{\theta_2}([\mathbf{z}; \mathbf{v}_x])$$



Two types of mirror estimation

① **Single mirror**

Estimate from an attribute $\mathbf{z}$

⇒ Some pairs are non-transferable

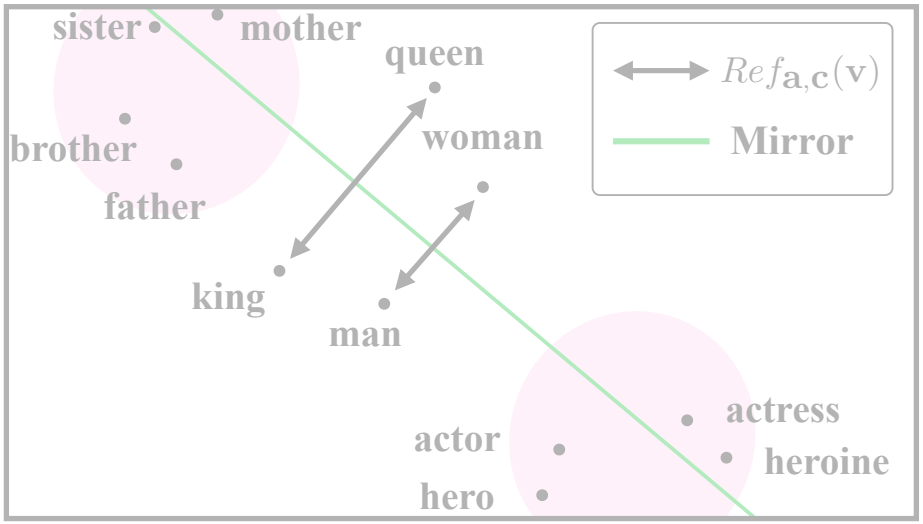
② **Parameterized mirrors**

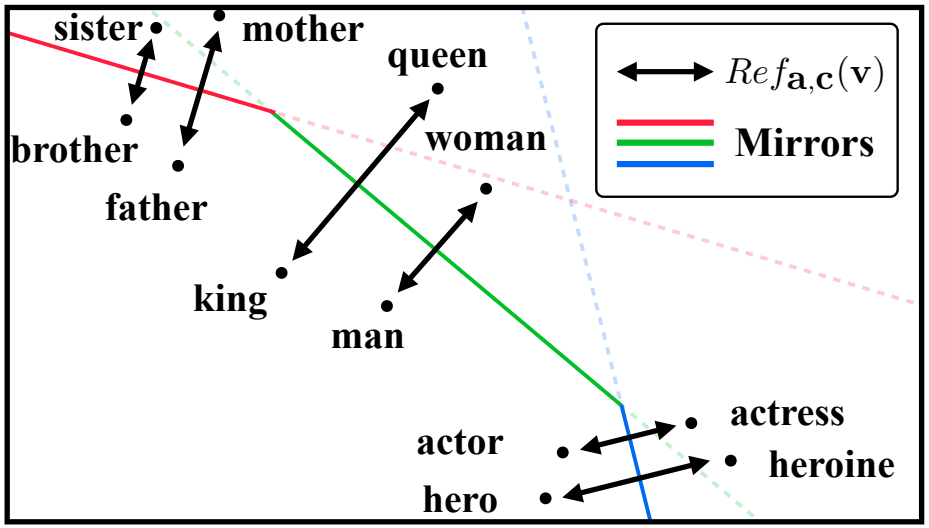
Estimate from $\mathbf{z}$ and an input word vector $\mathbf{v}_x$

⇒ Work more flexibly

$$\mathbf{a} = \mathrm{MLP}_{\theta_1}(\mathbf{z})$$
$$\mathbf{c} = \mathrm{MLP}_{\theta_2}(\mathbf{z})$$

$$\mathbf{a} = \mathrm{MLP}_{\theta_1}([\mathbf{z}; \mathbf{v}_x])$$
$$\mathbf{c} = \mathrm{MLP}_{\theta_2}([\mathbf{z}; \mathbf{v}_x])$$





Experiments

Purpose

Compare reflection and baselines

- **Four different attributes**
  - Male-Female, Singular-Plural
    Capital-Country, Antonyms
- **Two pre-trained word embeddings**
  - word2vec (SGNS), GloVe
- **Two evaluation metrics**
  - Accuracy, Stability

Dataset

- Attribute words … Four different binary attributes

Attribute (z)	Train	Val	Test	Example (x, t)
Male-Female (MF)	29	12	12	(king, queen)
Singular-Plural (SP)	90	25	25	(king, kings)
Capital-Country (CC)	59	25	25	(Japan, Tokyo)
Antonym (AN)	1354	290	290	(good, bad)

- Non-attribute words
  - Train $0 \leq |\mathcal{N}_{train}| \leq 50$ Test $|\mathcal{N}_{test}| = 1000$

