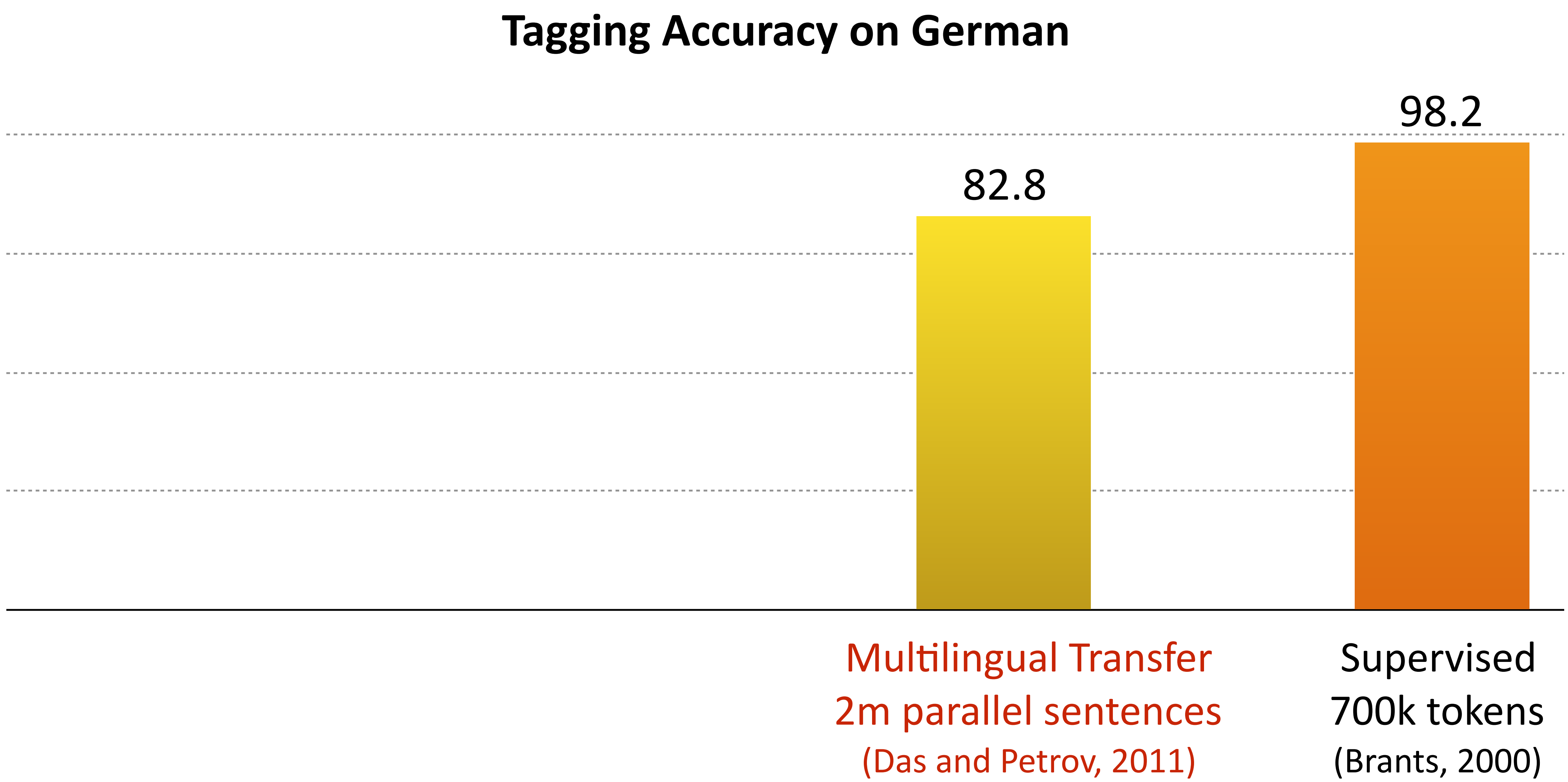
Ten Pairs to Tag – Multilingual POS Tagging via Coarse Mapping between Embeddings

