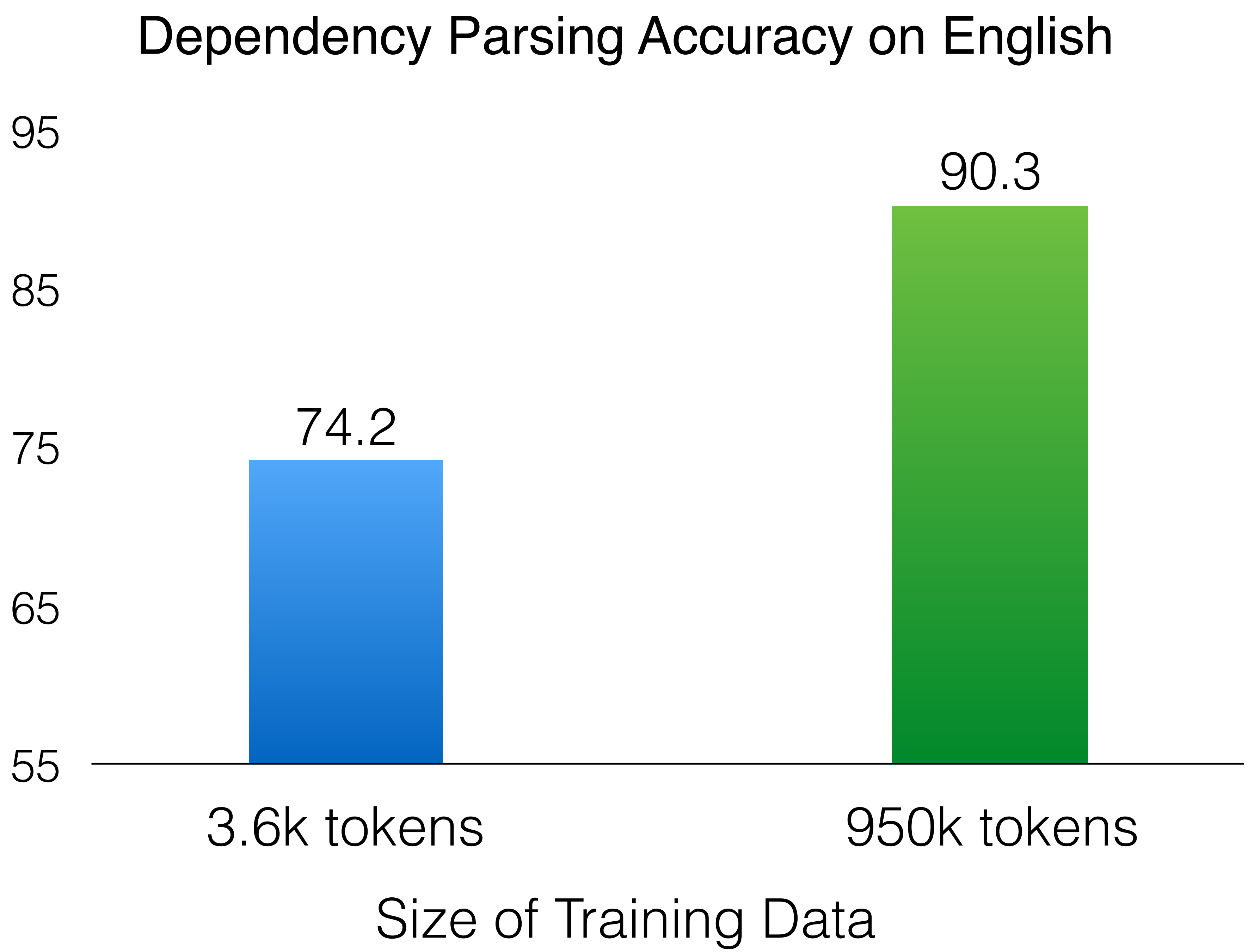
Transfer Learning for Low-resource Natural Language Analysis

Yuan Zhang

January 30, 2017

Low-resource Problem

- Top-performing systems need large amounts of annotated data



Low-resource Scenarios

Low-resource Languages:

Malagasy annotations
~1,000 tokens

English annotations
> 1 million tokens

Low-resource Scenarios

Low-resource Languages:

Malagasy annotations
~1,000 tokens

English annotations
> 1 million tokens

Low-resource Domains:

PubMed
Medical: ~ 500 sentences

THE WALL STREET JOURNAL.
WSJ
News articles:
> 100k sentences

Our Work: Transfer Learning

- Use rich resources in related source tasks to improve target performance

Target:

resource-poor

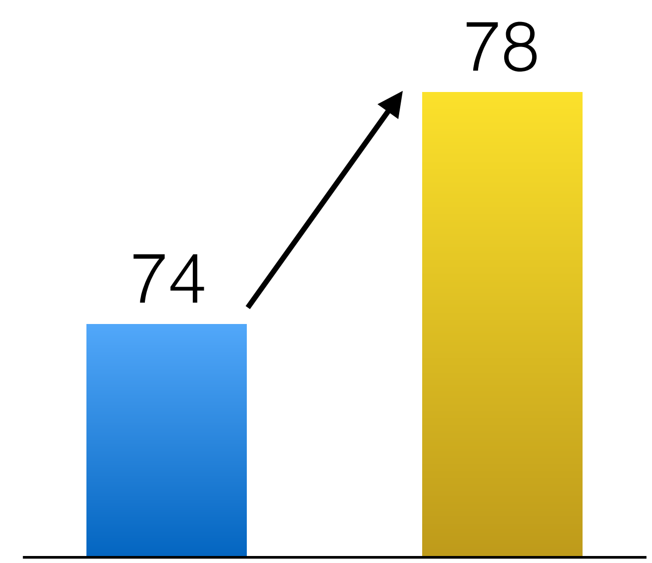
low accuracy

Source: resource-rich

Similar Languages

Related Domains

Improvement via transfer learning



Challenges in Transfer: Multilingual

- Part-of-speech (POS) tagging: different vocabulary

Source: English

DET	ADJ	NOUN
a	red	apple

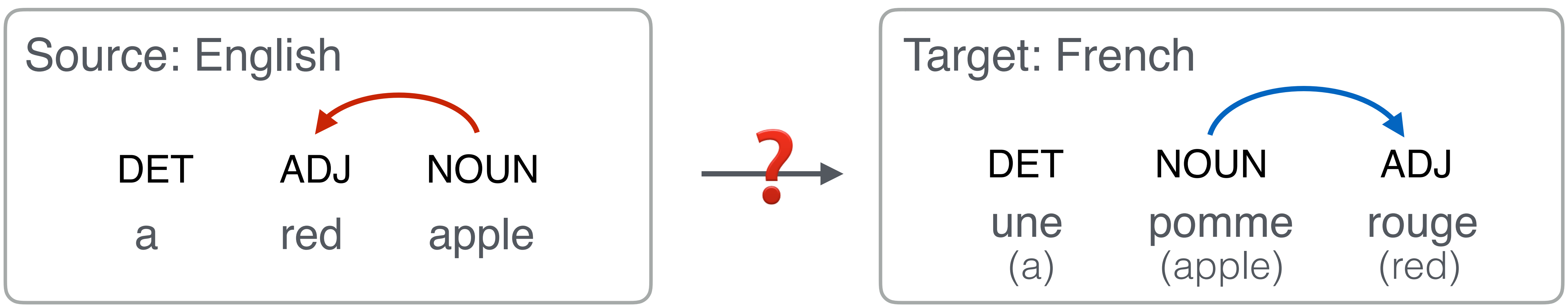
→

Target: French

DET	NOUN	ADJ
une (a)	pomme (apple)	rouge (red)

Challenges in Transfer: Multilingual

- Part-of-speech (POS) tagging: different vocabulary
- Dependency parsing: different word ordering



Challenges in Transfer: Monolingual

- Domain transfer: different writing-style

Source: Restaurant reviews

Target: Hotel reviews

The fries were **undercooked** → ? → The room **rained water from above**

Challenges in Transfer: Monolingual

- Domain transfer: different writing-style

Source: Restaurant reviews

Target: Hotel reviews

The fries were **undercooked**

The room **rained water from above**

- Aspect transfer: different aspects in the same domain

FINAL DIAGNOSIS: BREAST (LEFT) … **INVASIVE DUCTAL CARCINOMA (IDC) Tumor size: num x num x num cm Grade: 3. Lymphatic vessel invasion (LVI): Not identified.** Blood vessel invasion: Suspicious. Margin of invasive carcinoma …

Source Aspect: IDC

Target Aspect: LVI

General Setup: Low-resource Transfer

- No annotations for the target task

	Source	Target
Labeled	✓	✗
Unlabeled	✓	✓

- No parallel data, or a few word translation pairs
- Low level of human effort
  - ✦ Existing external resources
  - ✦ No feature engineering

General Setup: Low-resource Transfer

- No annotations for the target task

	Source	Target
Labeled	✓	✗
Unlabeled	✓	✓

- No parallel data, or a few word translation pairs
- Low level of human effort
  - Existing external resources
  - No feature engineering

Contribution: Improve low-resource transfer in multilingual and monolingual scenarios

Our Approach

Multilingual Transfer:

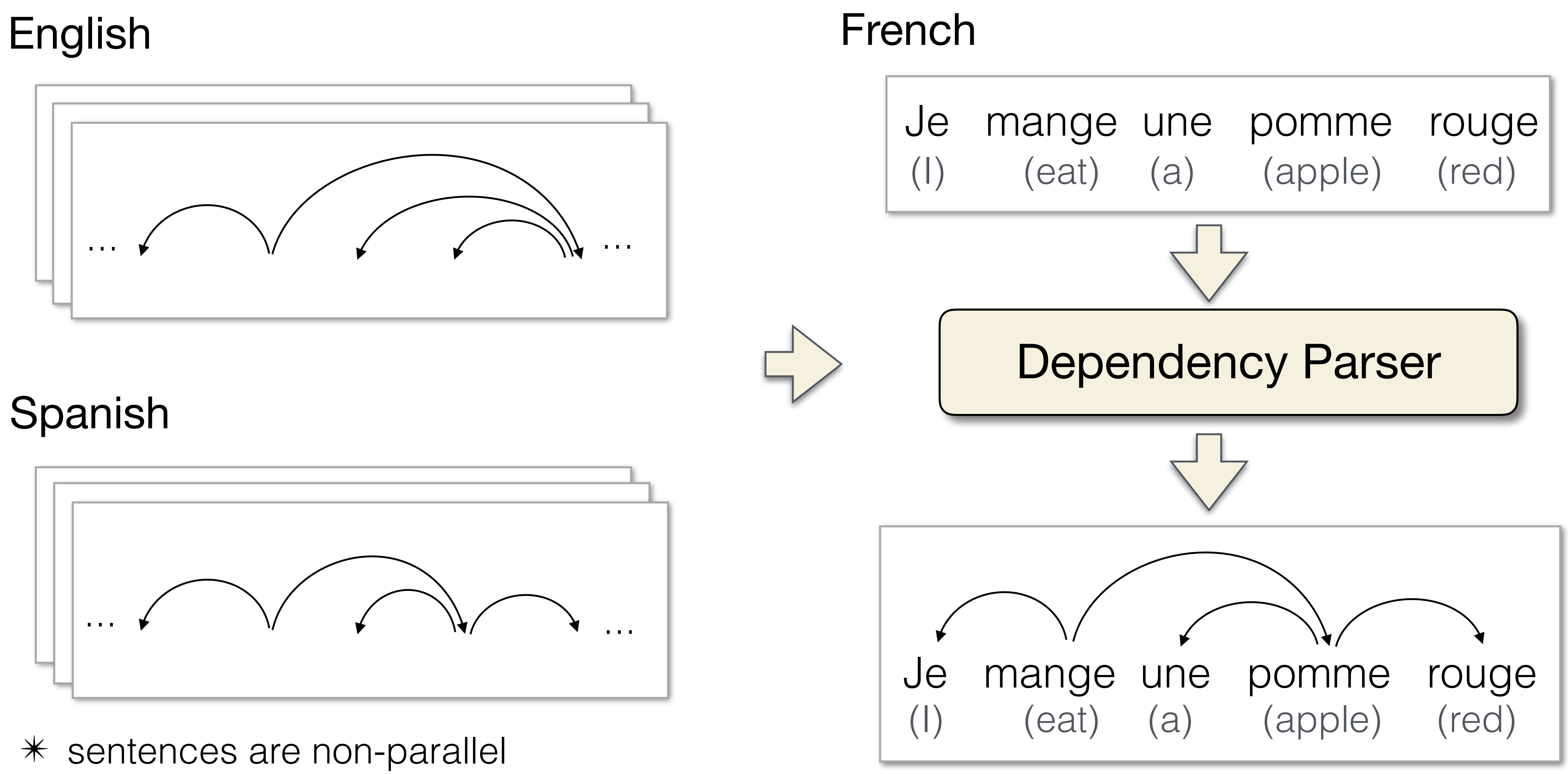
- Hierarchical tensors for dependency parsing
  - *Prior knowledge incorporation without feature engineering*
- Multilingual embeddings for POS tagging

Monolingual Transfer:

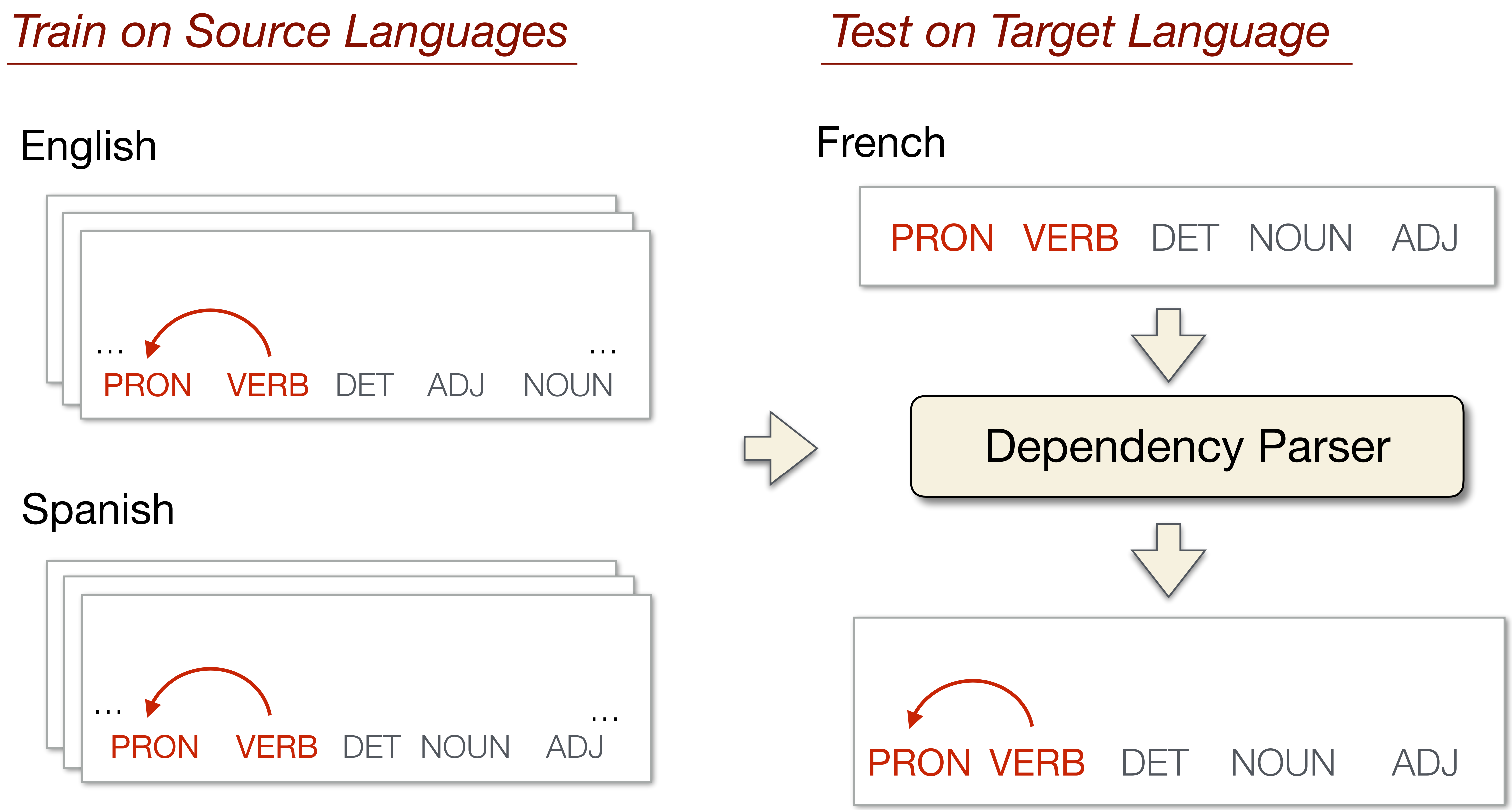
- Adversarial networks for aspect transfer

Multilingual Transfer for Dependency Parsing

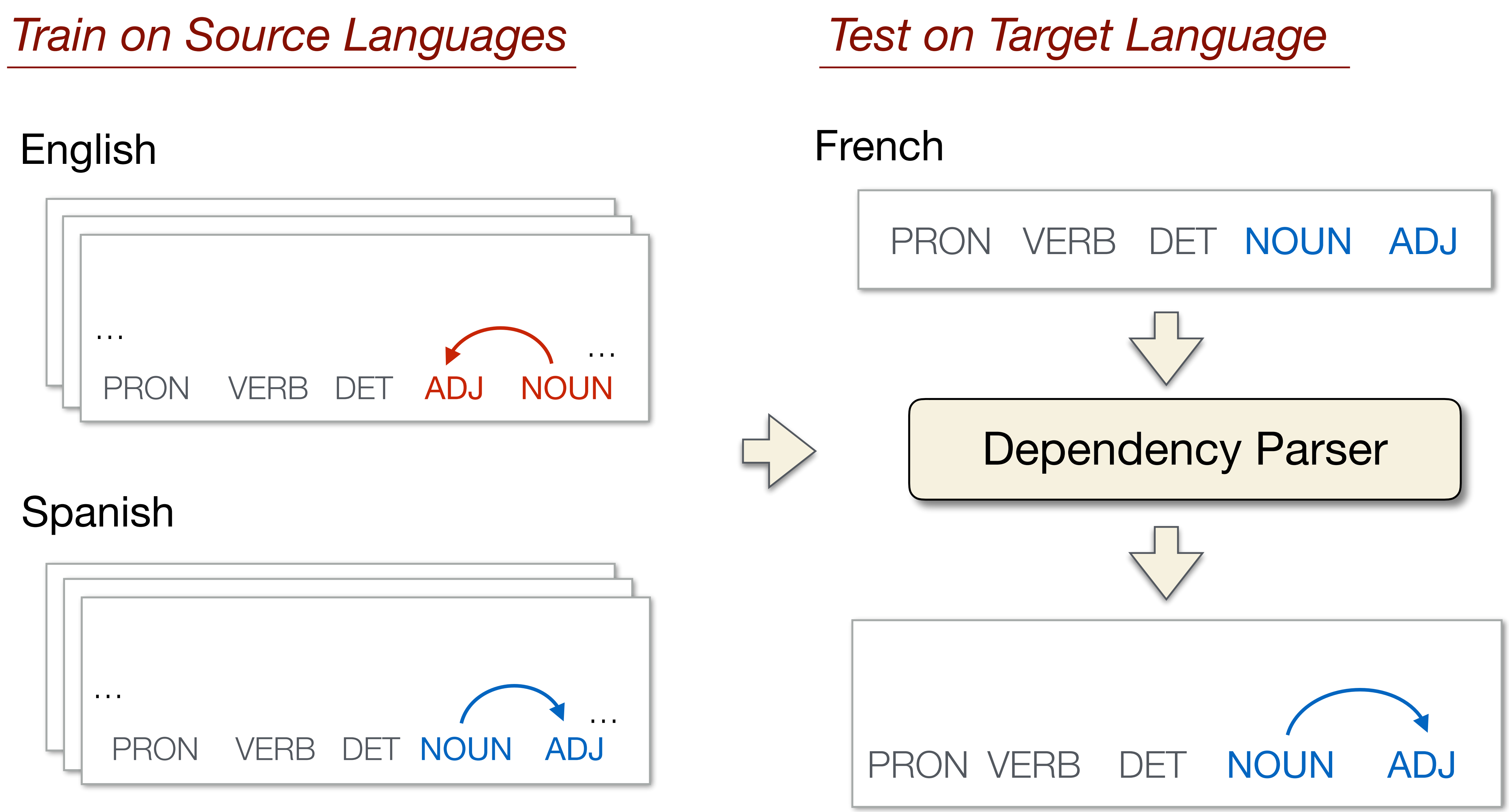
Train on Source Languages | *Test on Target Language*



Non-lexical Transfer via Universal POS



Challenge: Different Word Ordering



Solution: Linguistic Typology

- Form of typological features

Typological Feature	English	French
87A: Order of Noun and Adjective	ADJ-NOUN	NOUN-ADJ