Evaluation metrics

① **Accuracy:** Ratio of attribute words transferred

② **Stability:** Ratio of **non**-attribute words **not** transferred

Results (Accuracy)

Best method: Reflection with parameterized mirrors

→ High performance in both accuracy and stability

Worst method: MLP

Method	GloVe							
	Accuracy (%)				Stability (%)			
	MF	SP	CC	AN	MF	SP	CC	AN
Ref	12.5	2.0	26.0	0.0	100.0	100.0	100.0	100.0
Ref+PM	**45.8**	**50.0**	**76.0**	33.5	99.7	99.1	99.2	100.0
MLP	4.2	10.0	18.0	**36.7**	5.1	7.0	5.2	1.2
Diff+	25.0	2.0	26.0	-	99.3	94.2	99.3	-
Diff-	25.0	2.0	24.0	-	100.0	99.9	99.5	-

MF: Male-Female, SP: Singular-Plural, CC: Country-Capital, AN: Antonym

Results (Stability)

Best method: Reflection with parameterized mirrors

→ High performance in both accuracy and stability

Worst method: MLP

Method	GloVe							
	Accuracy (%)				Stability (%)			
	MF	SP	CC	AN	MF	SP	CC	AN
Ref	12.5	2.0	26.0	0.0	**100.0**	**100.0**	**100.0**	**100.0**
Ref+PM	45.8	50.0	76.0	33.5	99.7	99.1	99.2	**100.0**
MLP	4.2	10.0	18.0	36.7	**5.1**	**7.0**	**5.2**	**1.2**
Diff+	25.0	2.0	26.0	-	99.3	94.2	99.3	-
Diff-	25.0	2.0	24.0	-	**100.0**	99.9	99.5	-

MF: Male-Female, SP: Singular-Plural, CC: Country-Capital, AN: Antonym

Transfer examples

Reflection with parameterized mirrors (Ref+PM) can **selectively** transfer words without the knowledge

boy → **girl**　　**you** → **you**

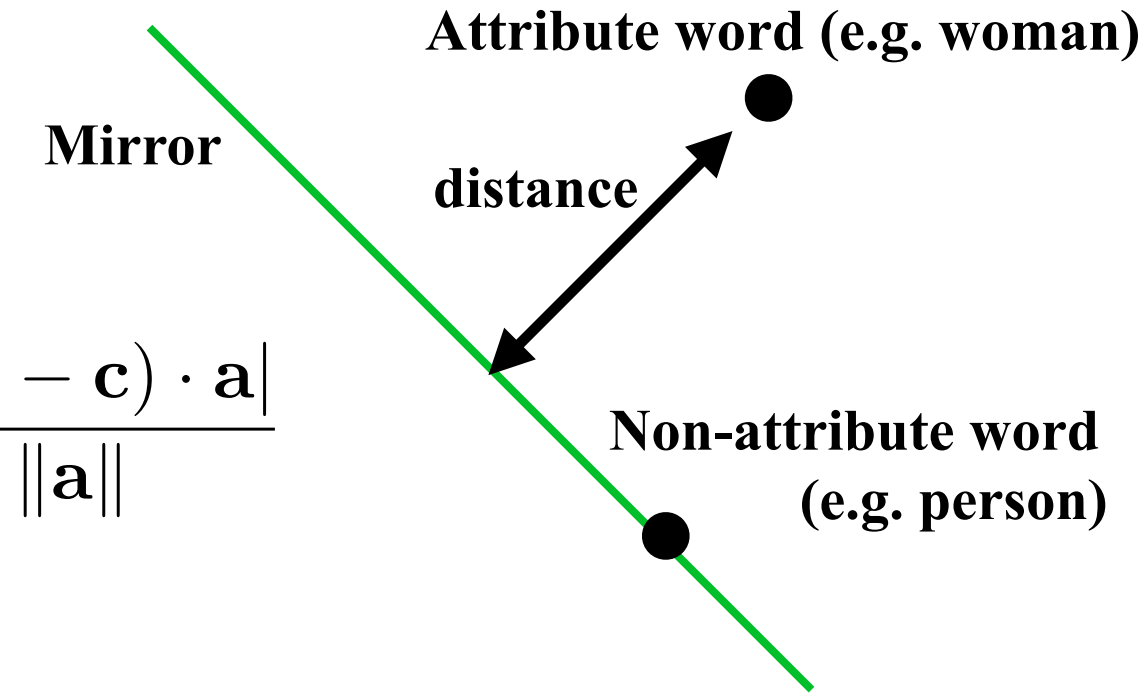
Input	the **woman** got married when you were a **boy**.
Ref	the **woman** got married when you were a **boy**.
Ref+PM	the **man** got married when you were a **girl**.
MLP	By_Katie_Klingsporn **girlfriend** Valerie_Glodowski fiancee Doughty_Evening_Chronicle ma'am Bob_Grossweiner_& a **mother**.
Diff+	the **man** got married when you were a **boy**.
Diff-	the **woman** got married when you were a **girl**.