Yuan Zhang, David Gaddy, Regina Barzilay, Tommi Jaakkola

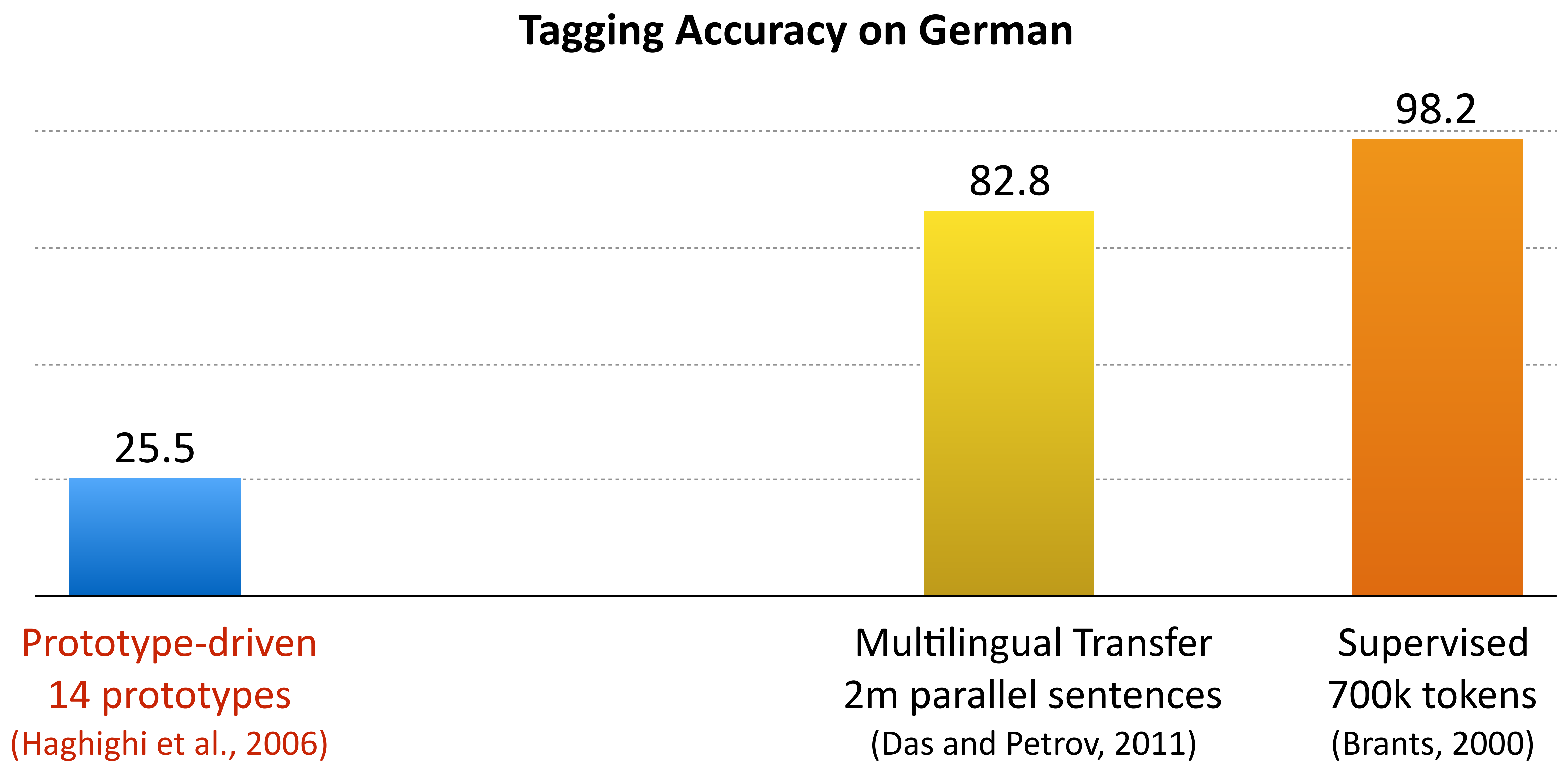
MIT, CSAIL

Multilingual Transfer of POS Tagging

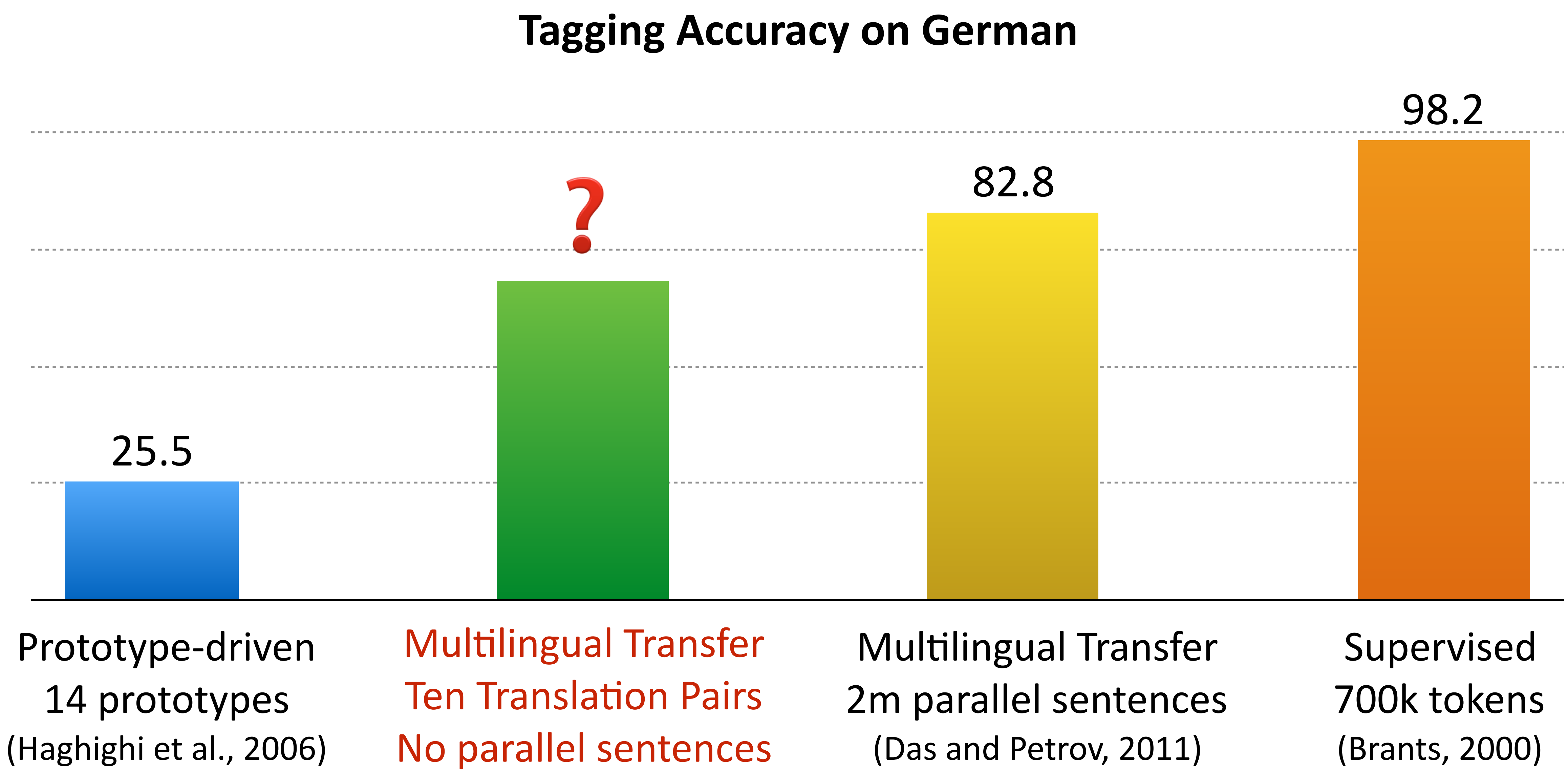
Multilingual Transfer of POS Tagging



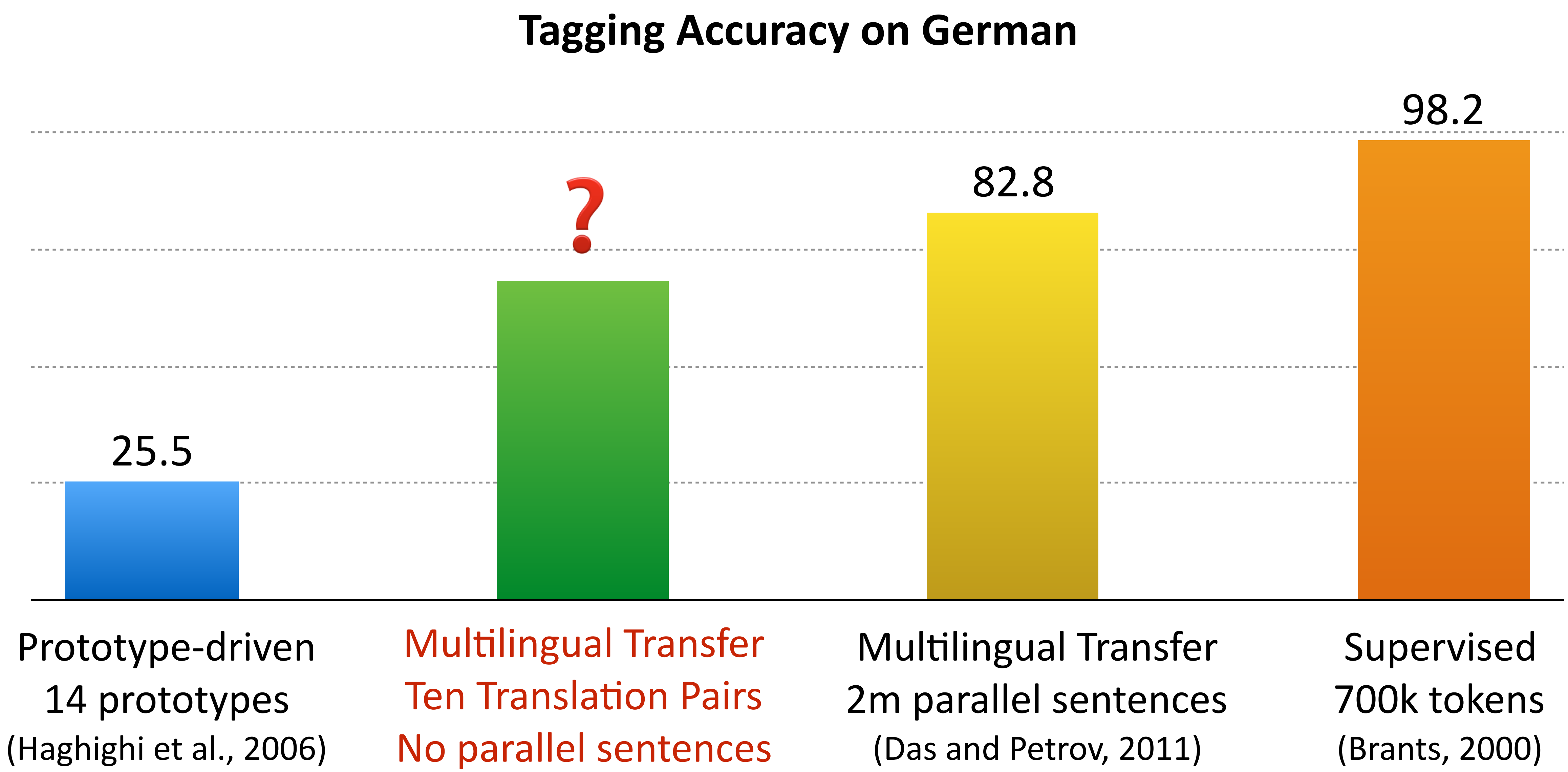
Multilingual Transfer of POS Tagging



Multilingual Transfer of POS Tagging



Multilingual Transfer of POS Tagging



How little parallel data is necessary to enable multilingual transfer?

Our Work

- Task: multilingual transfer of part-of-speech (POS) tagging
- Data:

	Source	Target
Labeled	✓	✗
Unlabeled	✓	✓ *(non-parallel data)*

Our Work

- Task: multilingual transfer of part-of-speech (POS) tagging
- Data:

	Source	Target
Labeled	✓	✗
Unlabeled	✓	✓ *(non-parallel data)*

Ten Translation Pairs

. || .	und || and
, || ,	dem || the
der || the	von || from
die || the	- || -
in || in	zu || to

Our Work

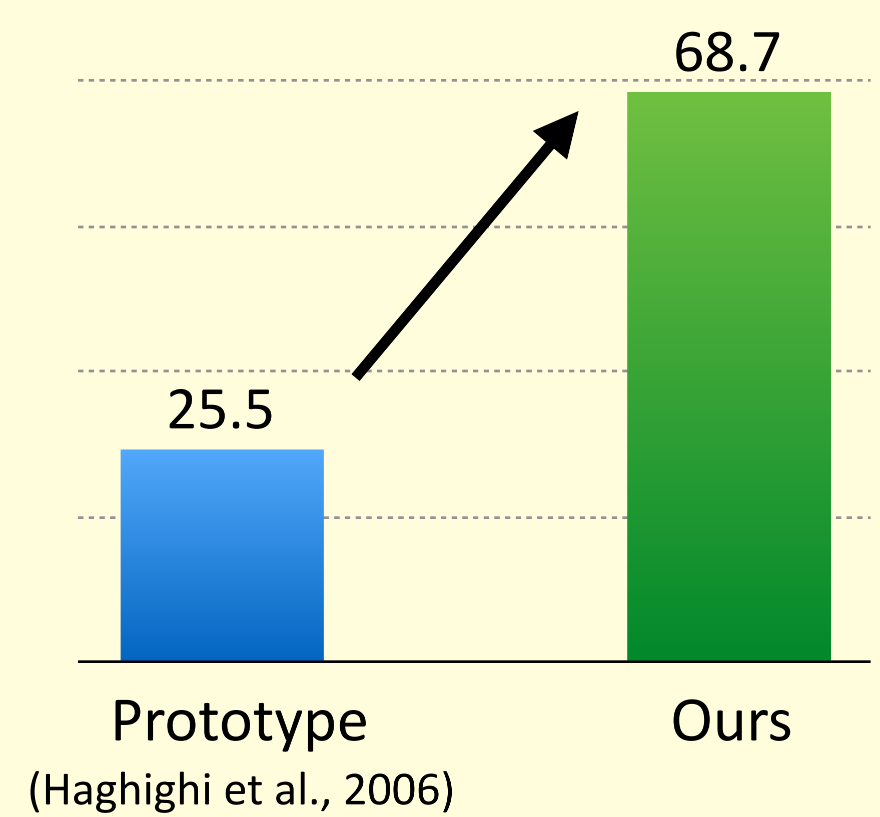
- Task: multilingual transfer of part-of-speech (POS) tagging
- Data:

	Source	Target
Labeled	✓	✗
Unlabeled	✓	✓ *(non-parallel data)*

Ten Translation Pairs

. || .	und || and
, || ,	dem || the
der || the	von || from
die || the	- || -
in || in	zu || to

POS Accuracy on German



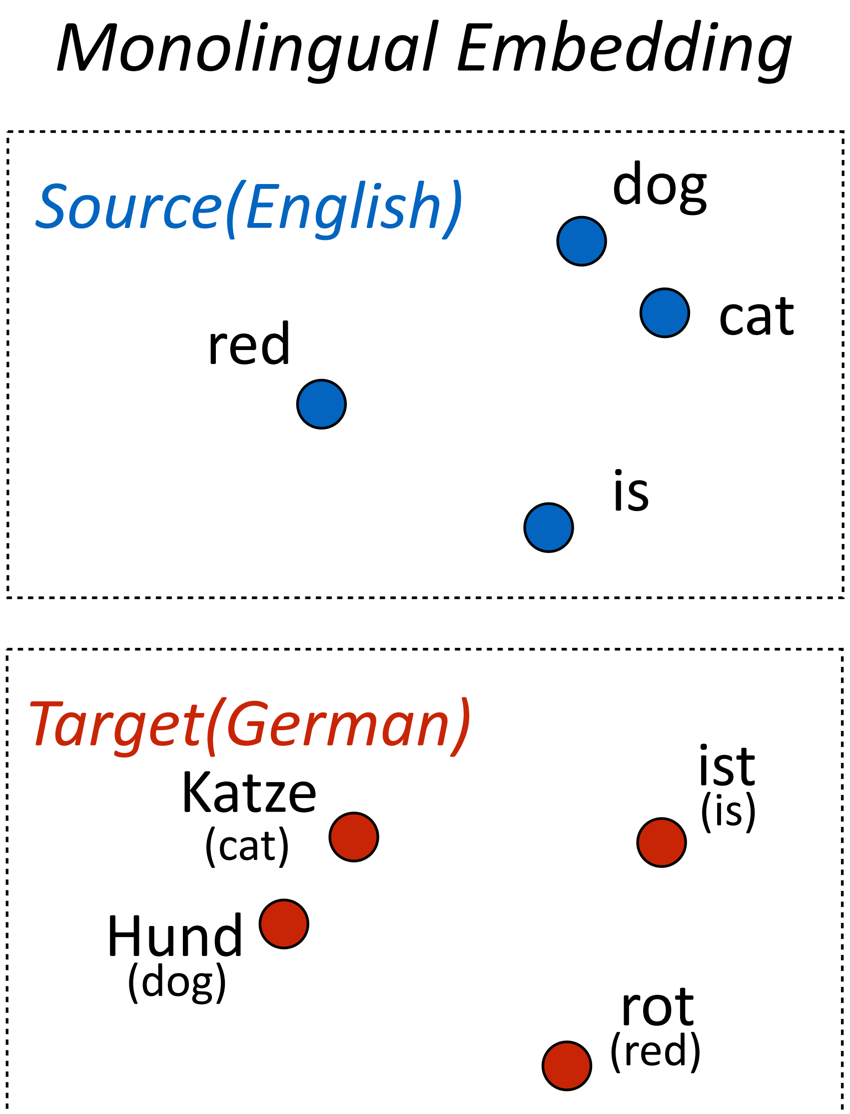
Our Two-phase Method

1. Learn coarse mapping between embeddings via ten translation pairs

2. Refine embedding transformations and model parameters via unsupervised learning on the target language

Coarse Mapping between Embeddings

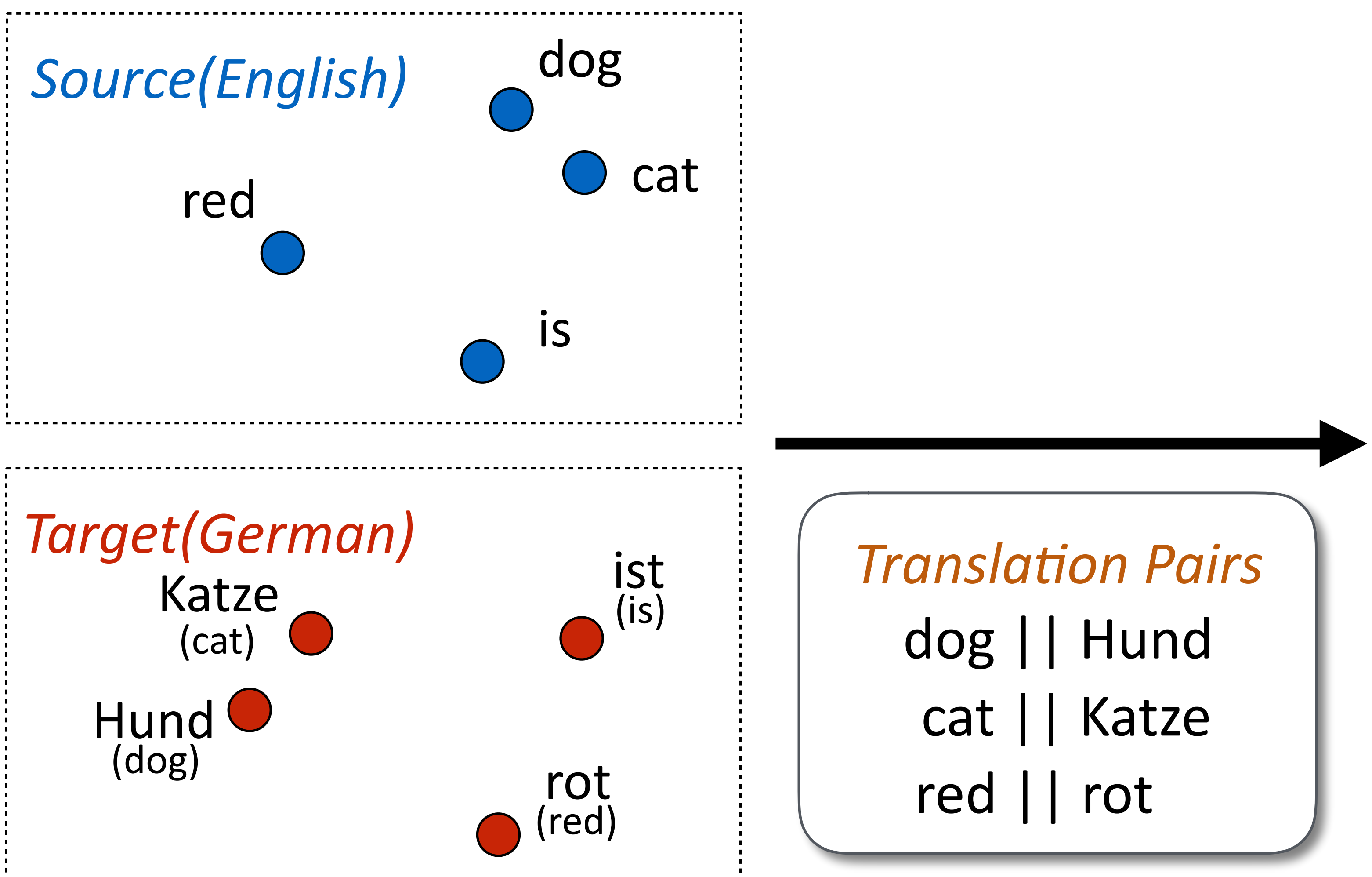
- Goal: find a linear transformation from target to source embedding space
- Objective: minimize the distance between translation pairs



Coarse Mapping between Embeddings

- Goal: find a linear transformation from target to source embedding space
- Objective: minimize the distance between translation pairs