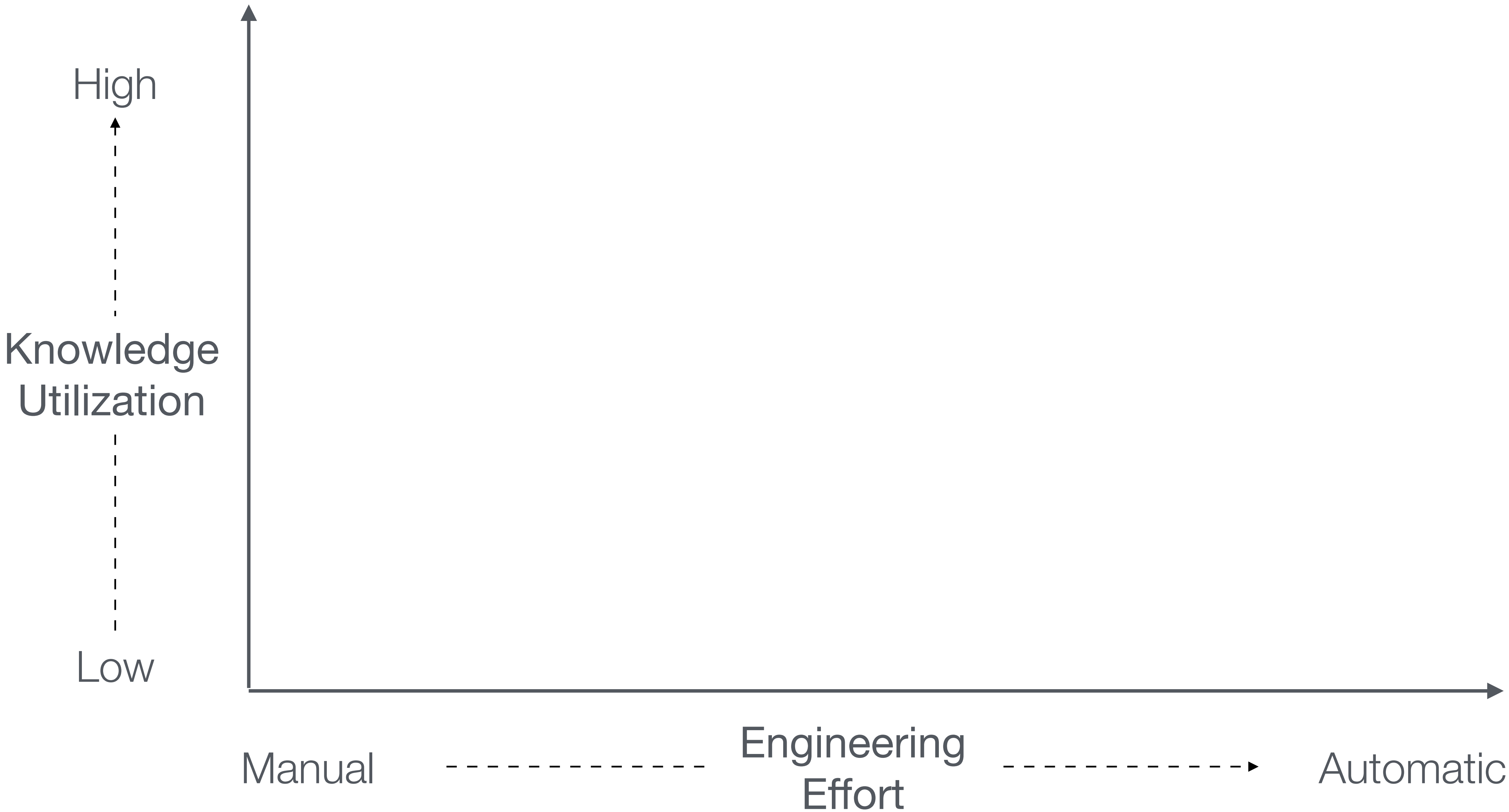
- Idea of selective transfer

English: 87A=ADJ-NOUN

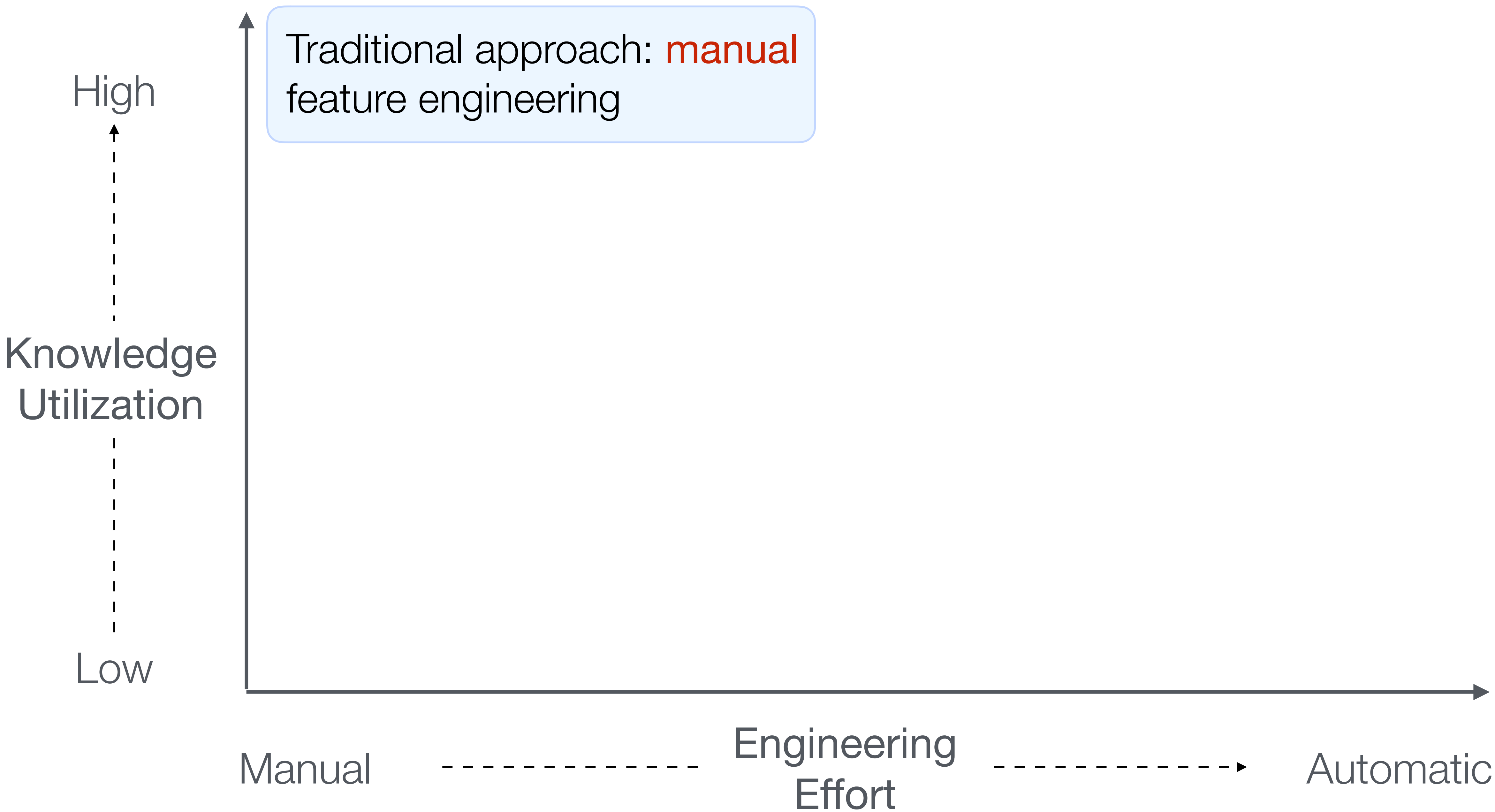
French: 87A=NOUN-ADJ

Spanish: 87A=NOUN-ADJ

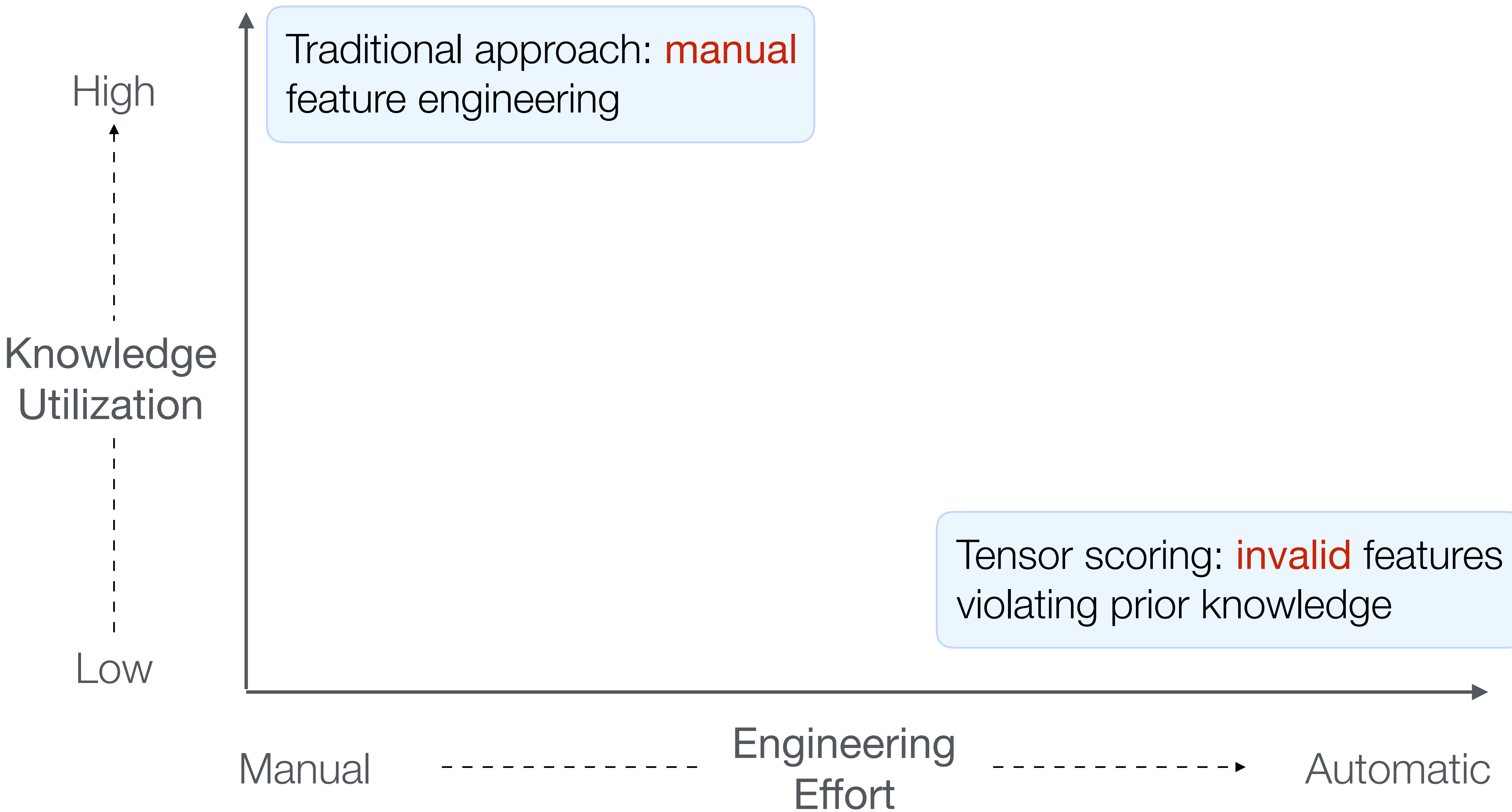
Utilizing Typology Knowledge



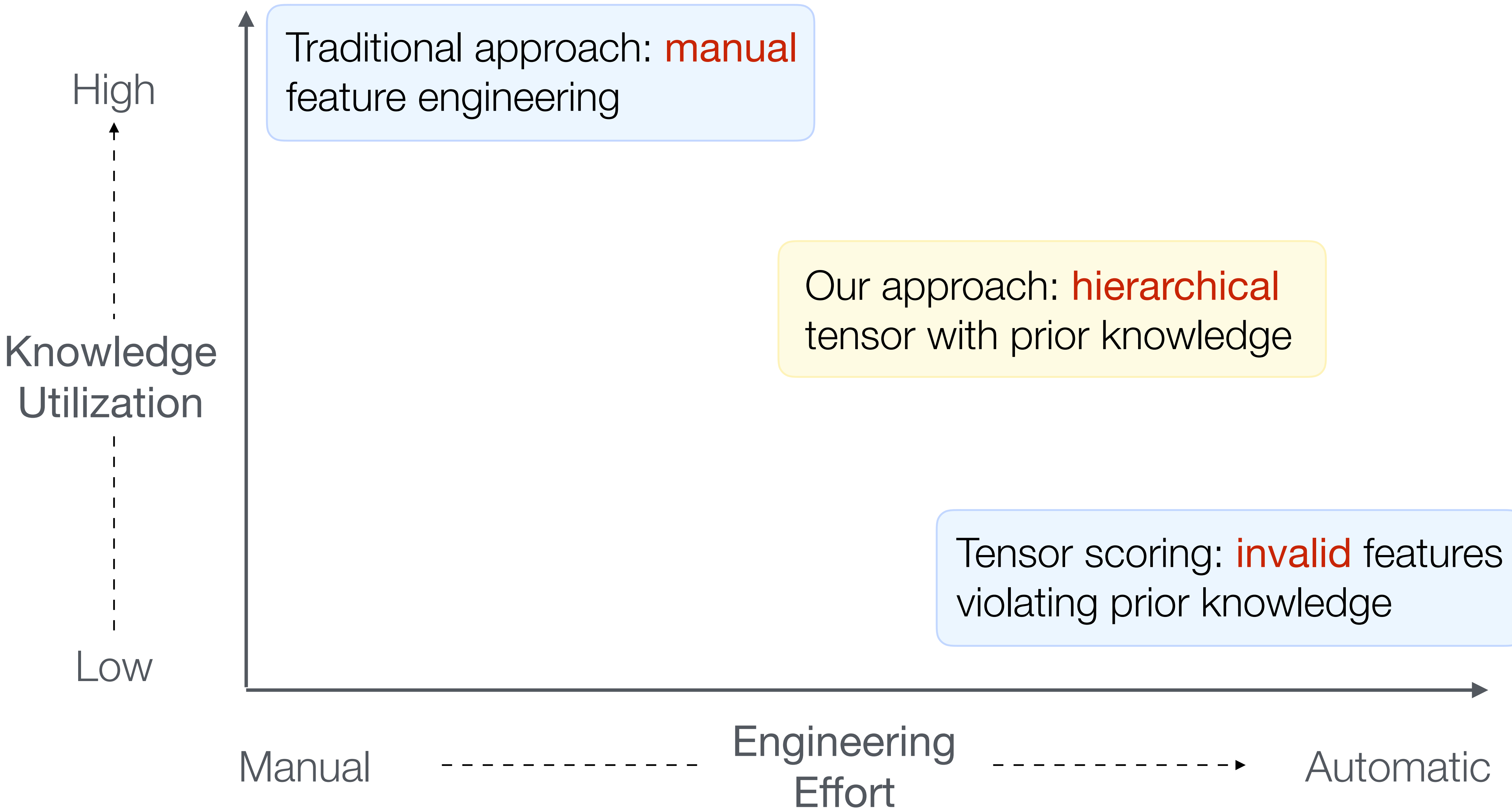
Utilizing Typology Knowledge



Utilizing Typology Knowledge



Utilizing Typology Knowledge



Traditional Approach: Feature Engineering

- Manually conjoin standard parsing features with typological features (Täckström et al., 2013)

$$f_{100}(\cdot) = \mathbb{I}\{\text{head POS=NOUN, modifier POS=ADJ, direction=Right, 87A=NOUN-ADJ}\}$$

✳ 87A: code of noun-adjective typological feature

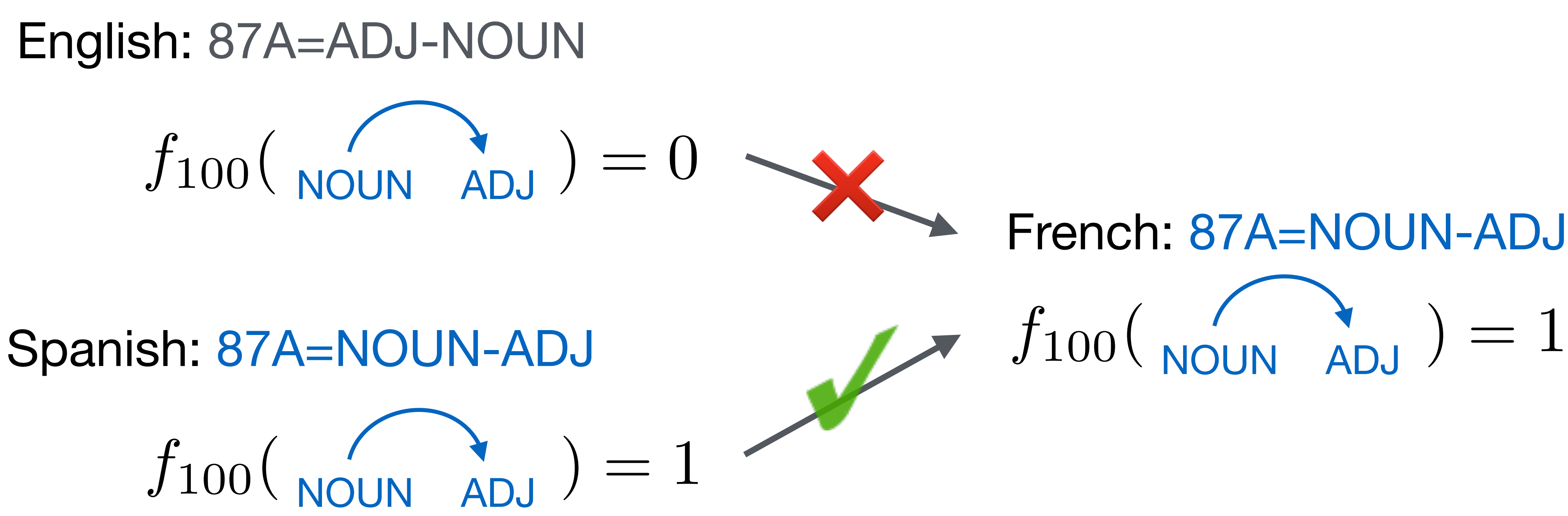
Traditional Approach: Feature Engineering

- Manually conjoin standard parsing features with typological features (Täckström et al., 2013)

$$f_{100}(\cdot) = \mathbb{I}\{\text{head POS=NOUN, modifier POS=ADJ, direction=Right, 87A=NOUN-ADJ}\}$$

✳ 87A: code of noun-adjective typological feature

- Features are selectively shared



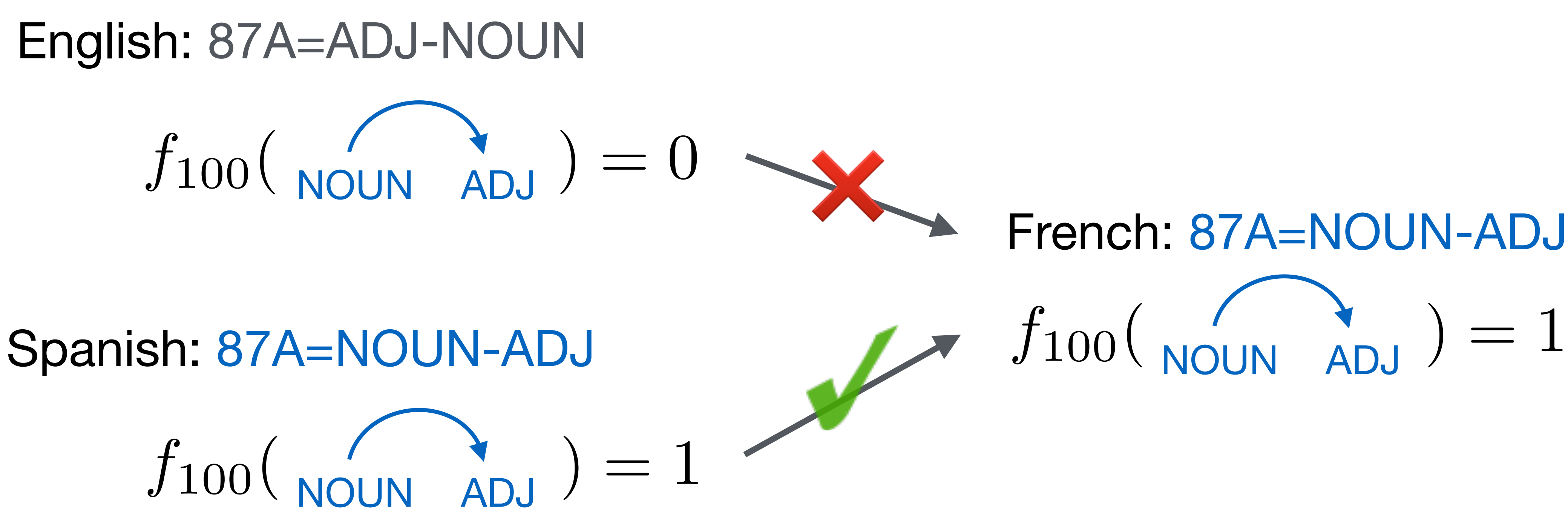
Traditional Approach: Feature Engineering