Why is the reflection very stabile? 21

Hypothesis: Non-attribute word distributes on its mirror

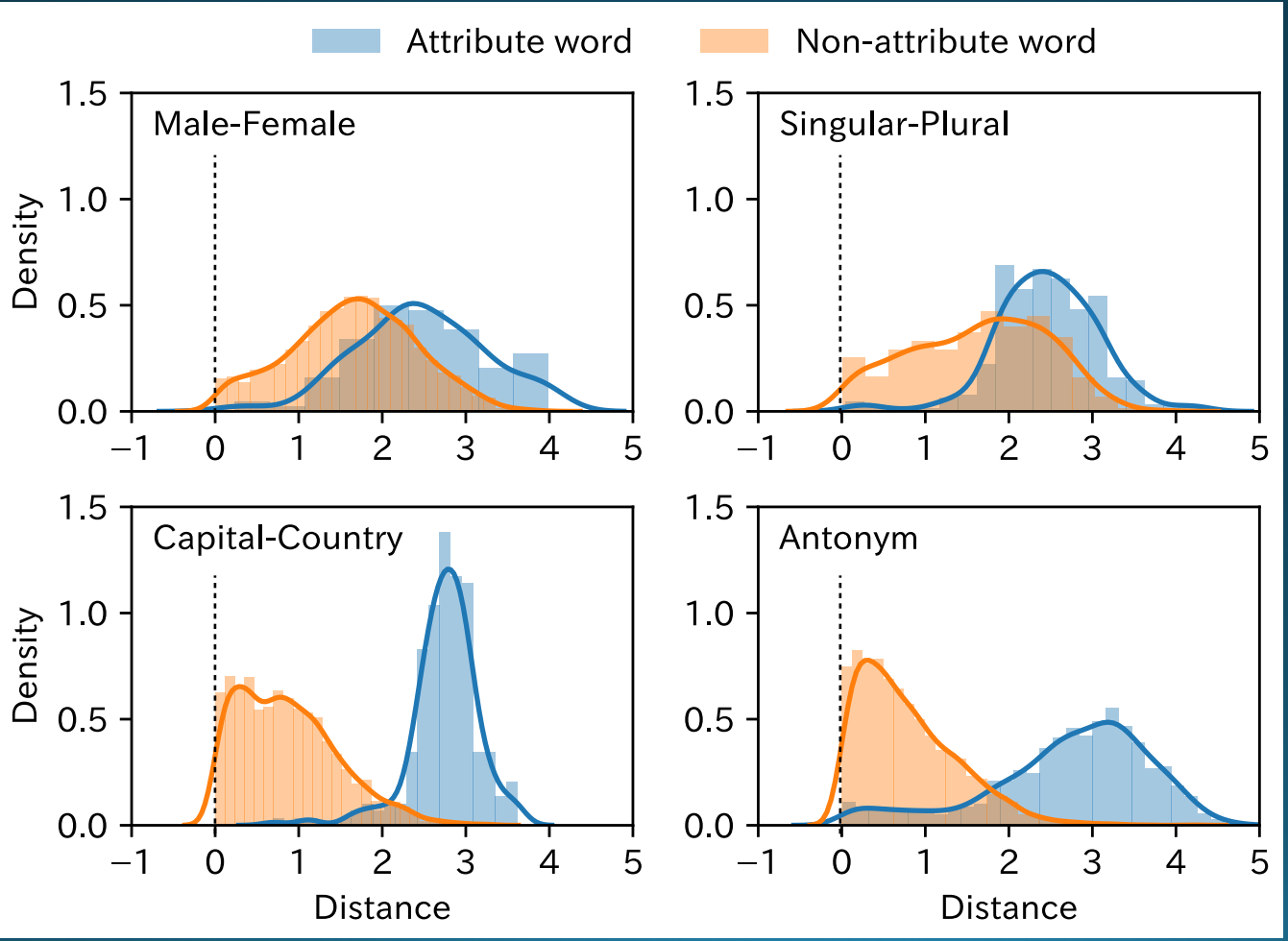
→ Visualize the distance between a word vector and its mirror

$$\text{distance} = \frac{|(\mathbf{v}_x - \mathbf{c}) \cdot \mathbf{a}|}{\|\mathbf{a}\|}$$



Distance between the word and its mirror 22

- Attribute words distributed apart from the mirror
- Non-attribute words distributed near the mirror



Summary

Background

- Word attribute transfer task
- Analogy can be used for the transfer
- Analogy-based transfer requires the explicit knowledge

Proposed method

- Reflection-based word attribute transfer
- Reflection is an ideal mapping for word attribute transfer

Experimental results

- Reflection-based transfer achieved best performance
- Reflection transfers attribute words
  e.g. man → woman
- Reflection does not transfer non-attribute words
  e.g. person → person