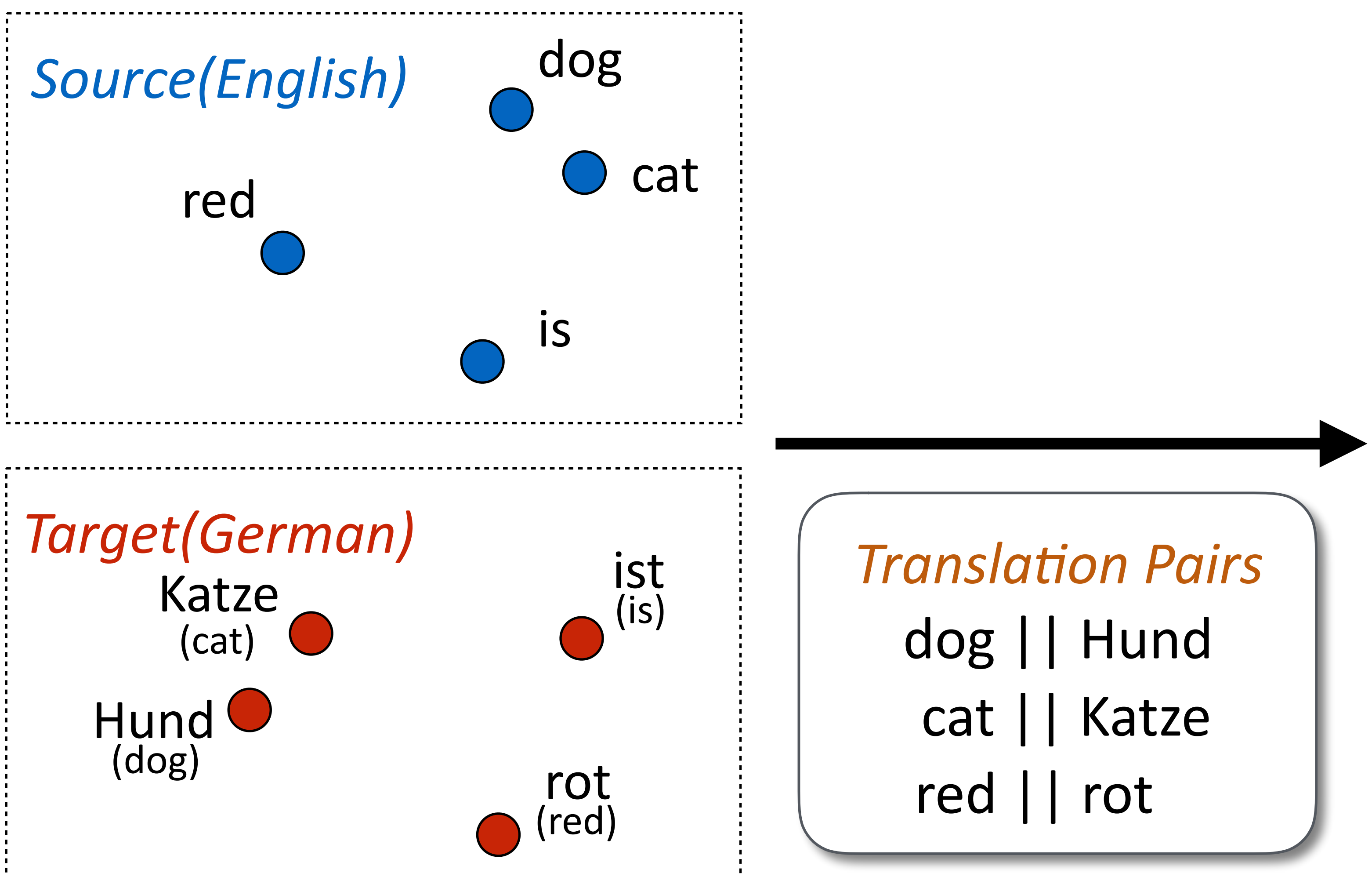
Monolingual Embedding



Coarse Mapping between Embeddings

- Goal: find a linear transformation from target to source embedding space
- Objective: minimize the distance between translation pairs

Monolingual Embedding

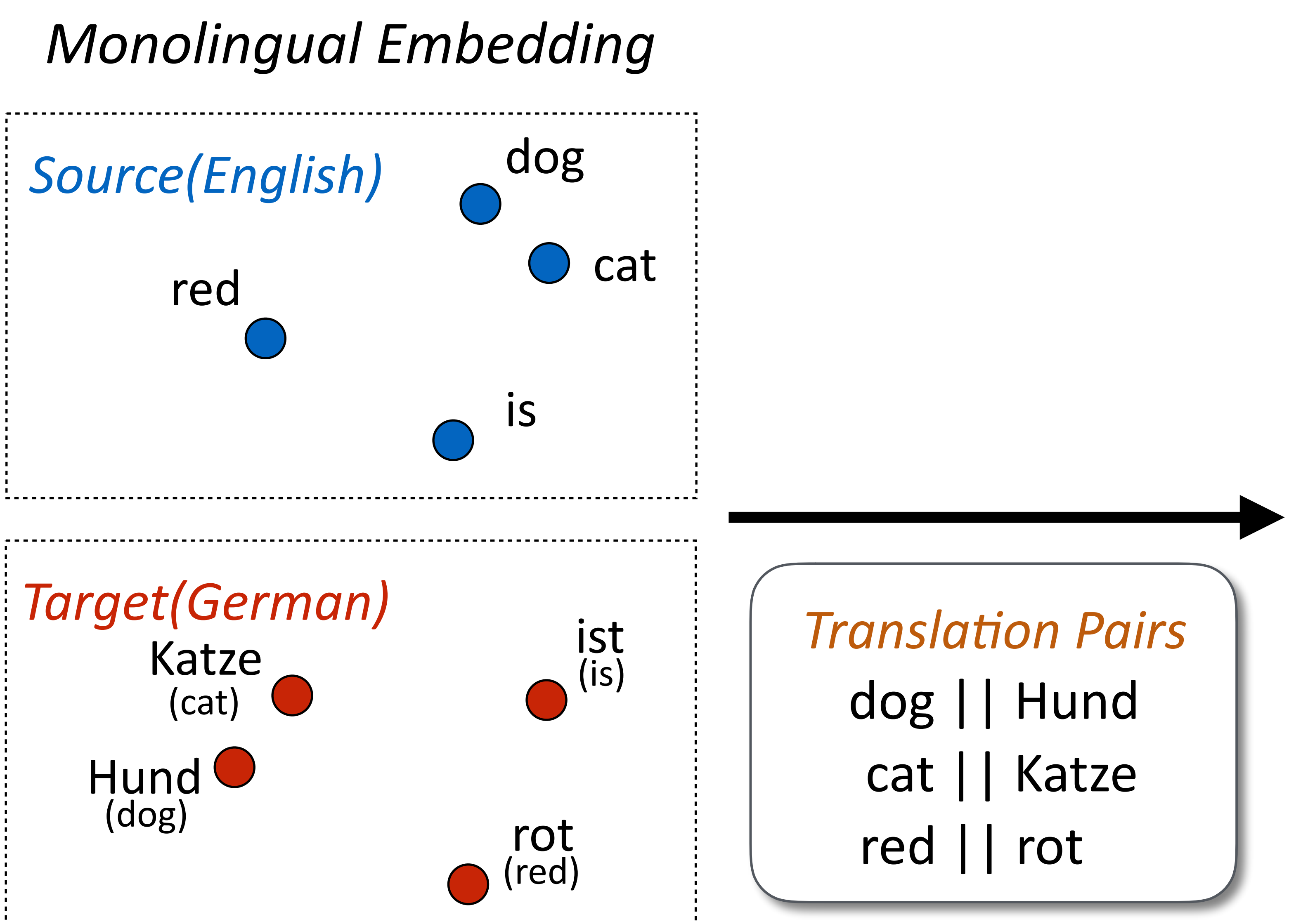


Too many degrees of freedom

dimension:	20
# pairs:	10
degree of freedom:	10

Coarse Mapping between Embeddings

- Goal: find a linear transformation from target to source embedding space
- Objective: minimize the distance between translation pairs



Too many degrees of freedom

dimension: 20

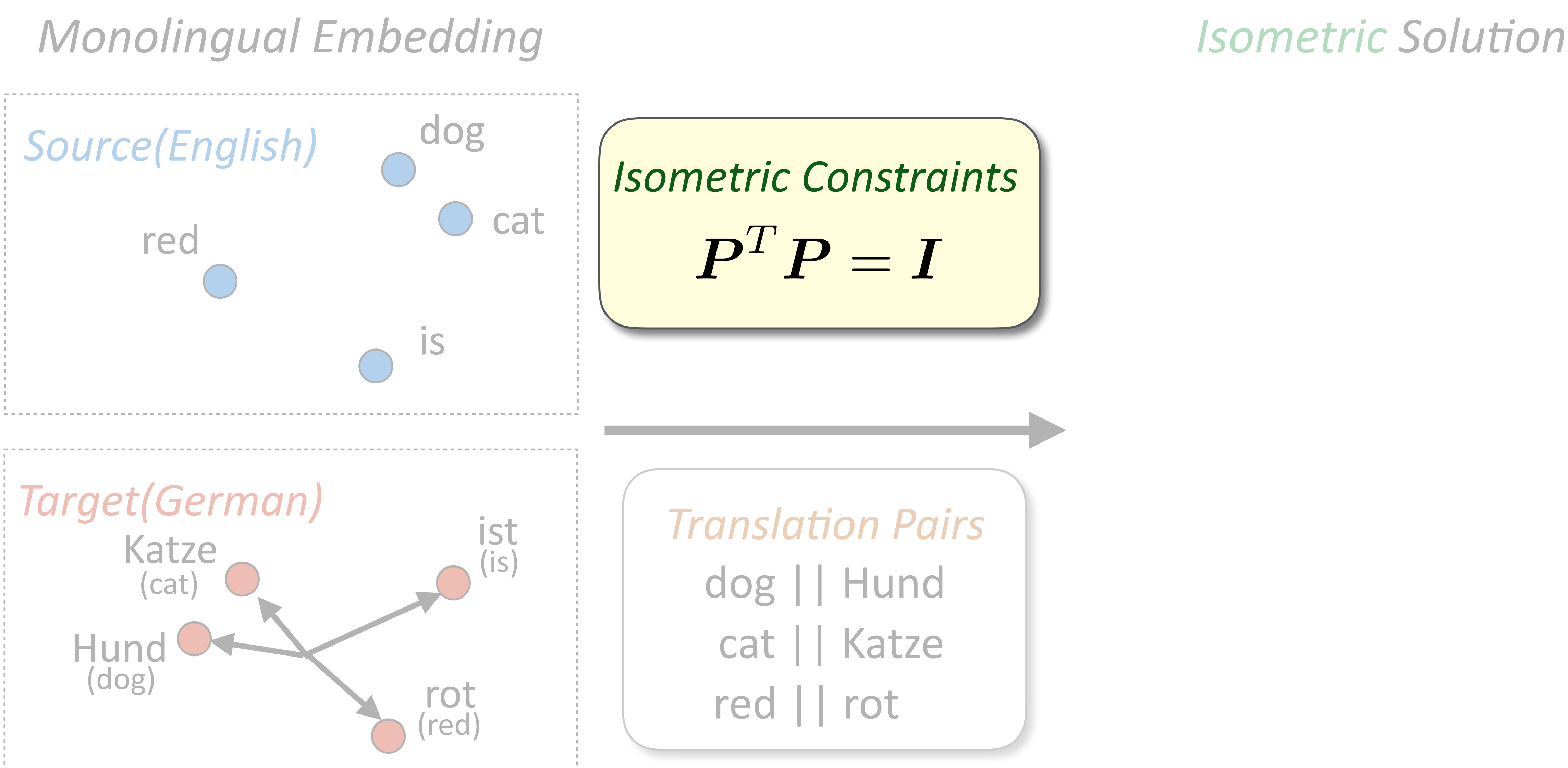
 # pairs: 10

degree of freedom: 10

Solutions need to be constrained!