- Manually conjoin standard parsing features with typological features (Täckström et al., 2013)

$$f_{100}(\cdot) = \mathbb{I}\{\text{head POS=NOUN, modifier POS=ADJ, direction=Right, 87A=NOUN-ADJ}\}$$

✳ 87A: code of noun-adjective typological feature

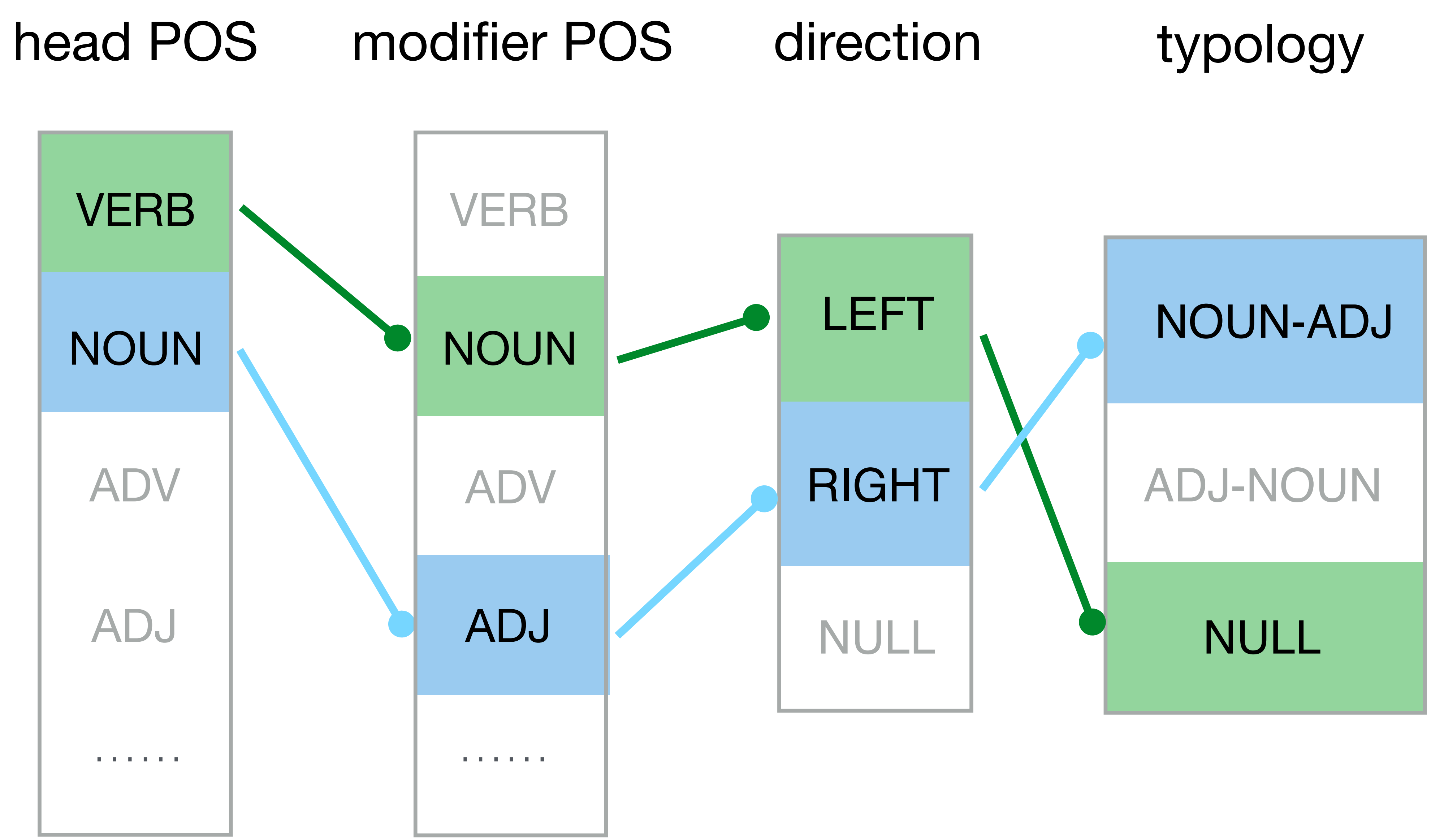
- Features are selectively shared



- In practice, need to manually construct hundreds of features

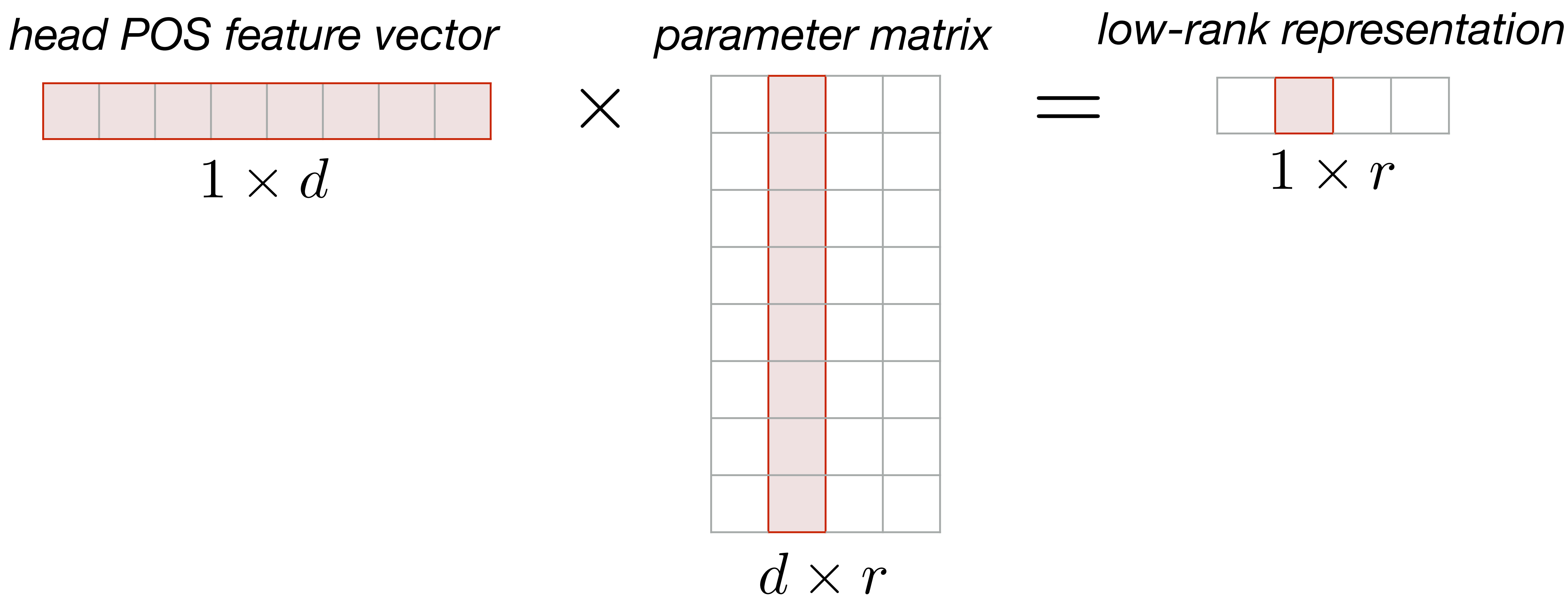
Tensor Scoring Method

- Represent arc features in a tensor view (e.g., 4-way tensor)
- Automatically capture all possible feature combinations

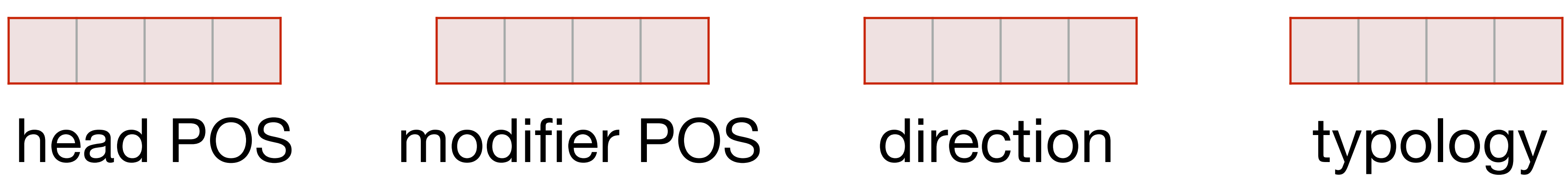


Low-rank Feature Representation

- Avoid parameter explosion via low-rank factorization
- Learn feature mappings to a low-rank representation