Our Solution: Isometric Constraints

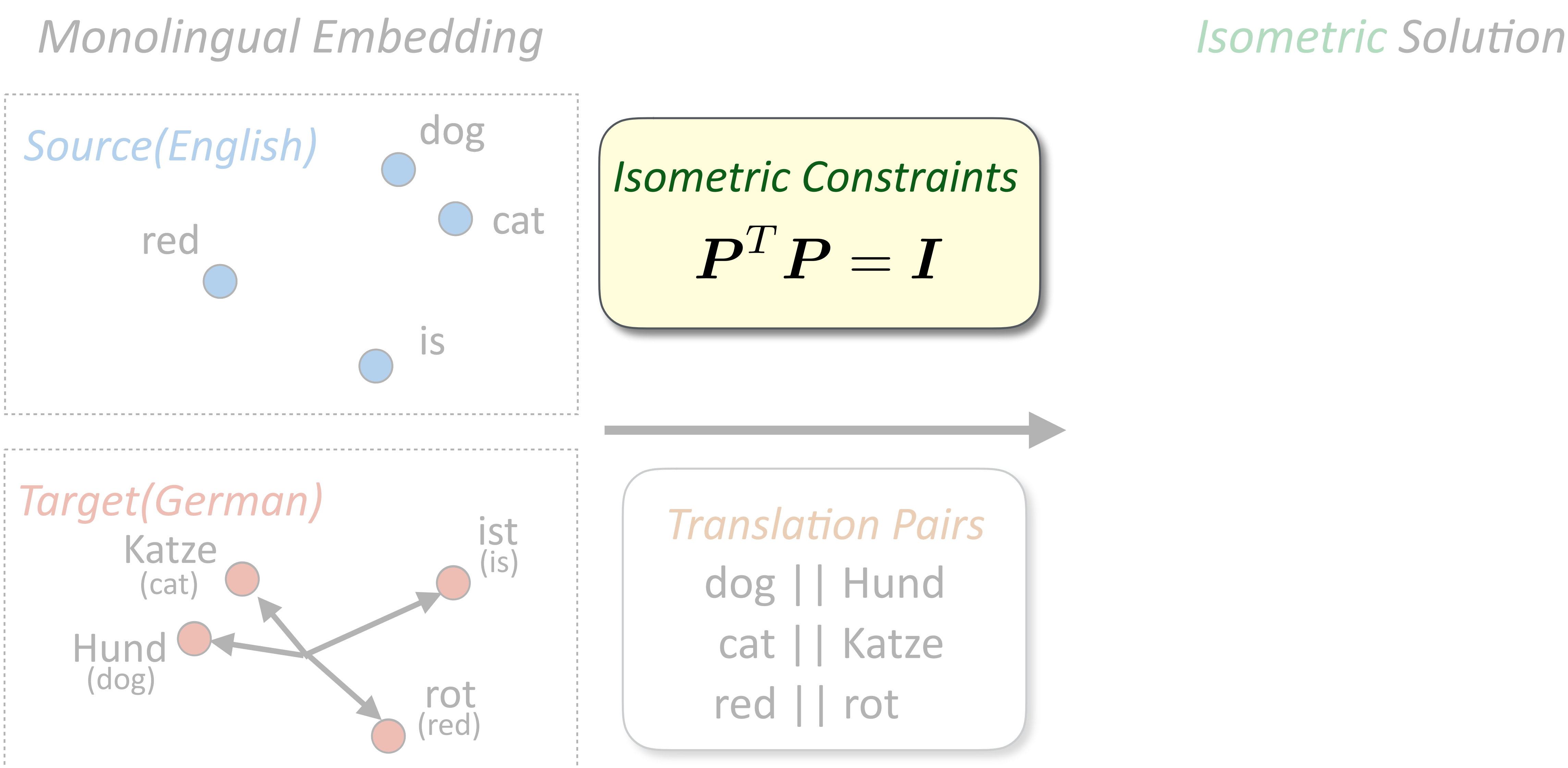
- Transformation P is an isometric (orthonormal) matrix
- Transformation preserves angles and lengths (cosine similarity) of word vectors, thus preserving semantic relations



Our Solution: Isometric Constraints

- Transformation $\boldsymbol{P}$ is an isometric (orthonormal) matrix
- Transformation preserves angles and lengths (cosine similarity) of word vectors, thus preserving semantic relations

$$\cos\langle \text{cat}, \text{dog}\rangle \approx \cos\langle \text{Katze}, \text{Hund}\rangle, \;\; \cos\langle \text{dog}, \text{red}\rangle \approx \cos\langle \text{Hund}, \text{rot}\rangle$$

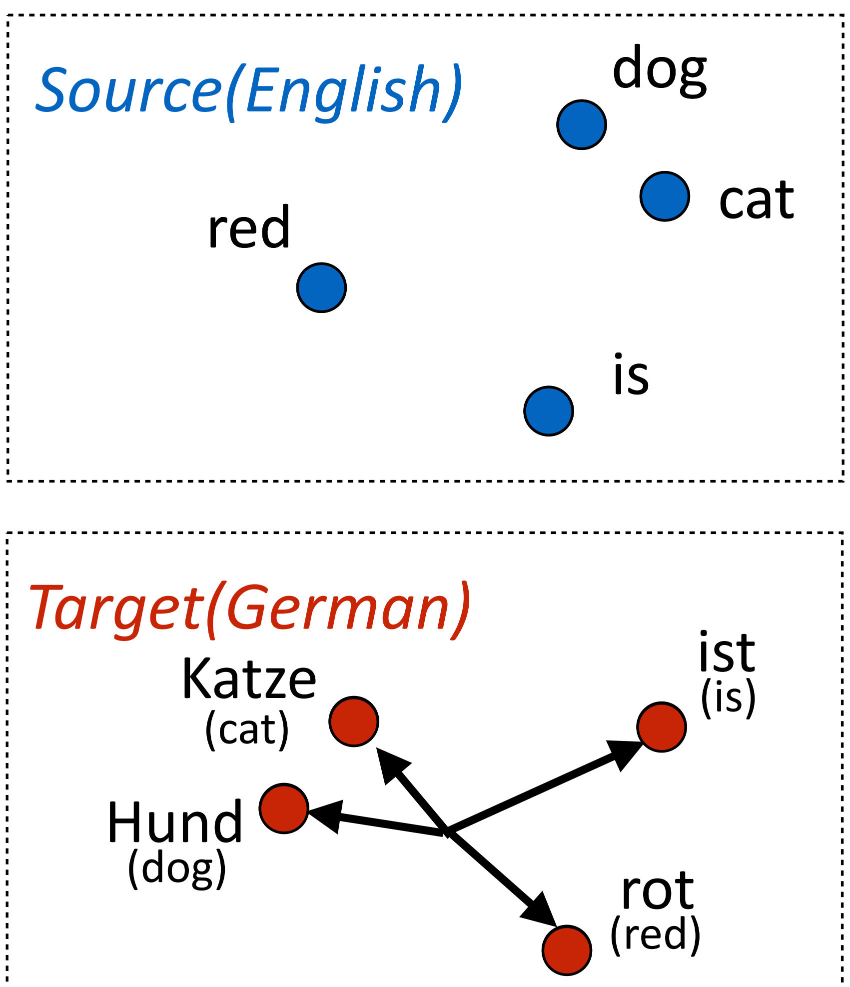


Our Solution: Isometric Constraints

- Transformation P is an isometric (orthonormal) matrix
- Transformation preserves angles and lengths (cosine similarity) of word vectors, thus preserving semantic relations

Monolingual Embedding

Isometric Solution



Isometric Constraints

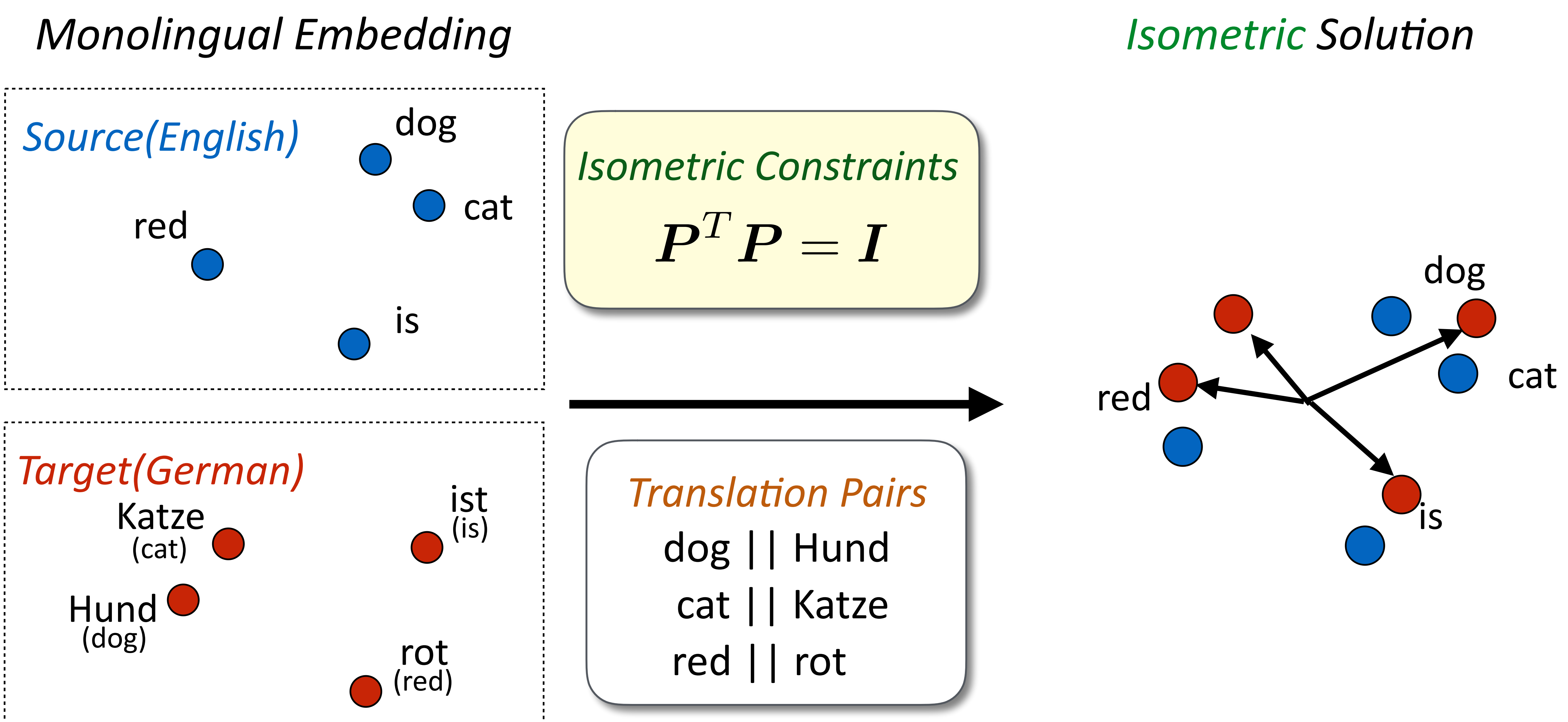
$$P^T P = I$$

Translation Pairs

dog || Hund
cat || Katze
red || rot

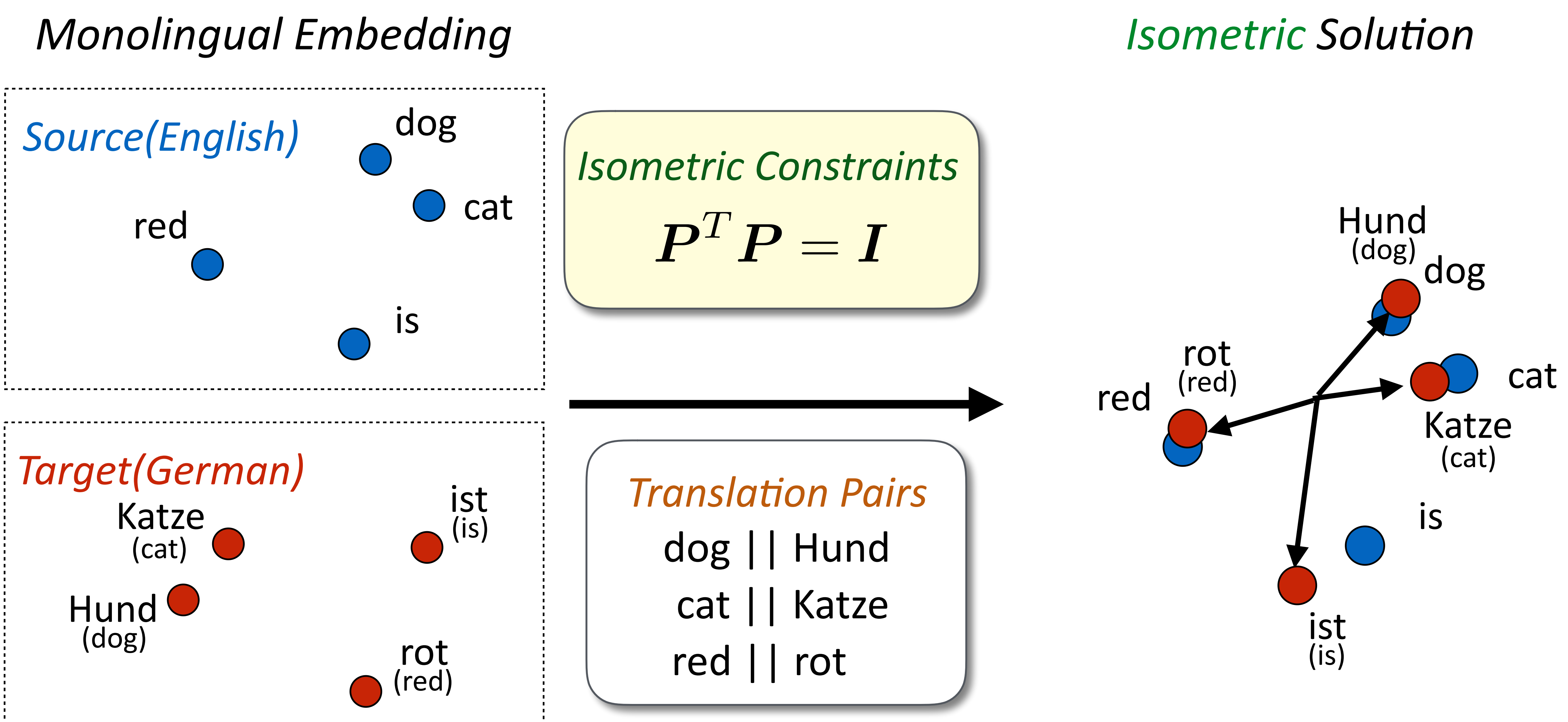
Our Solution: Isometric Constraints

- Transformation P is an isometric (orthonormal) matrix
- Transformation preserves angles and lengths (cosine similarity) of word vectors, thus preserving semantic relations



Our Solution: Isometric Constraints

- Transformation P is an isometric (orthonormal) matrix
- Transformation preserves angles and lengths (cosine similarity) of word vectors, thus preserving semantic relations
- Use the steepest descent algorithm (Abrudan et al., 2008)

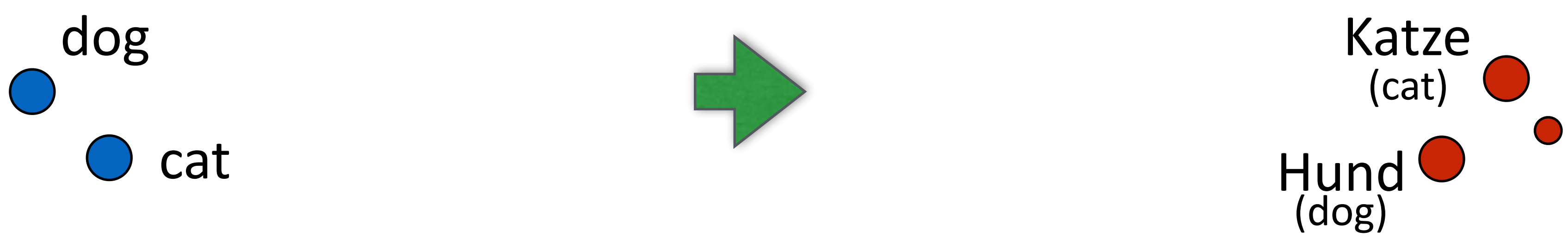


Validation of Isometric Constraints

- Validation for $\cos\langle \text{cat}, \text{dog} \rangle \approx \cos\langle \text{Katze}, \text{Hund} \rangle$
- Verify whether nearest neighbors are preserved after translations

English: nearest neighbor

German: k-th (k≤2) nearest neighbor?



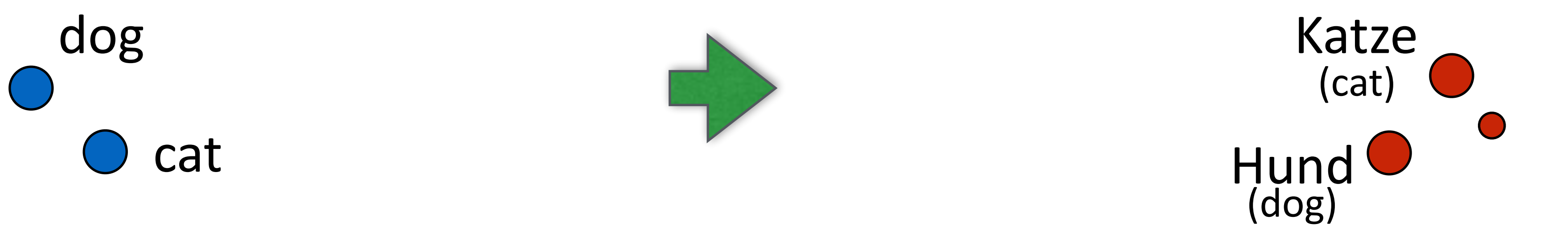
✦ For 50% of word pairs, $k \leq 2$

Validation of Isometric Constraints

- Validation for $\cos\langle\text{cat, dog}\rangle \approx \cos\langle\text{Katze, Hund}\rangle$
- Verify whether nearest neighbors are preserved after translations

English: nearest neighbor

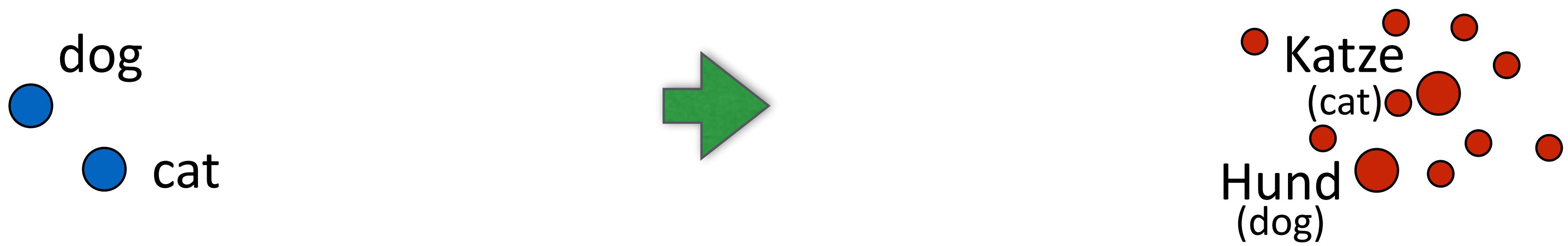
German: k-th (k≤2) nearest neighbor?



✦ For 50% of word pairs, $k \leq 2$

English: nearest neighbor

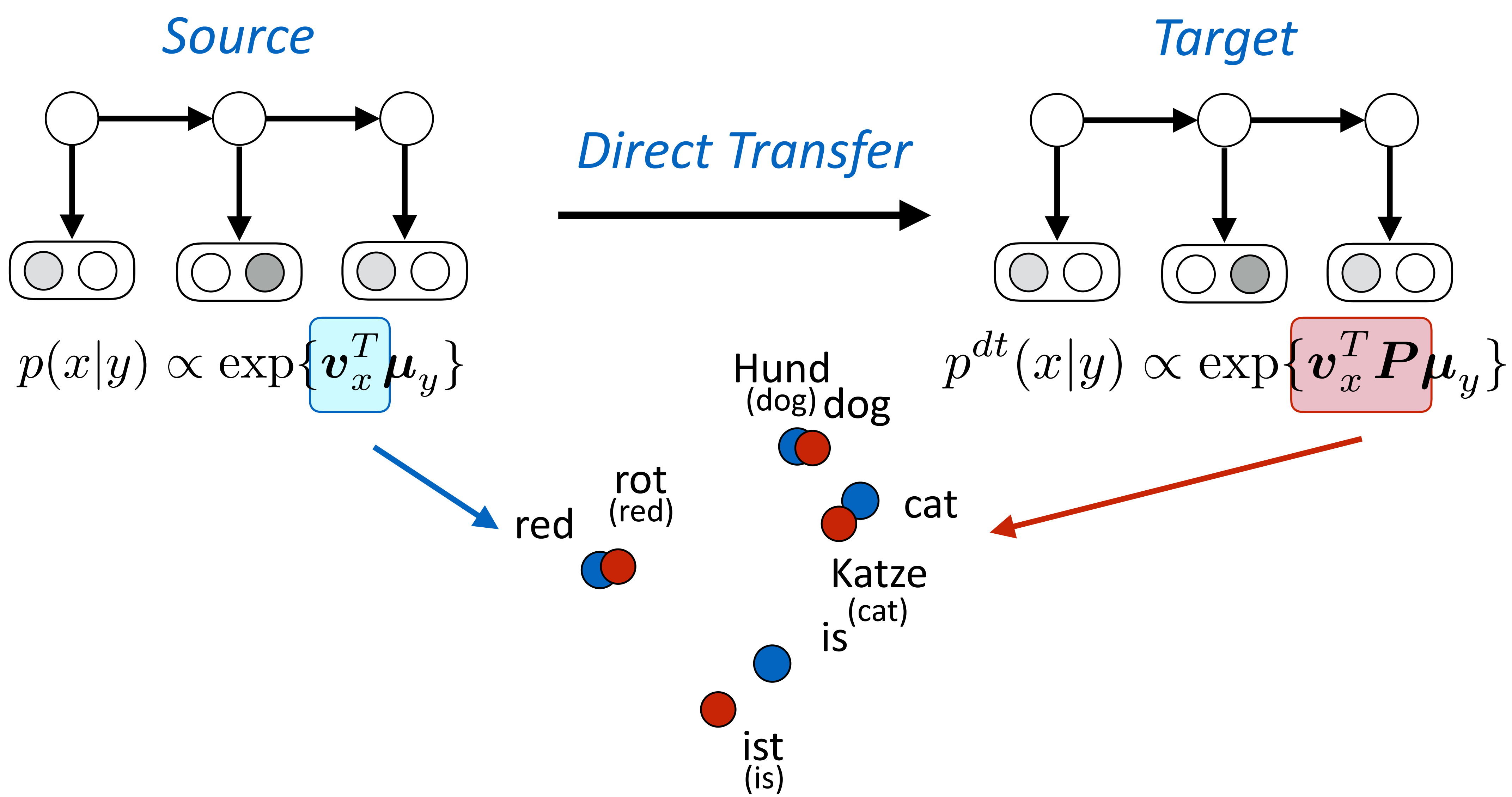
German: k-th (k≤10) nearest neighbor?



✦ For 90% of word pairs, $k \leq 10$

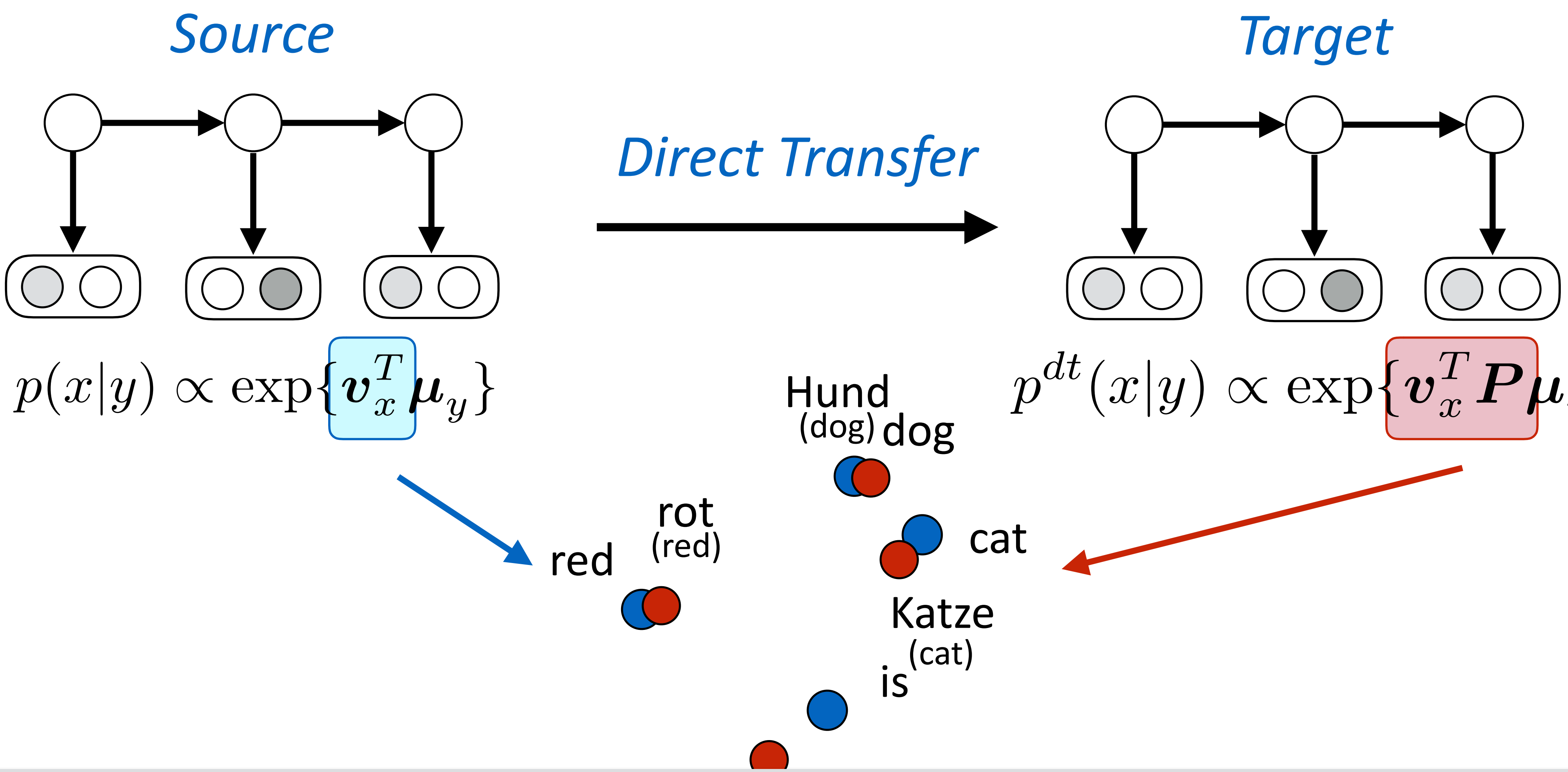
Direct Transfer Model

- Supervised source language HMM
  - ✦ Feature-based HMM (Berg-Kirkpatrick et al., 2010)
  - ✦ Word embeddings as emission features