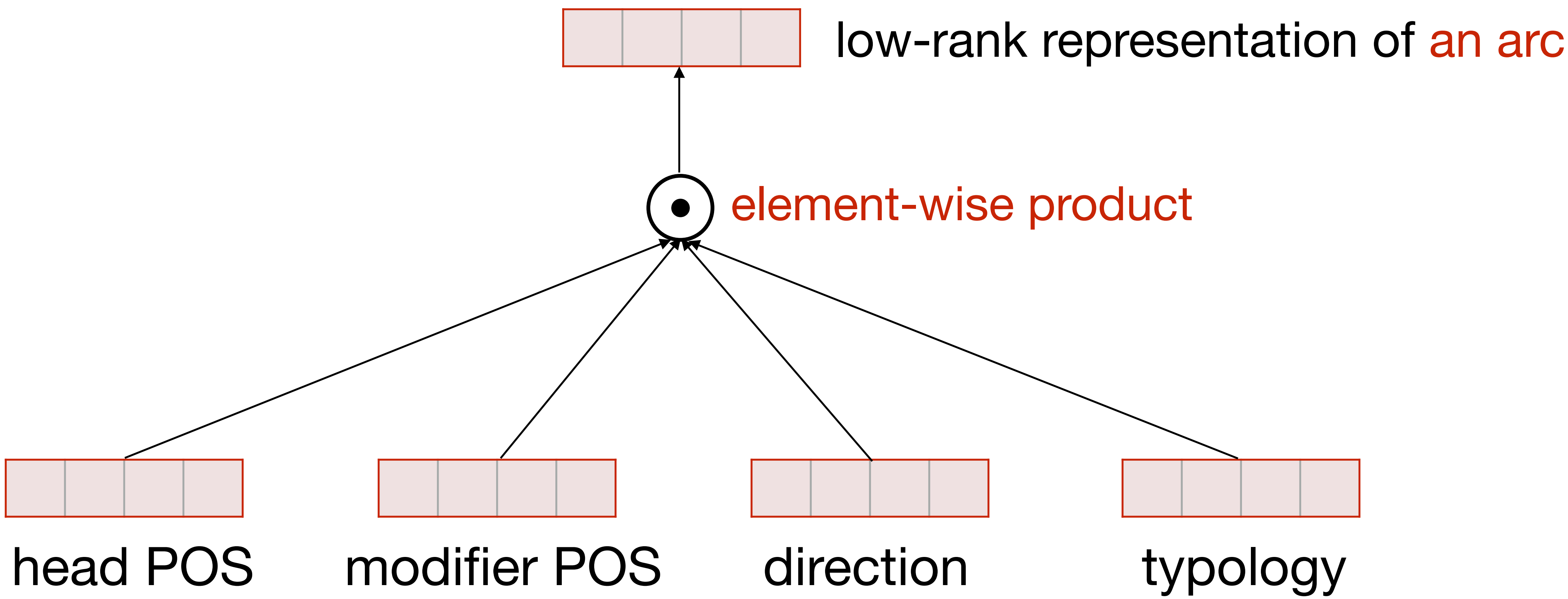


Low-rank Feature Representation



Low-rank Feature Representation

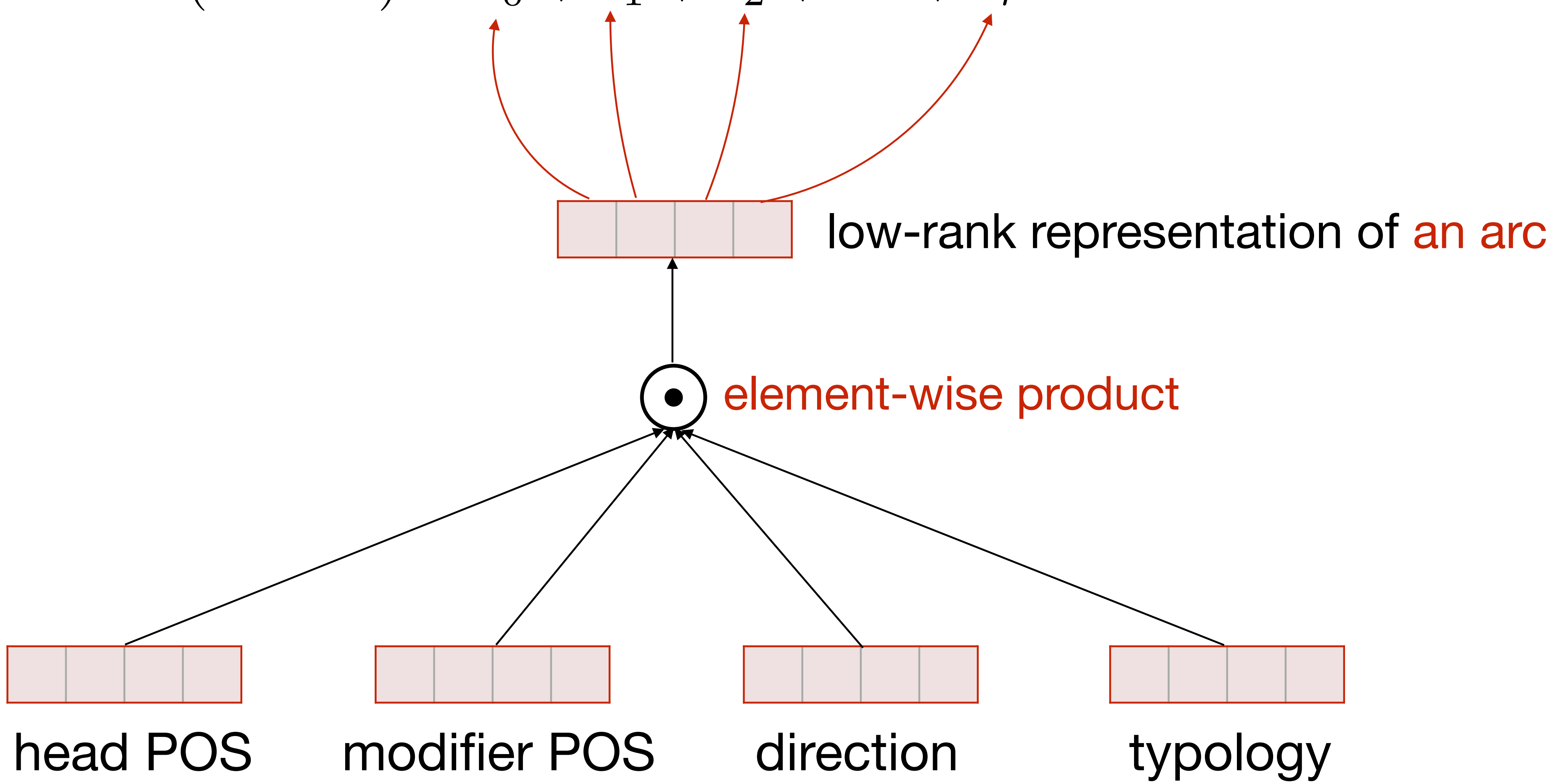
- Compute low-rank representation of an arc via element-wise product



Low-rank Feature Representation

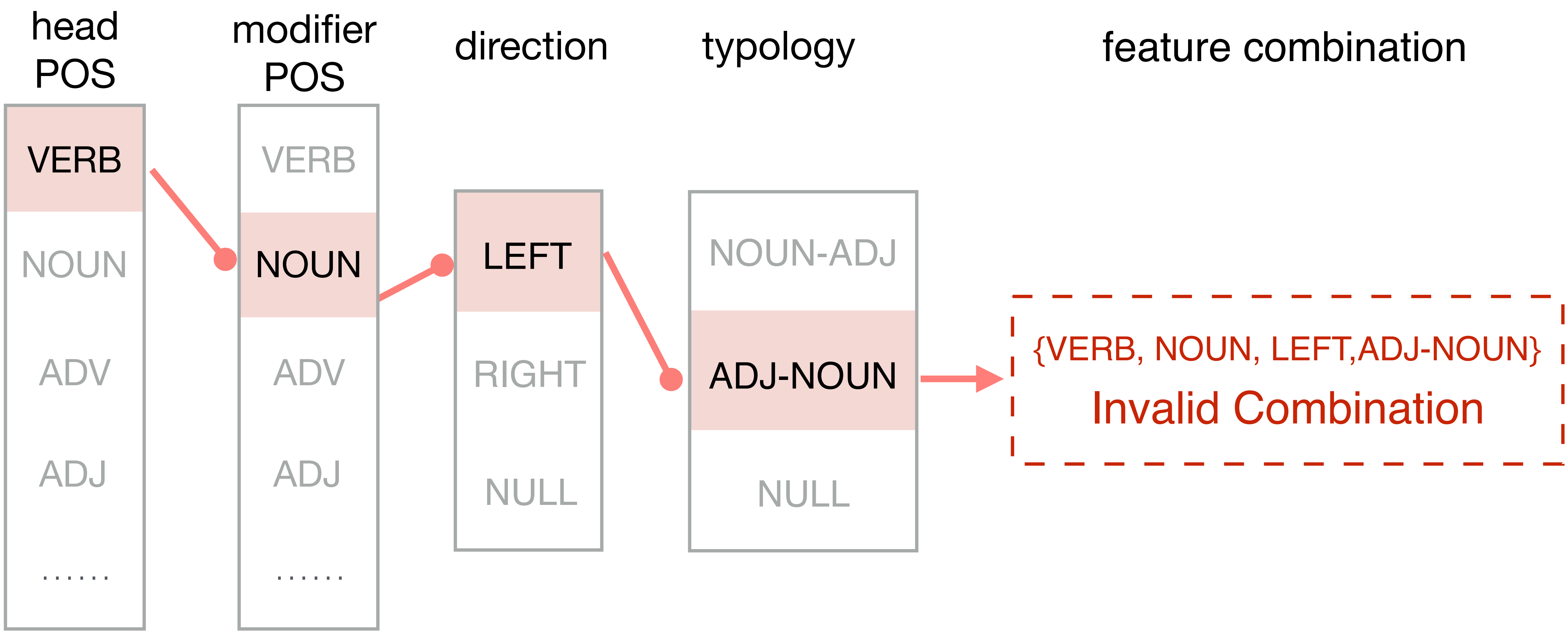
- Compute low-rank representation of an arc via element-wise product
- Compute arc score as:

$$S(h \rightarrow m) = e_0 + e_1 + e_2 + \cdots + e_r$$



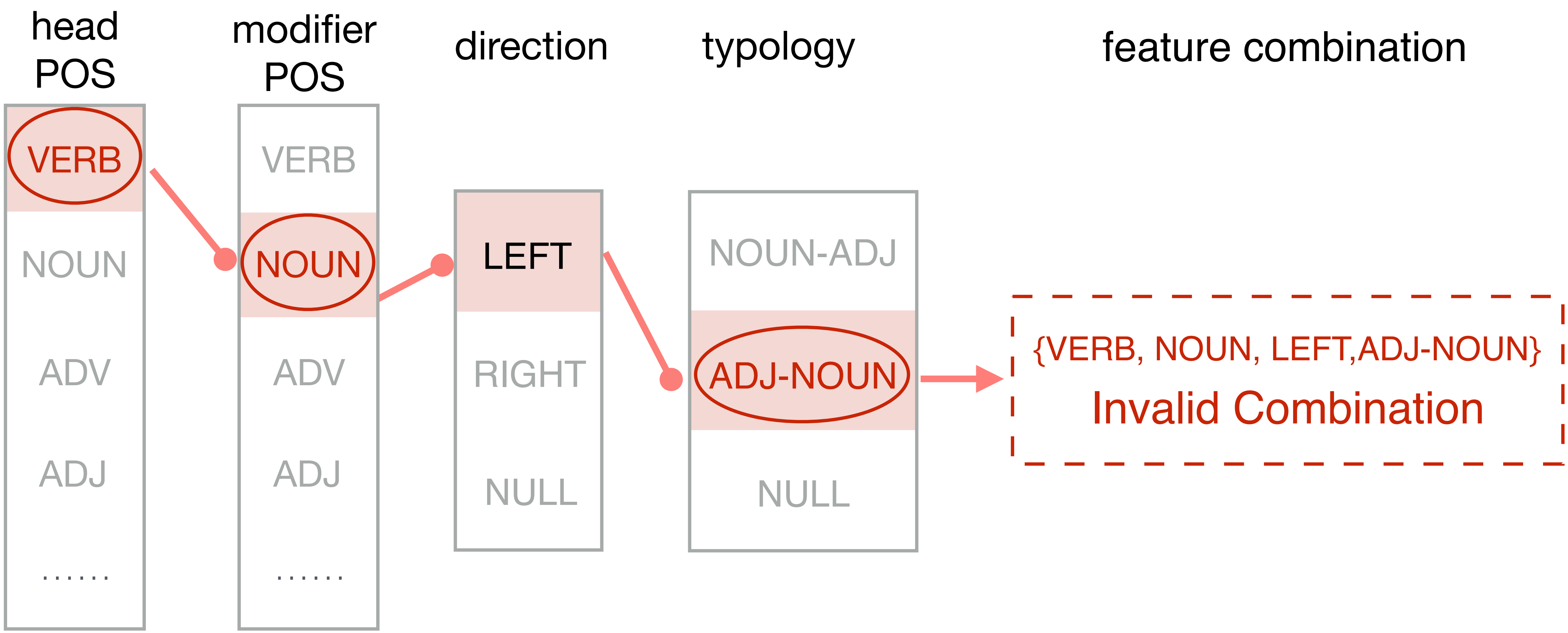
Issue of Tensor Methods

- Capture invalid feature combinations and assign non-zero weights

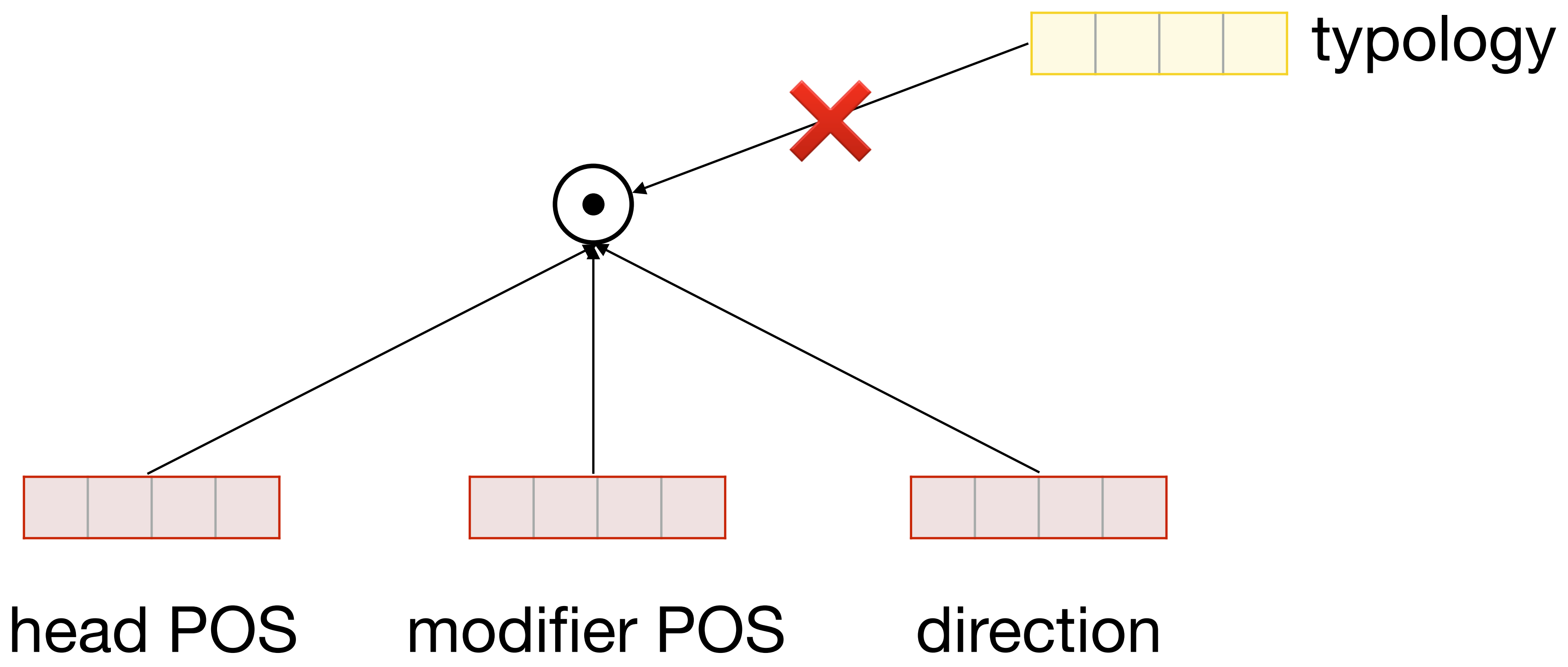


Issue of Tensor Methods

- Capture invalid feature combinations and assign non-zero weights
- Should avoid directly taking tensor-product between typology and others