Direct Transfer Model

- Supervised source language HMM
  - ✦ Feature-based HMM (Berg-Kirkpatrick et al., 2010)
  - ✦ Word embeddings as emission features



$p(x|y) \propto \exp\{\boxed{\boldsymbol{v}_x^T}\boldsymbol{\mu}_y\}$

$p^{dt}(x|y) \propto \exp\{\boxed{\boldsymbol{v}_x^T \boldsymbol{P}}\boldsymbol{\mu}_y\}$

Coarse mapping is not accurate

Our Two-phase Method

1. Learn coarse mapping between embeddings via ten translation pairs

2. Refine embedding transformations and model parameters via unsupervised learning on the target language

Unsupervised Target Language HMM

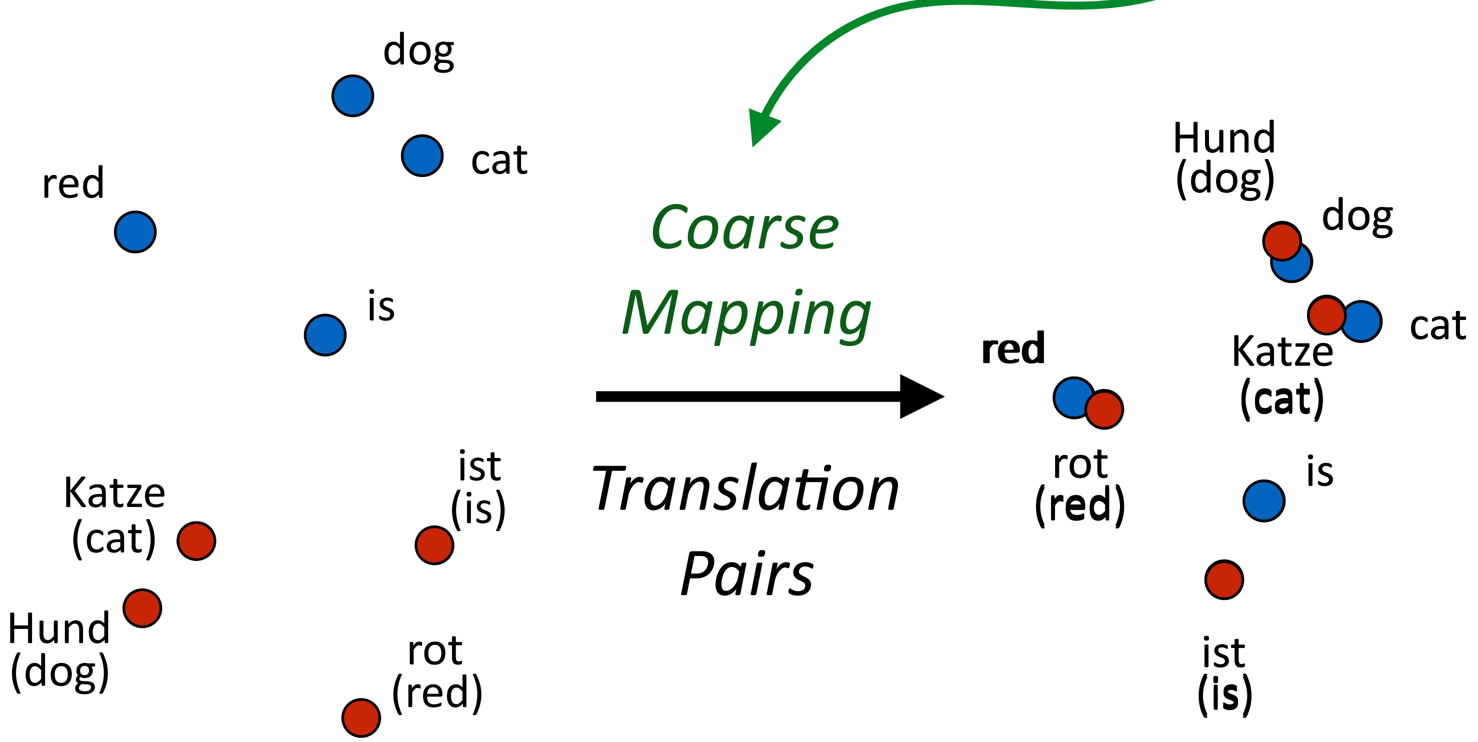
- Use the direct transfer model (based on the coarse mapping) to initialize and regularize the unsupervised tagger on the target language
- Refine mapping via global linear transformation M and local non-linear adjustment $\theta_{x,y}$

$$p(x|y) \propto \exp\{\boldsymbol{v}_x^T \boldsymbol{P}\boldsymbol{M}\boldsymbol{\mu}_y + \theta_{x,y}\}$$

Unsupervised Target Language HMM

- Use the direct transfer model (based on the coarse mapping) to initialize and regularize the unsupervised tagger on the target language
- Refine mapping via global linear transformation M and local non-linear adjustment $\theta_{x,y}$

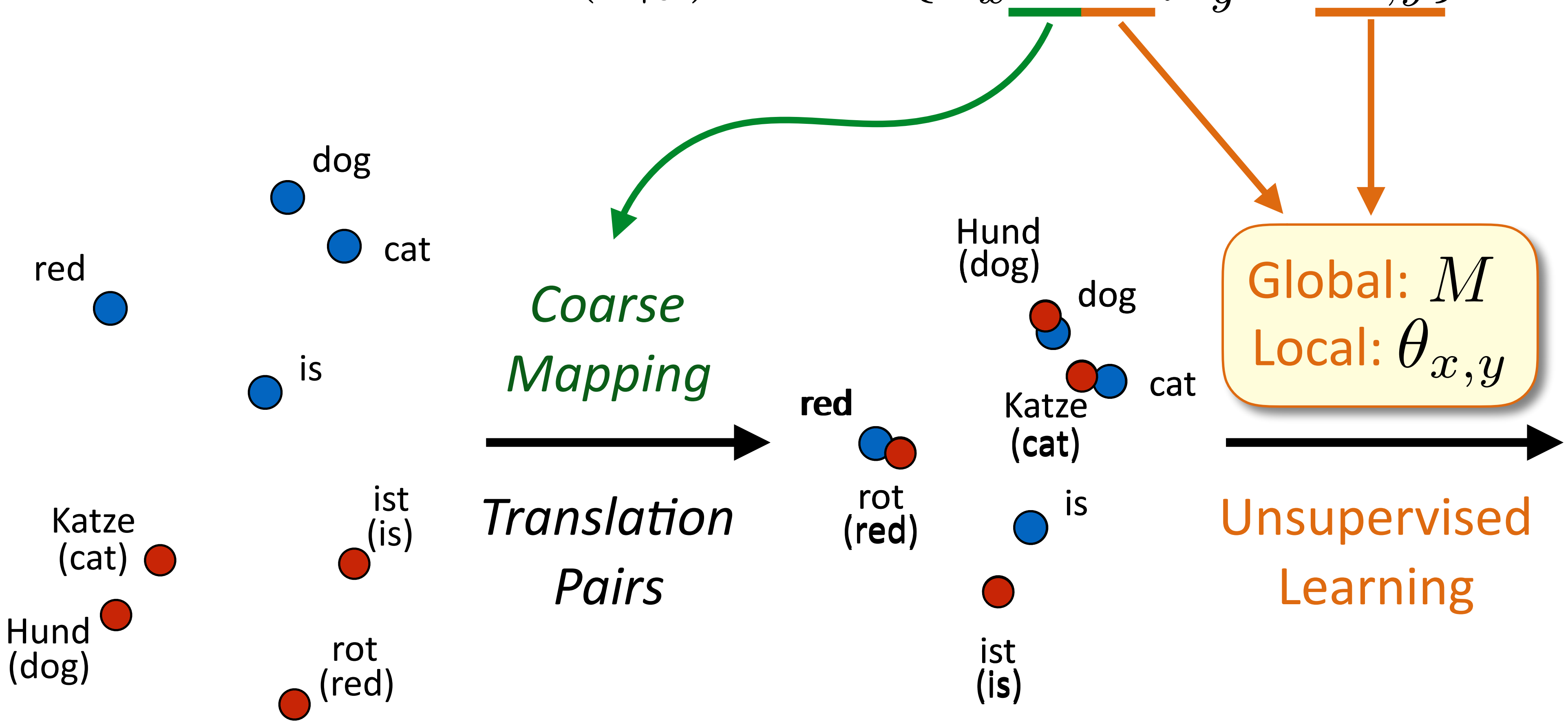
$$p(x|y) \propto \exp\{\boldsymbol{v}_x^T \underline{\boldsymbol{P}} \boldsymbol{M} \boldsymbol{\mu}_y + \theta_{x,y}\}$$



Unsupervised Target Language HMM

- Use the direct transfer model (based on the coarse mapping) to initialize and regularize the unsupervised tagger on the target language
- Refine mapping via global linear transformation M and local non-linear adjustment $\theta_{x,y}$

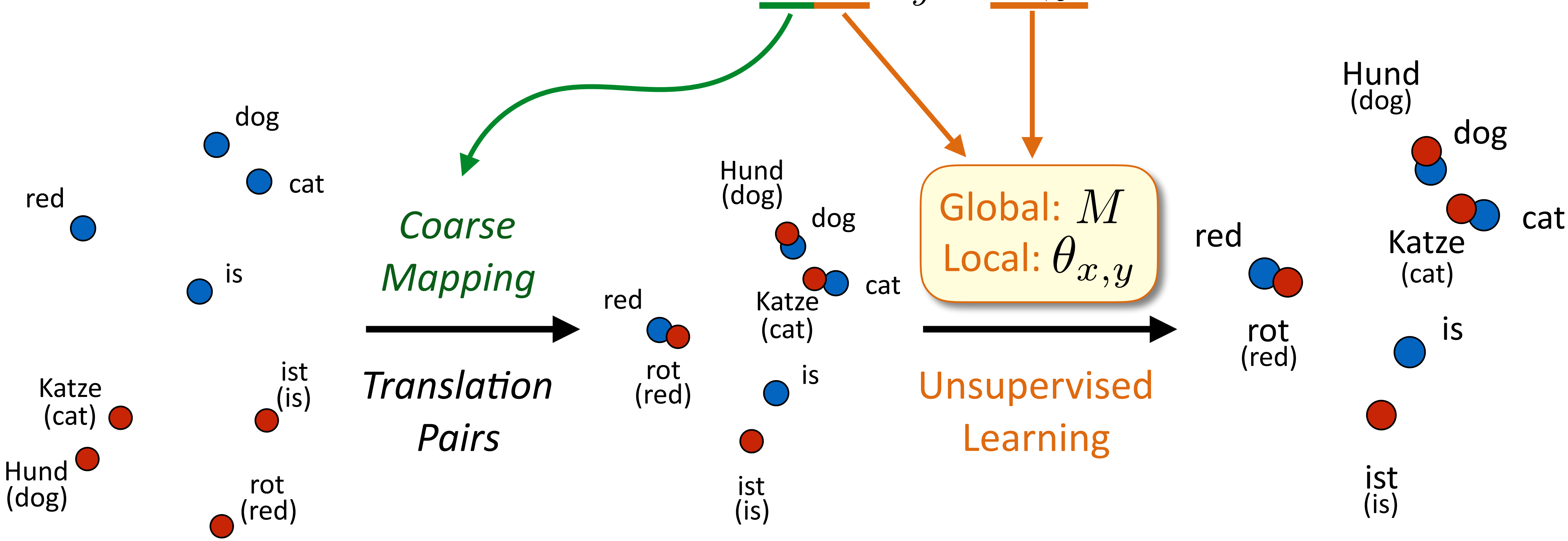
$$p(x|y) \propto \exp\{\boldsymbol{v}_x^T \underline{\boldsymbol{P}\boldsymbol{M}}\boldsymbol{\mu}_y + \underline{\theta_{x,y}}\}$$



Unsupervised Target Language HMM

- Use the direct transfer model (based on the coarse mapping) to initialize and regularize the unsupervised tagger on the target language
- Refine mapping via global linear transformation M and local non-linear adjustment $\theta_{x,y}$