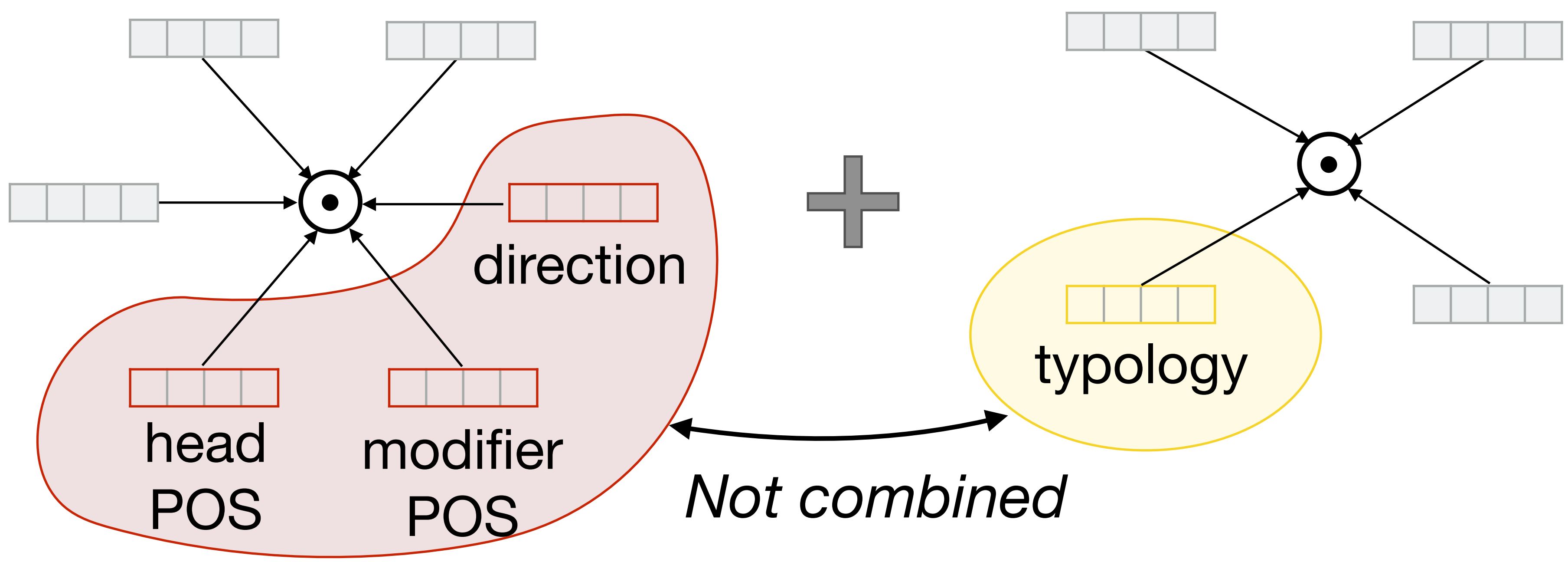


Avoid Product Operation



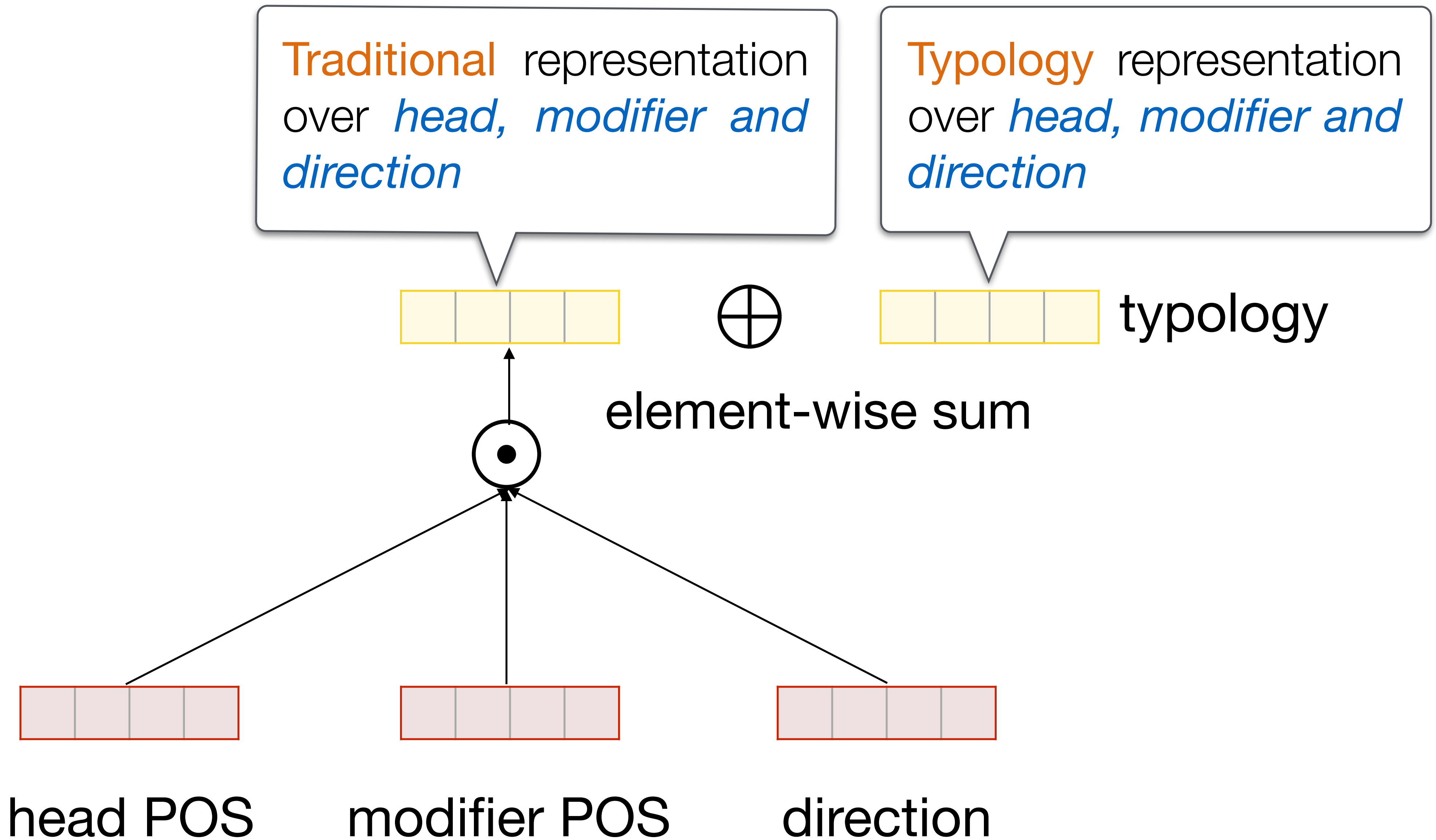
Target Feature Combination

- Union of different feature groups



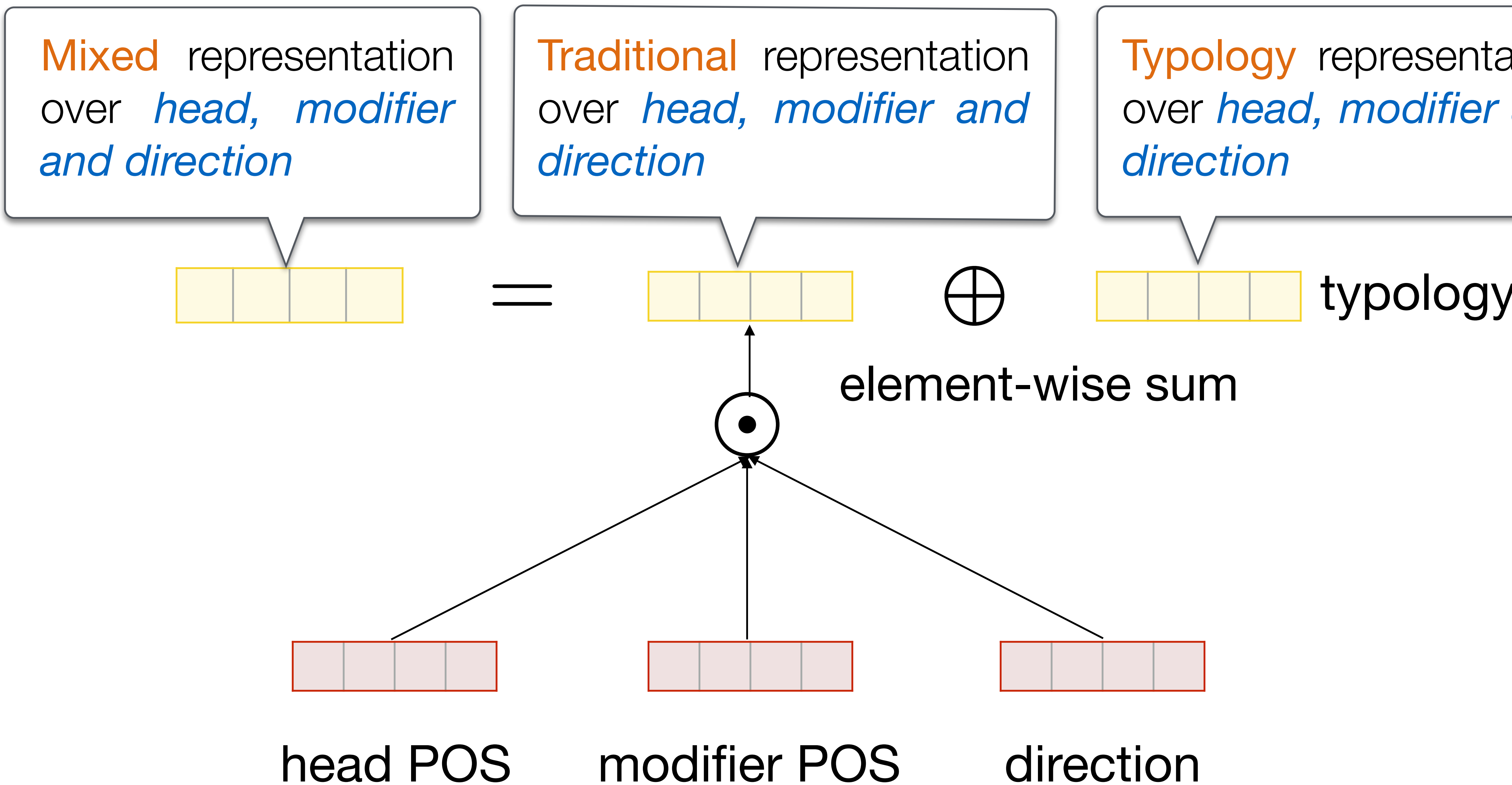
Solution: Hierarchical Structure

- Element-wise sum operation over different representations of the same set of atomic features