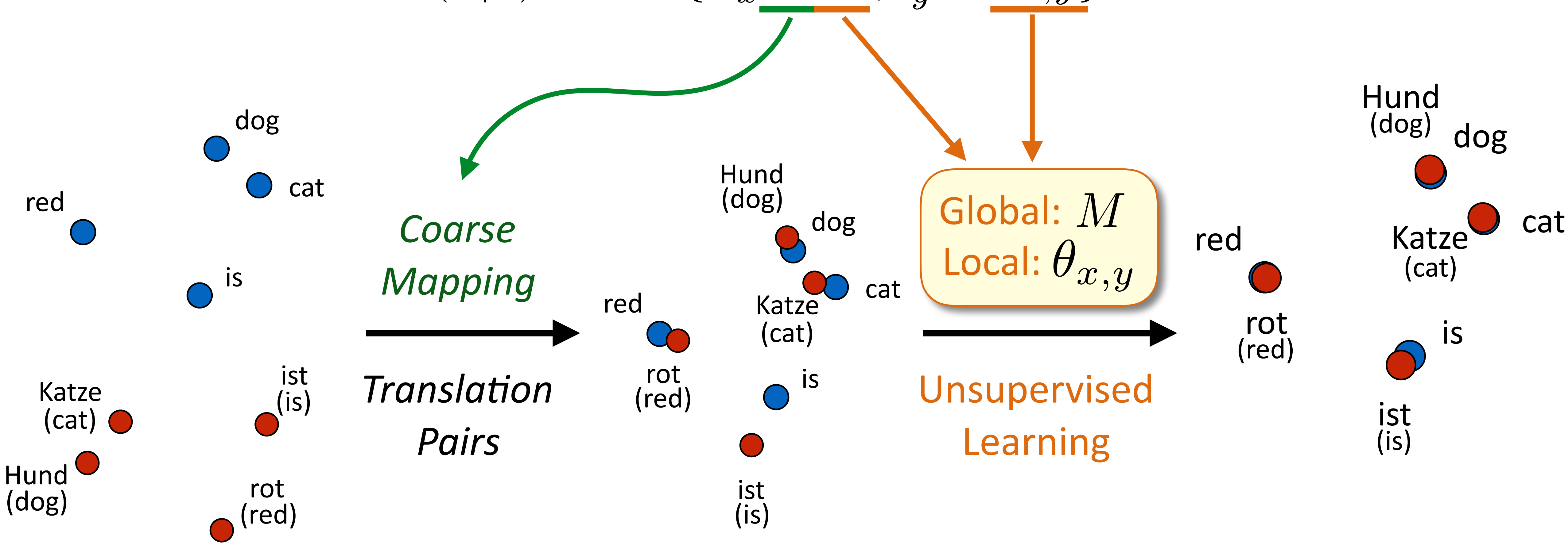
$$p(x|y) \propto \exp\{\boldsymbol{v}_x^T \boldsymbol{P}\boldsymbol{M}\boldsymbol{\mu}_y + \theta_{x,y}\}$$



Unsupervised Target Language HMM

- Use the direct transfer model (based on the coarse mapping) to initialize and regularize the unsupervised tagger on the target language
- Refine mapping via global linear transformation M and local non-linear adjustment $\theta_{x,y}$

$$p(x|y) \propto \exp\{\boldsymbol{v}_x^T \boldsymbol{P}\boldsymbol{M}\boldsymbol{\mu}_y + \theta_{x,y}\}$$



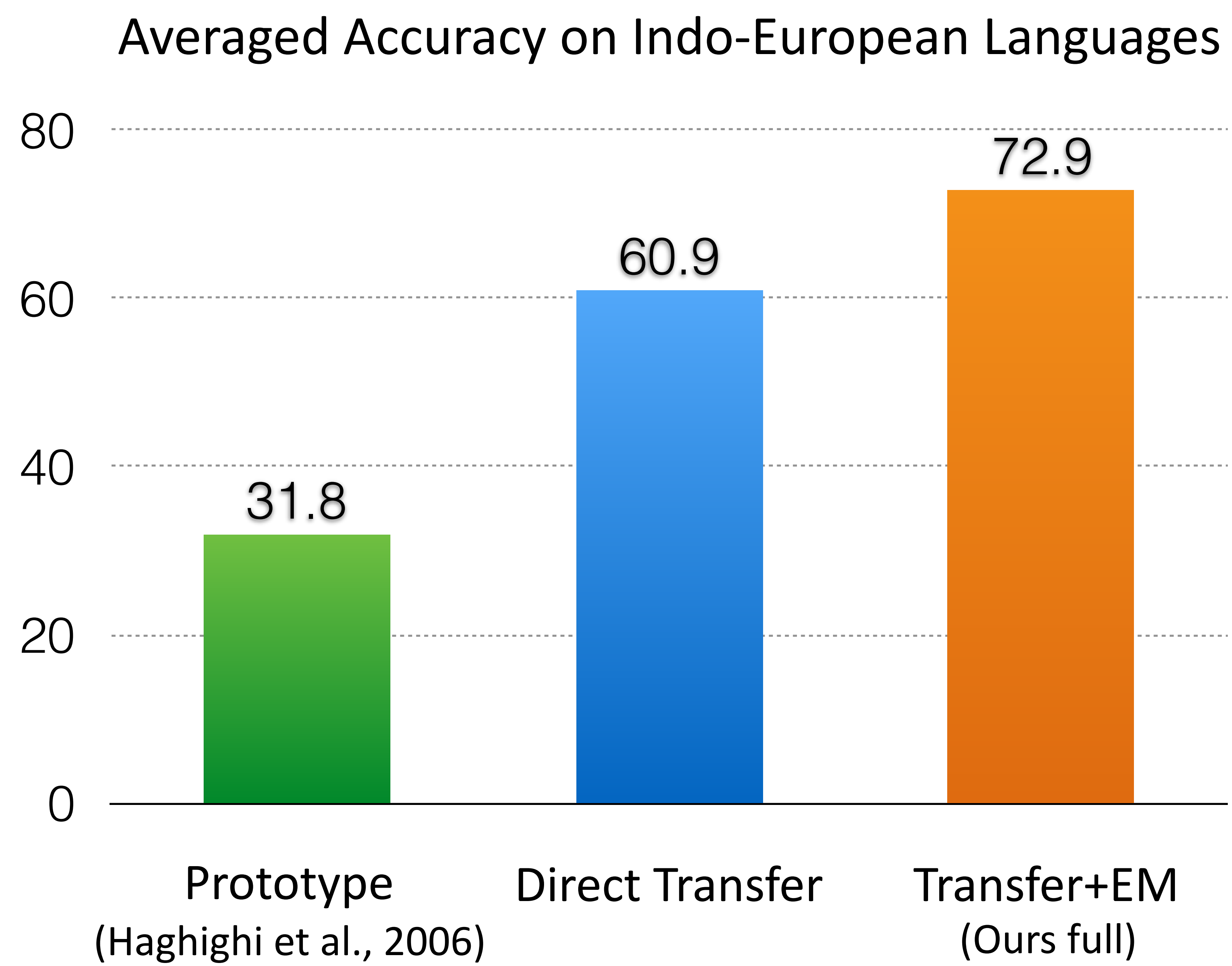
Learning

- Parameters: $\boldsymbol{\mu}_y, \theta_{y,y'}, \boldsymbol{M}, \theta_{x,y}$
- Optimization method: standard Expectation-Maximization (EM)
  - ✦ E-step: forward-backward
  - ✦ M-step: gradient ascent using L-BFGS

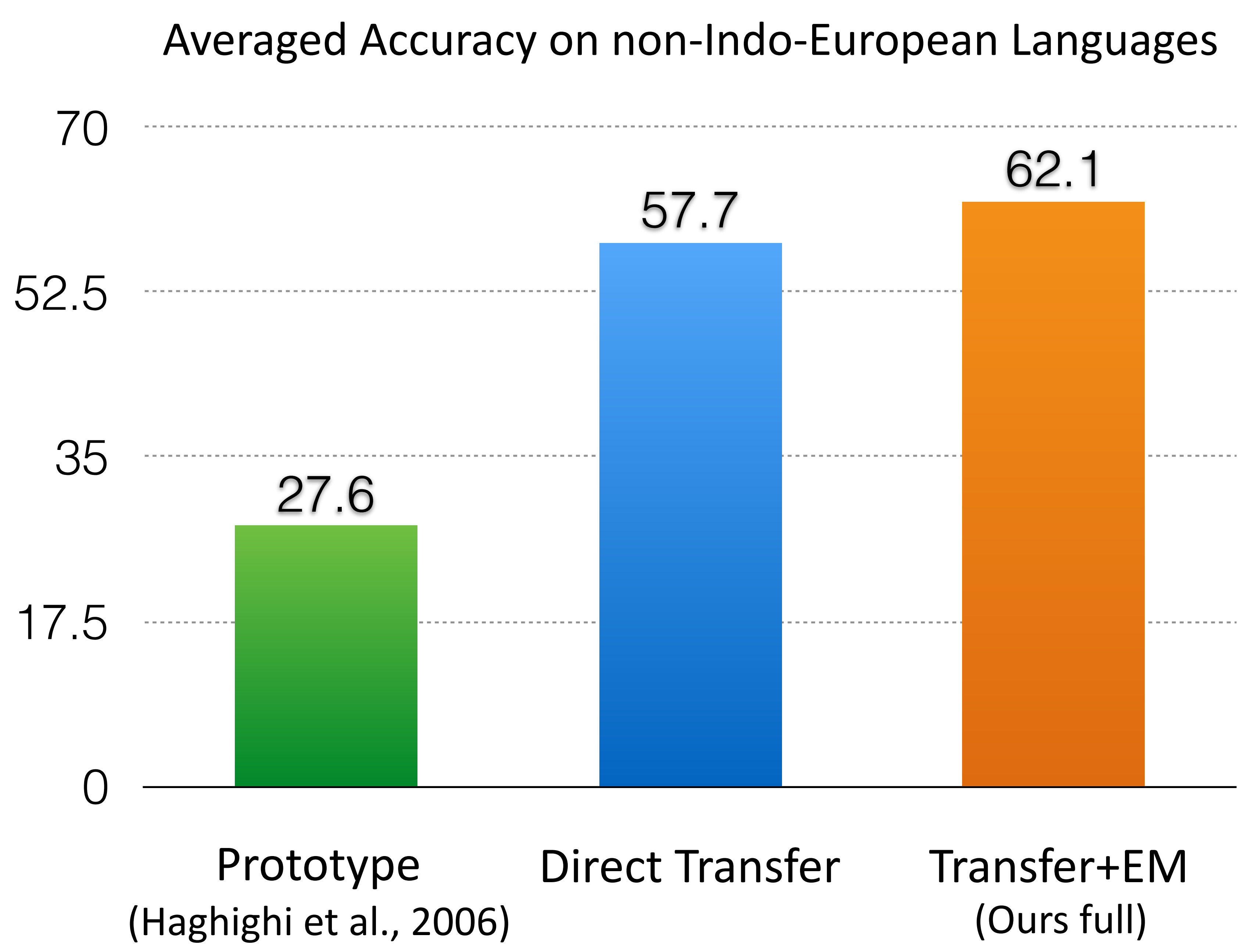
Experimental Setup

- Datasets: Universal Dependencies Treebanks v1.2
  - Source: English
  - Target (Indo-European): Danish (da), German (de), Spanish (es)
  - Target (non-Indo-European): Finnish (fi), Hungarian (hu), Indonesian (id)
- Universal tagset: 14 tags (noun, verb, adjective etc.)
- Word embeddings: 20-dimension vectors trained on Wiki dumps using word2vec