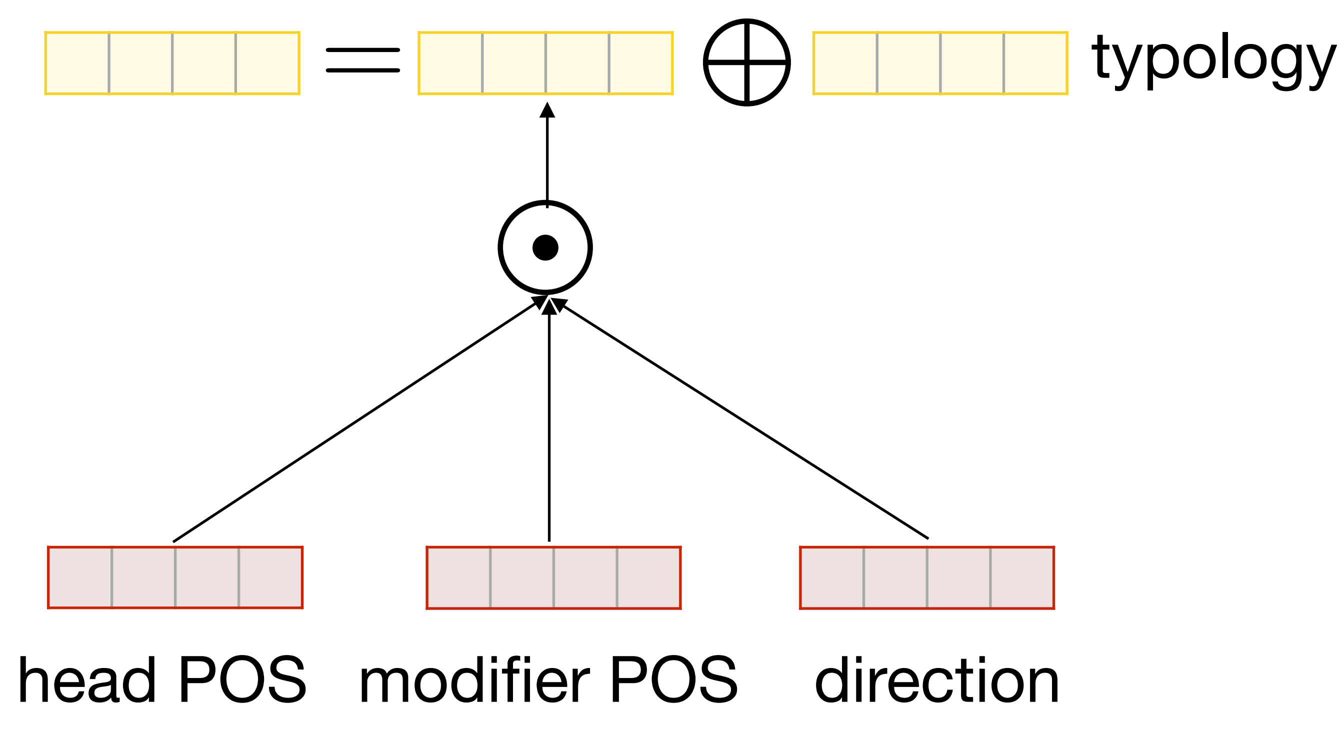


Solution: Hierarchical Structure

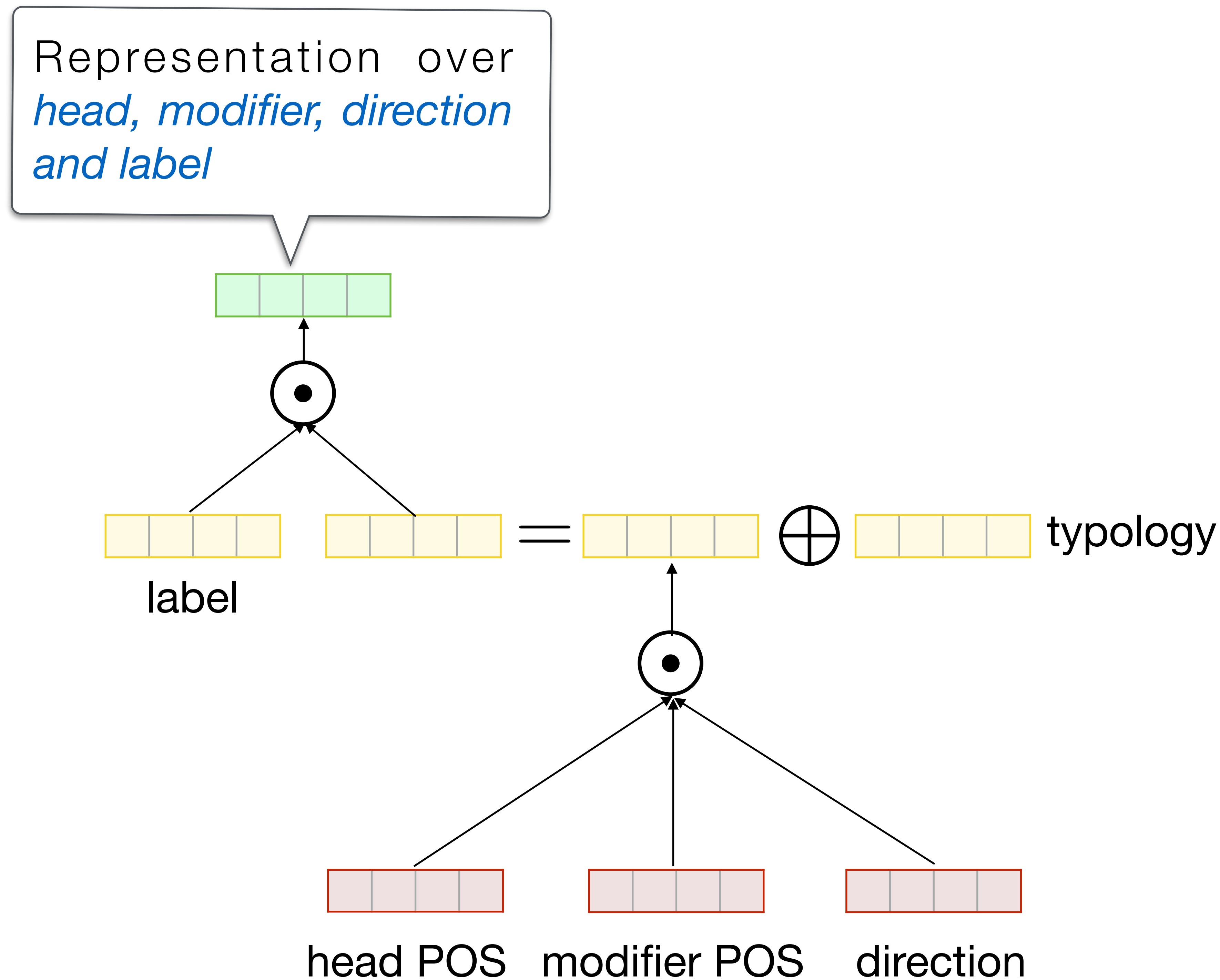
- Element-wise sum operation over different representations of the same set of atomic features