Indo-European Results

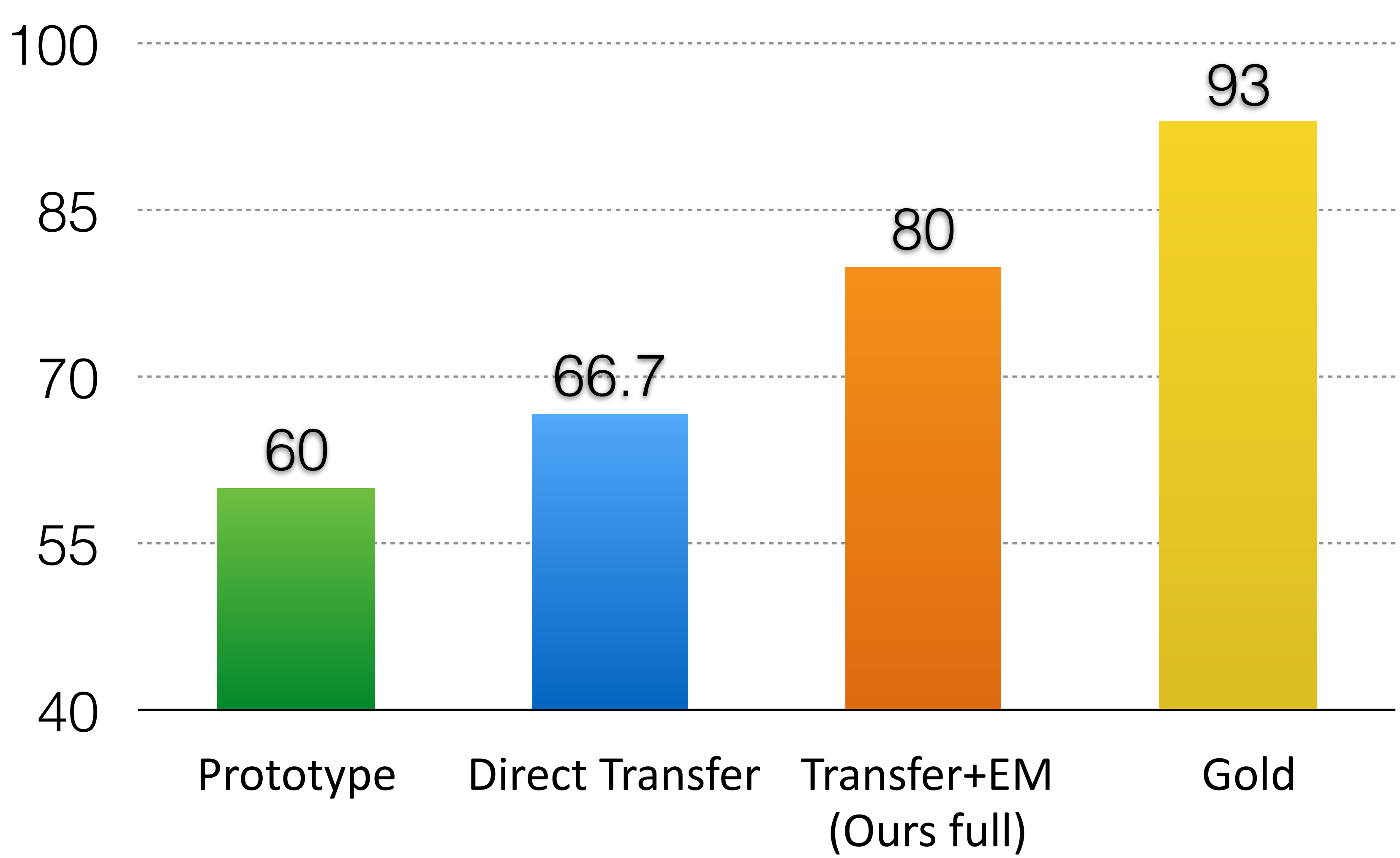


Non-Indo-European Results



Prediction of Linguistic Typology

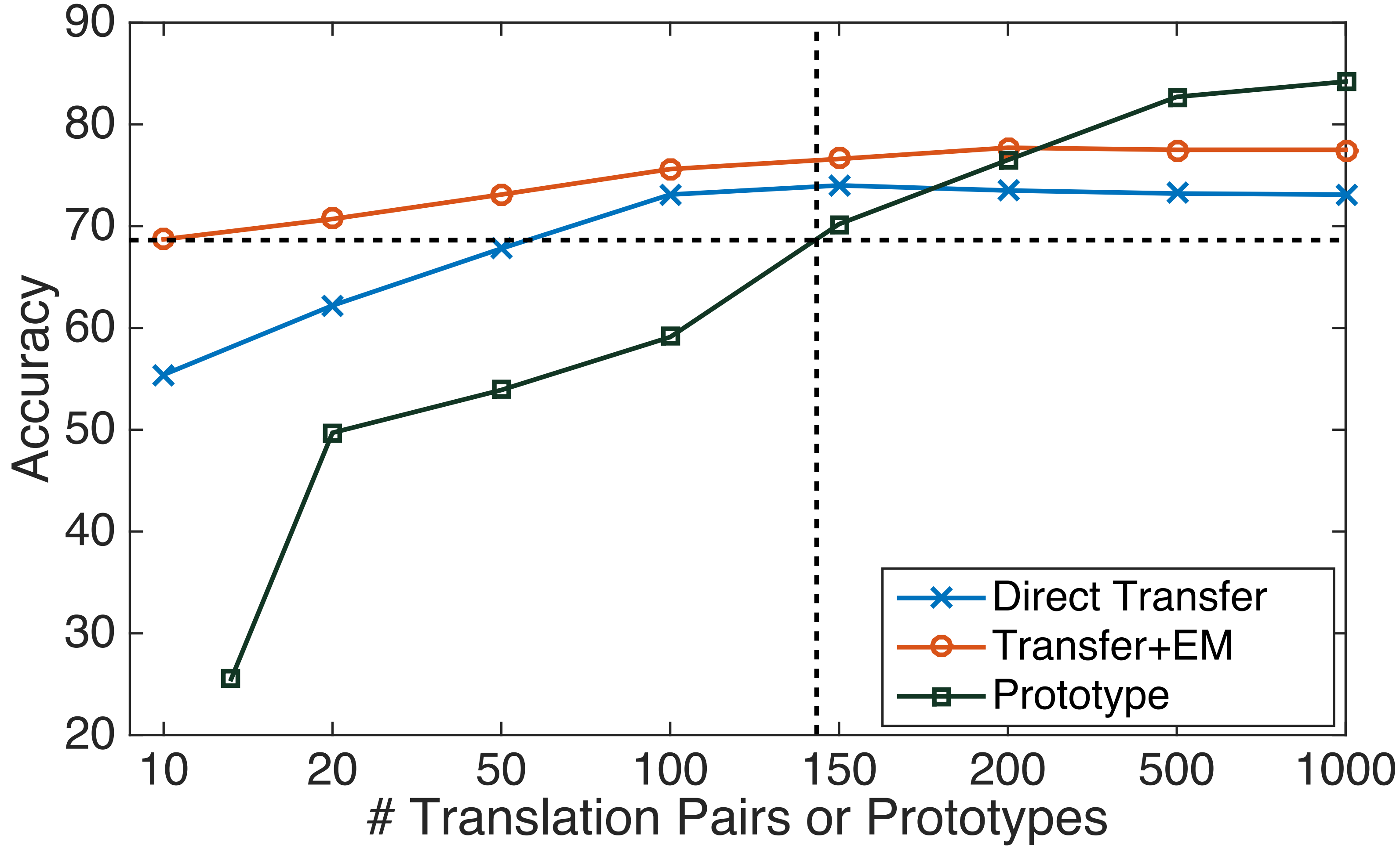
- Task: predict whether a language is verb-object or object-verb (five typological properties)
- Features: bigrams and trigrams of POS tags



Impact of Amount of Supervision

- Transfer+EM with 10 pairs = 150 prototypes

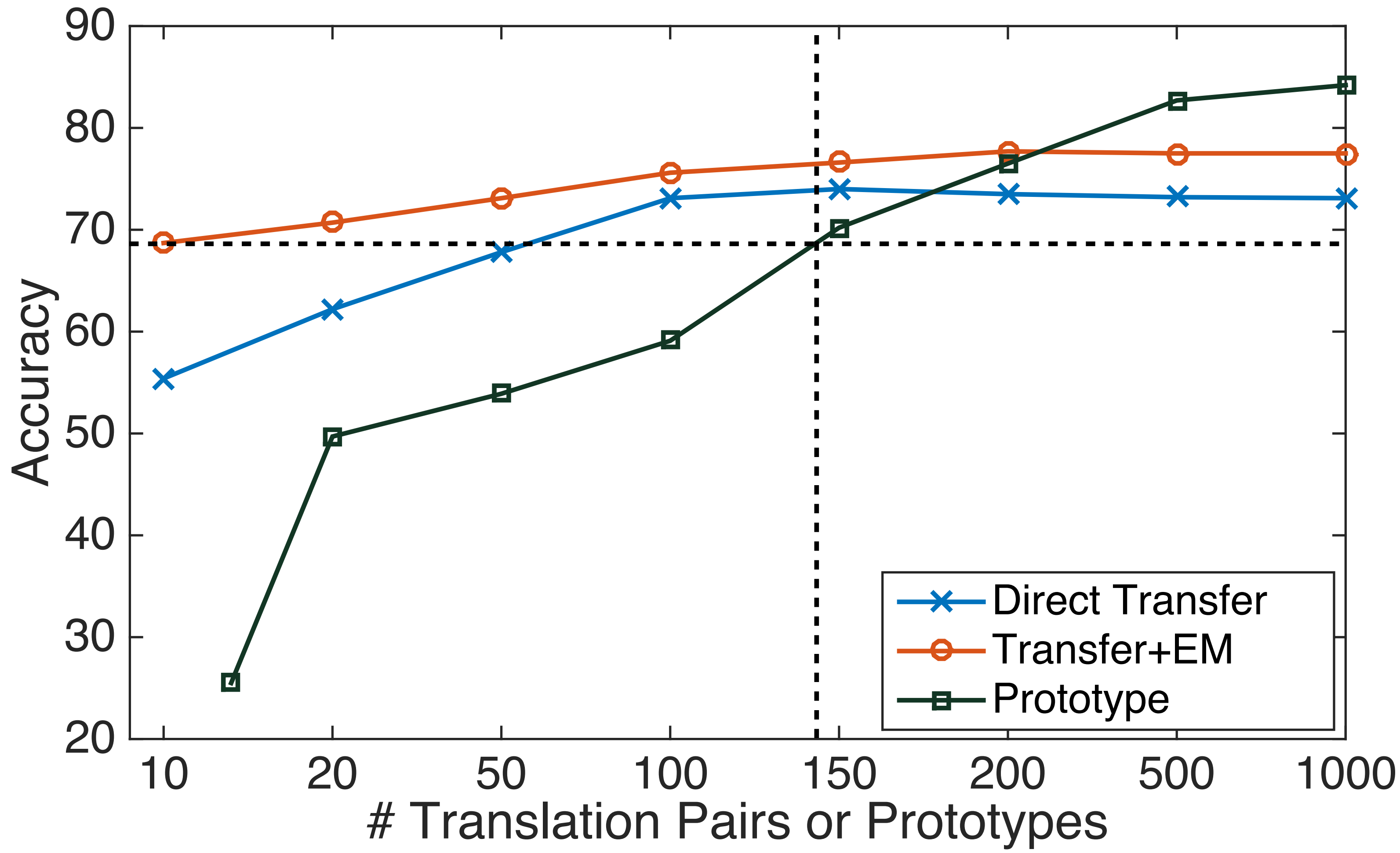
Accuracy on German



Impact of Amount of Supervision

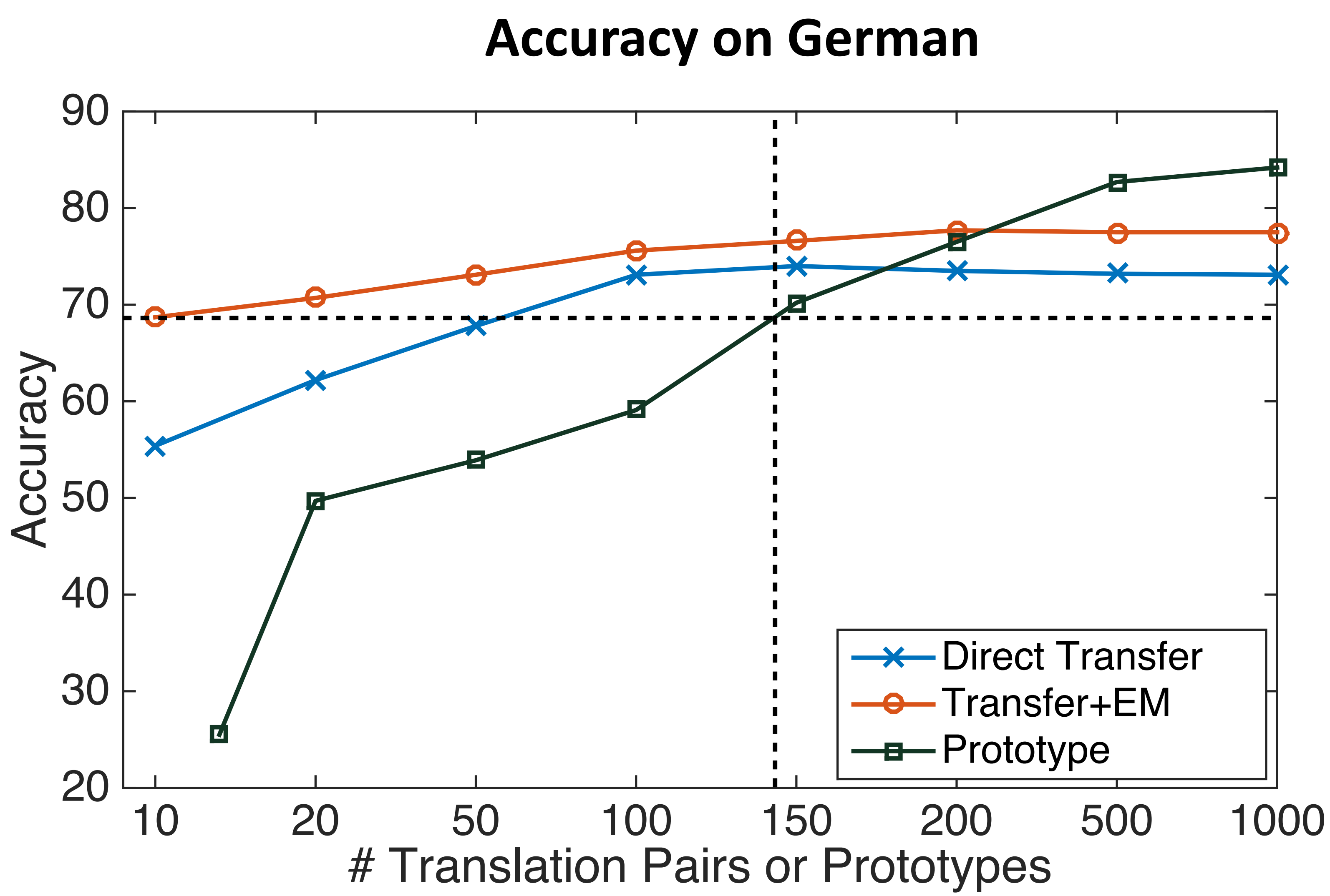
- Transfer+EM with 10 pairs = 150 prototypes
- Prototype improves with large amount of annotations

Accuracy on German



Impact of Amount of Supervision

- Transfer+EM with 10 pairs = 150 prototypes
- Prototype improves with large amount of annotations
- Transfer+EM consistently improves over Direct Transfer



Conclusion

- Ten translation pairs are sufficient to enable multilingual transfer of POS tagging
- Our model significantly outperforms the direct transfer and the prototype-driven method

Source code available at:

https://github.com/yuanzh/transfer_pos

Thank You!

Impact of Embedding Dimensions and Window Size