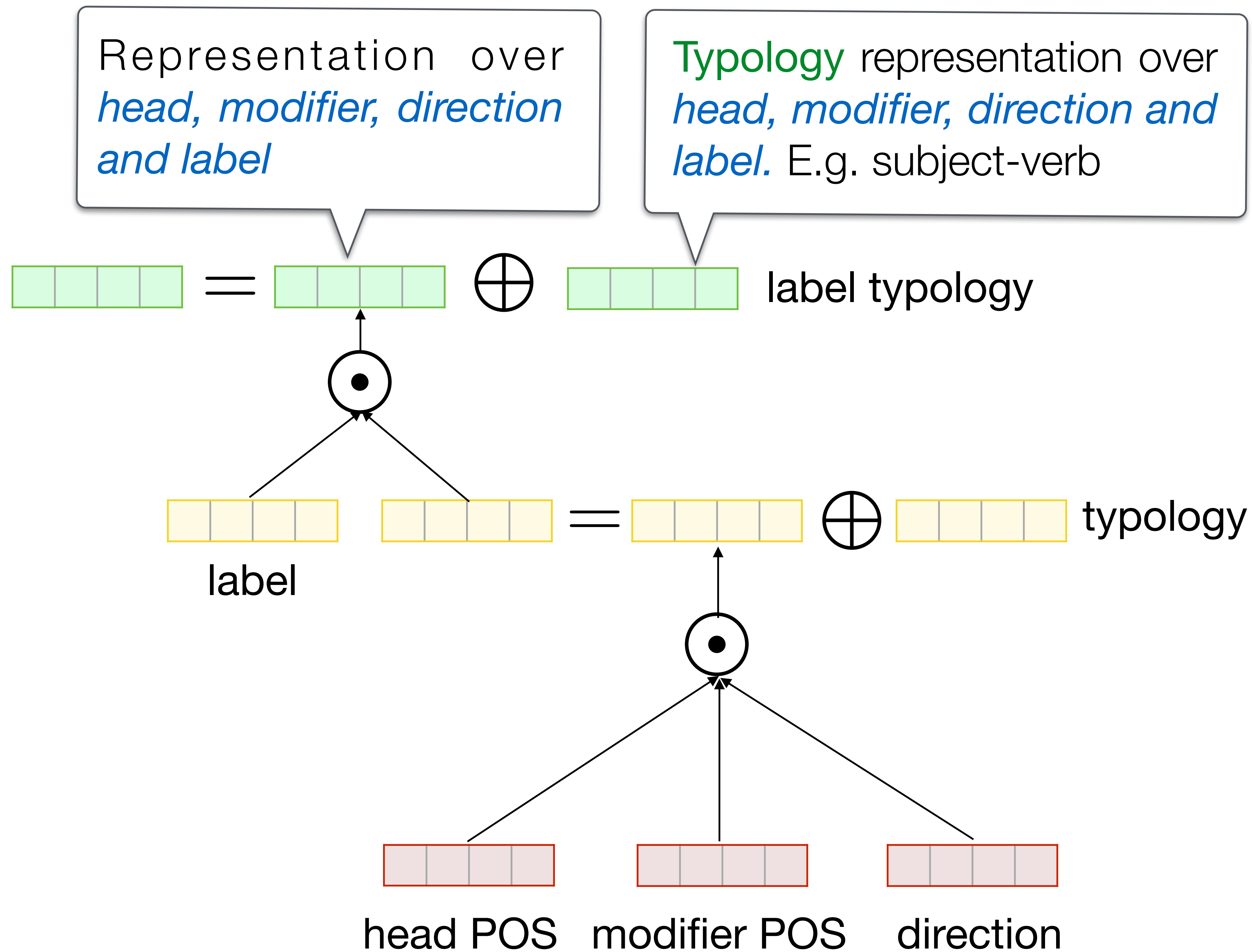
Solution: Hierarchical Structure



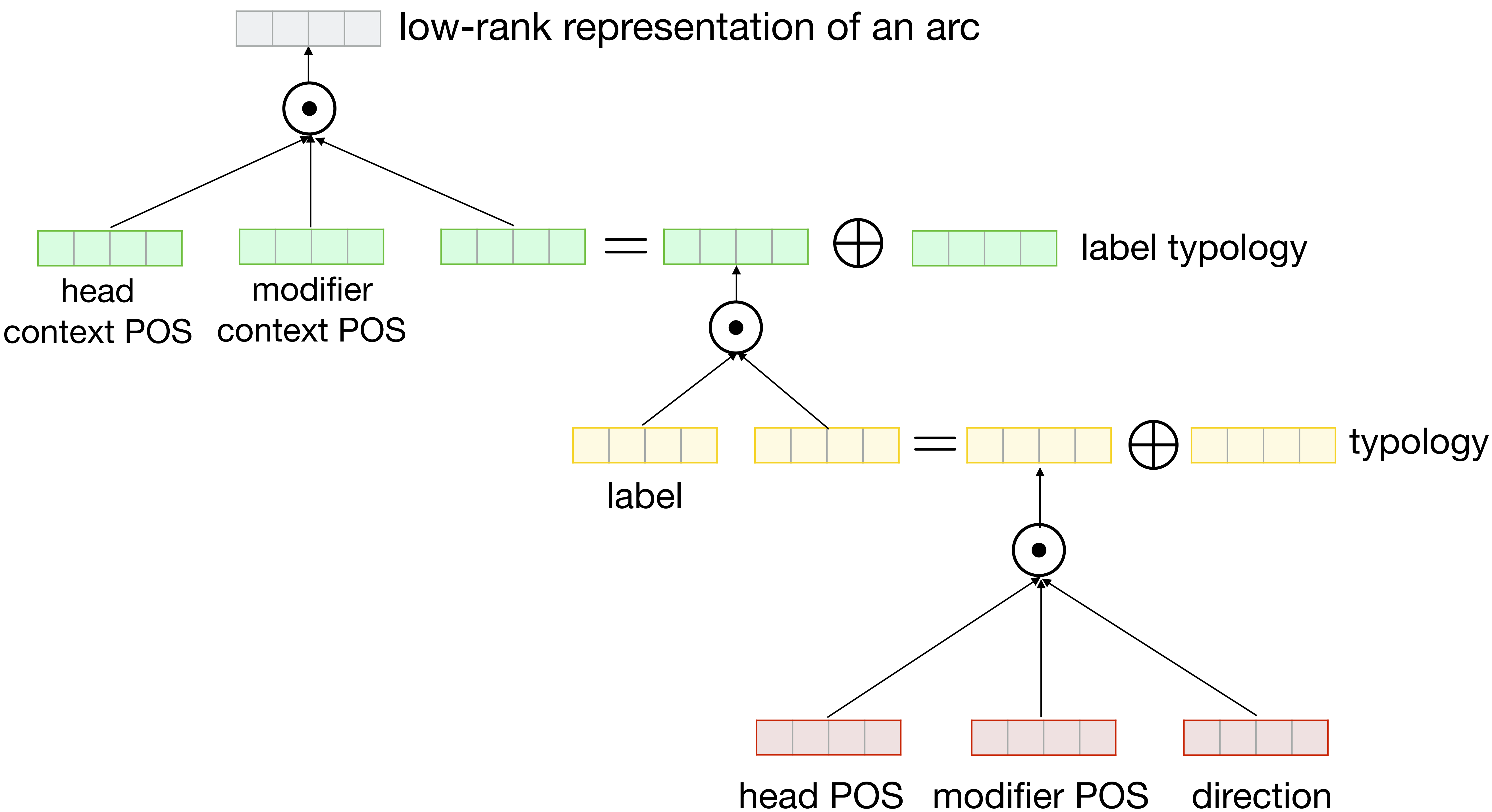
Solution: Hierarchical Structure



Solution: Hierarchical Structure

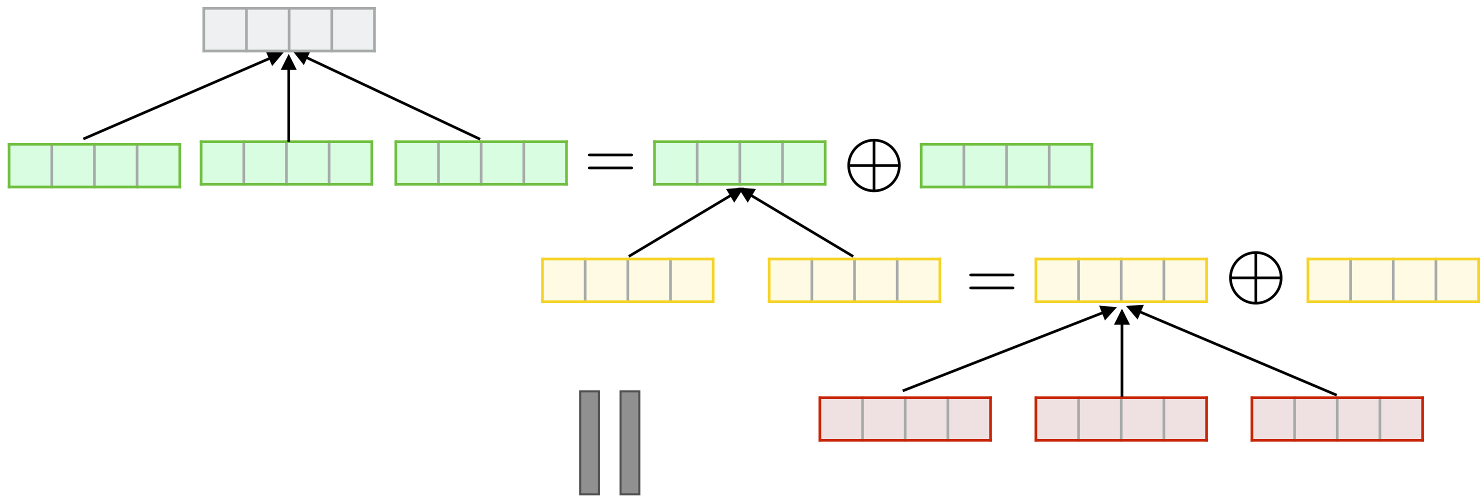


Solution: Hierarchical Structure



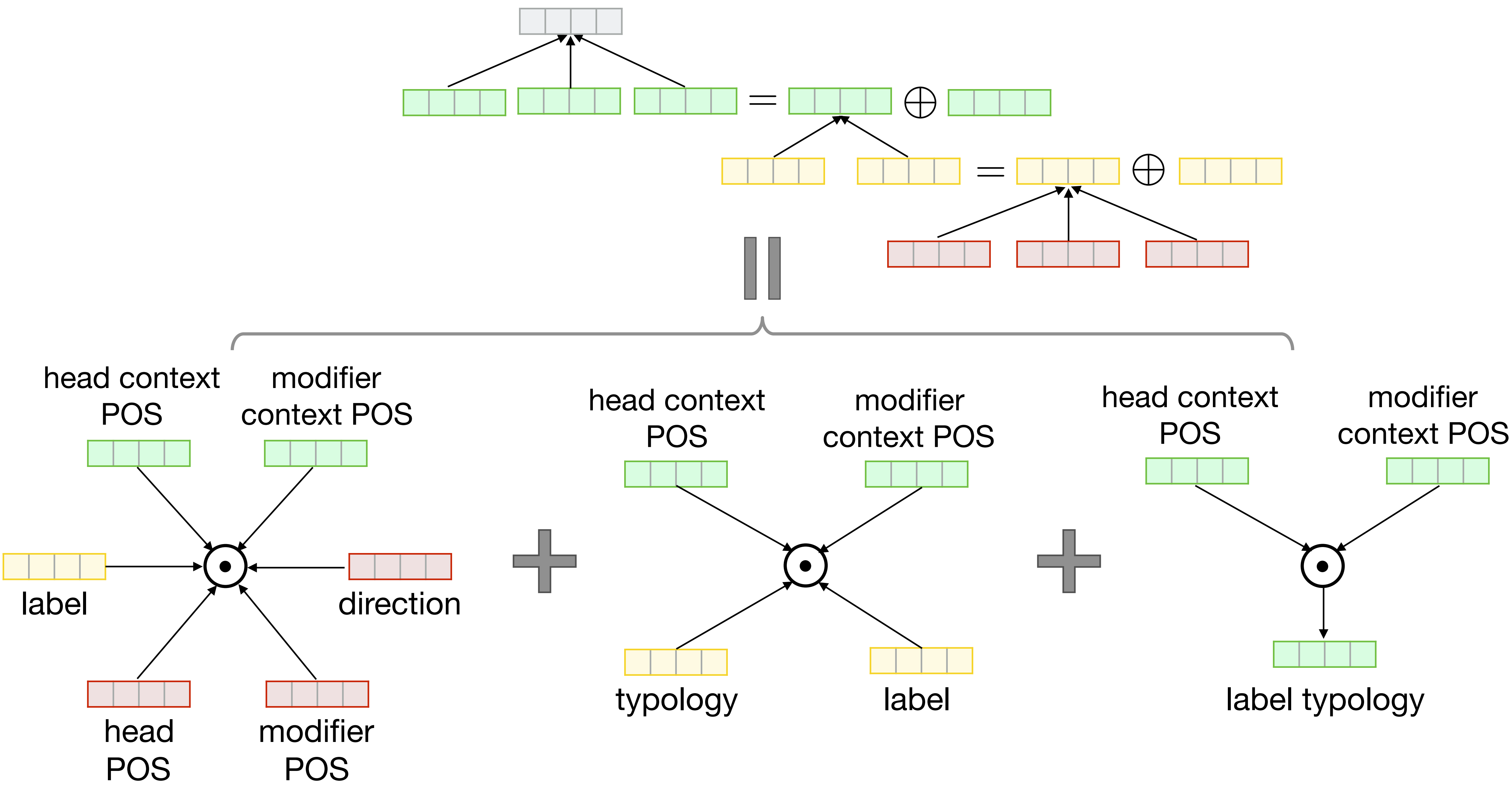
Algebraic Interpretation

- Algebraically equal the sum of three multiway tensors with shared parameters
- Capture three groups of feature combinations



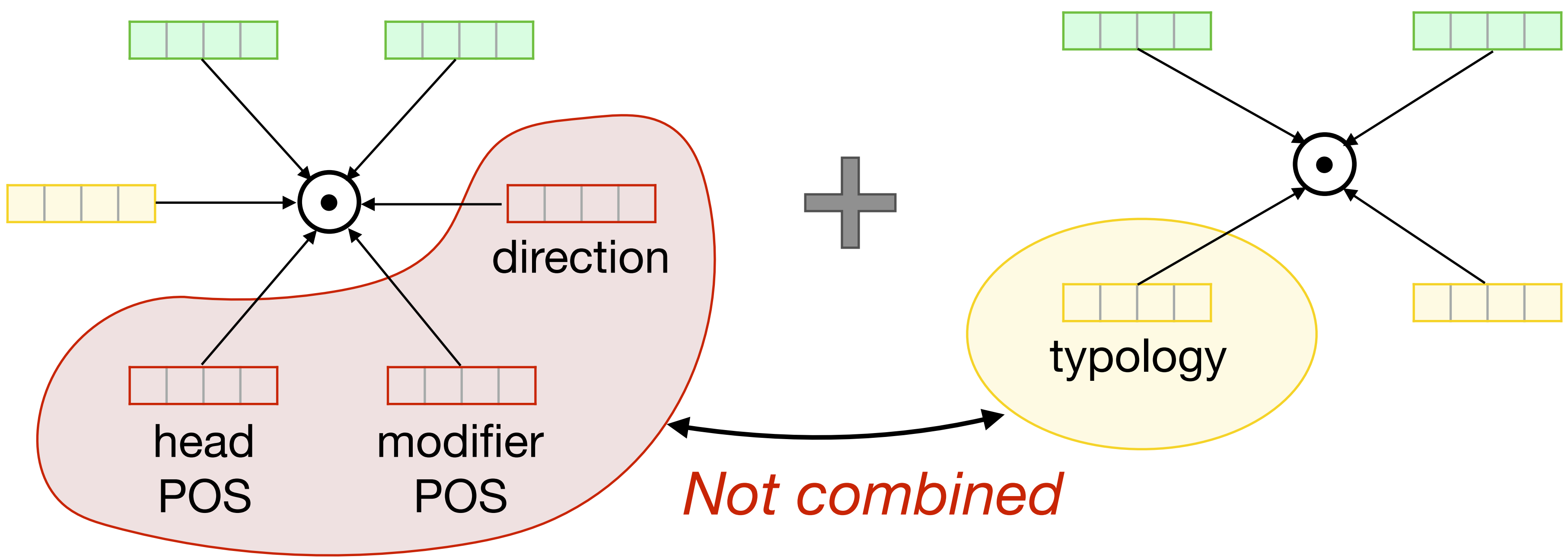
Algebraic Interpretation

- Algebraically equal the sum of three multiway tensors with shared parameters
- Capture three groups of feature combinations



Avoid Invalid Features

- Exclude the combination of typology with head, modifier and direction



- Assign zero weights to invalid features
  - Weight of {head POS=VERB, mod POS=NOUN, typology=ADJ-NOUN} is 0

Parameter Initialization and Learning

Algebraic view:

Compute the gradient for each multiway tensor and take the sum

Tensor initialization:

Use iterative power methods

Parameter learning:

Adopt online learning with passive-aggressive algorithm

Other details:

Follow previous work (Lei et al., 2015)

Experimental Setup

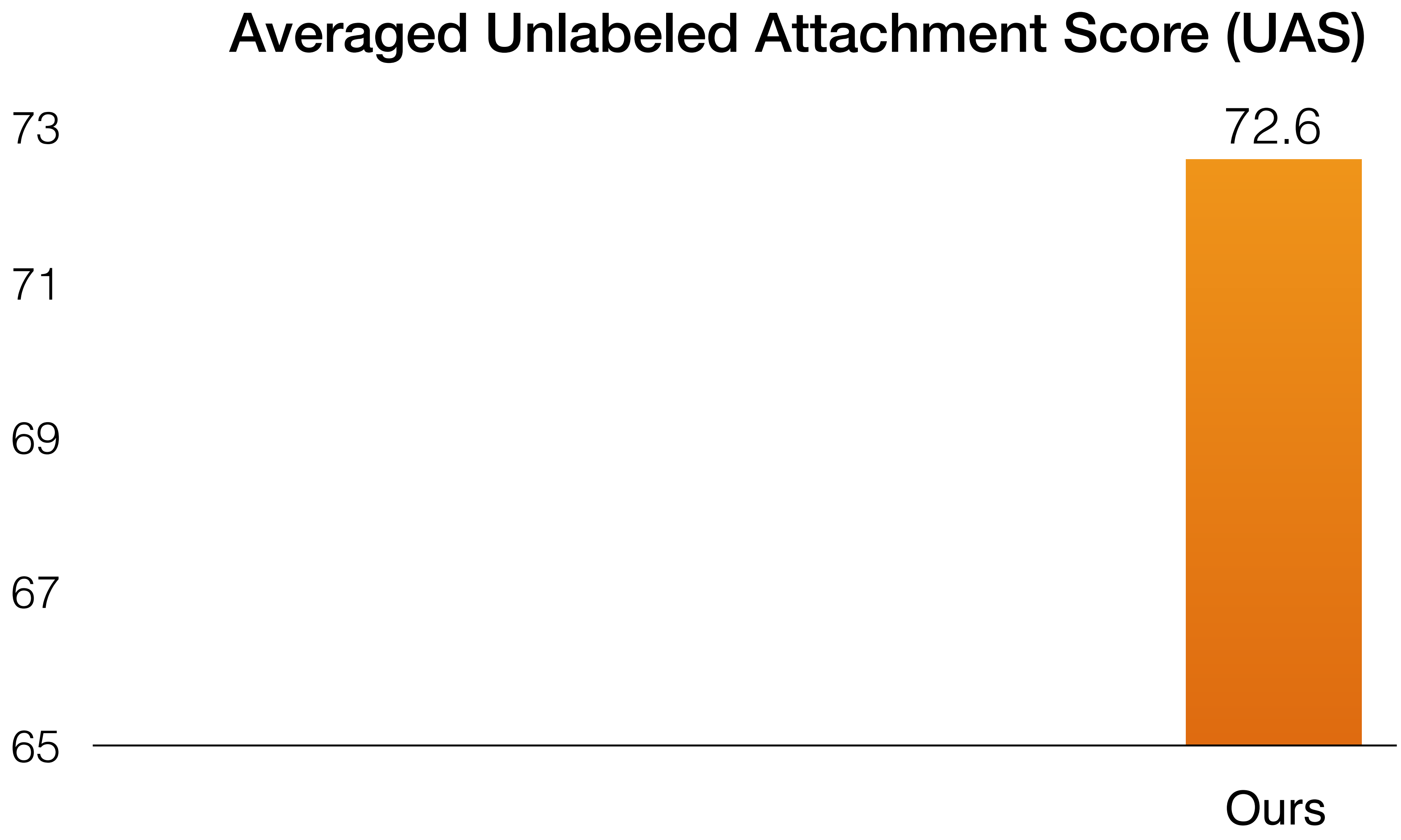
Dataset: Universal Dependency Treebank v2.0

- 10 languages
- Universal POS tags (12 tags)
- Stanford dependency labels (40 labels)

Baselines:

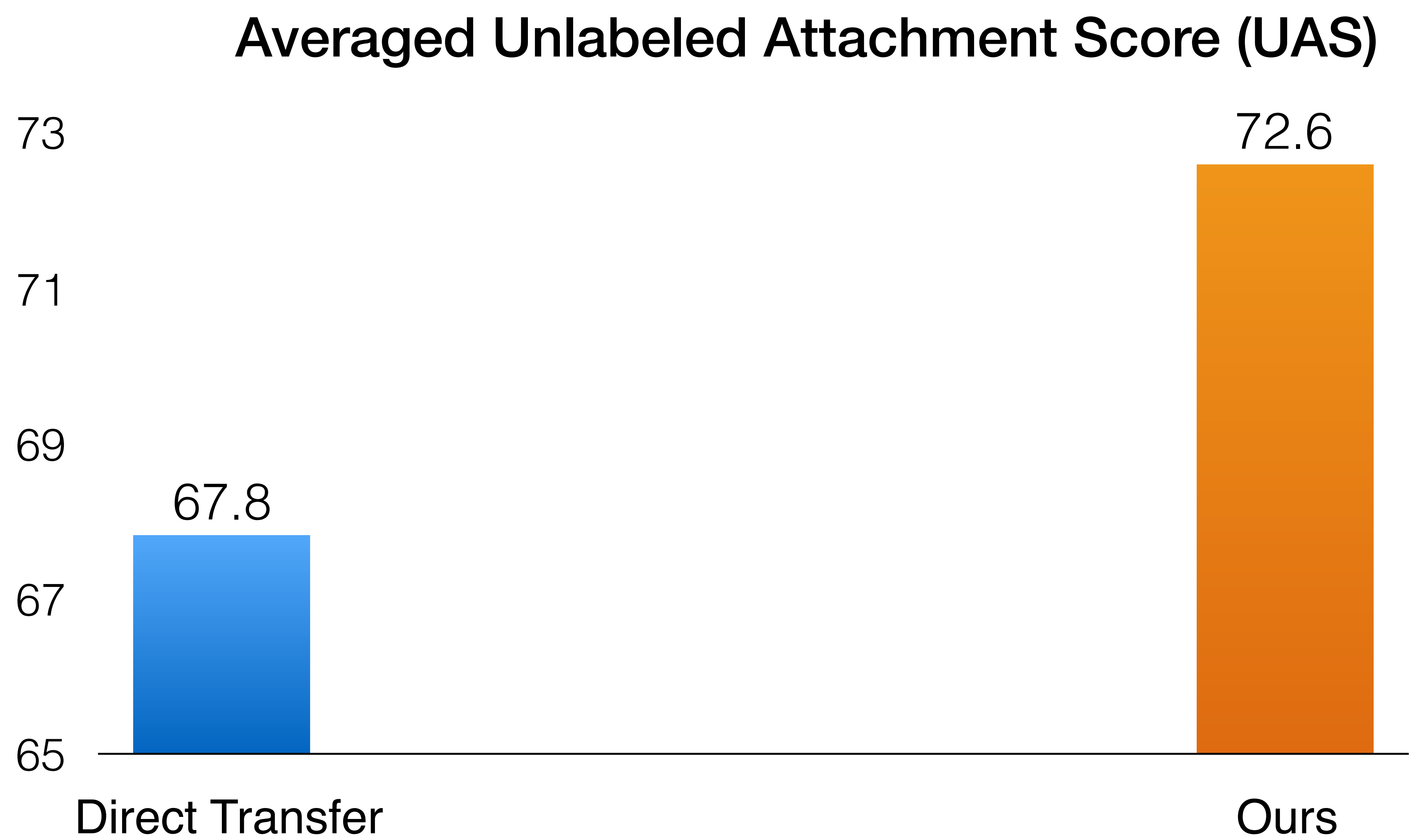
- Direct transfer (McDonald et al., 2005)
- Feature-based transfer (Täckström et al., 2013)
- Traditional multiway tensor

Unsupervised Results

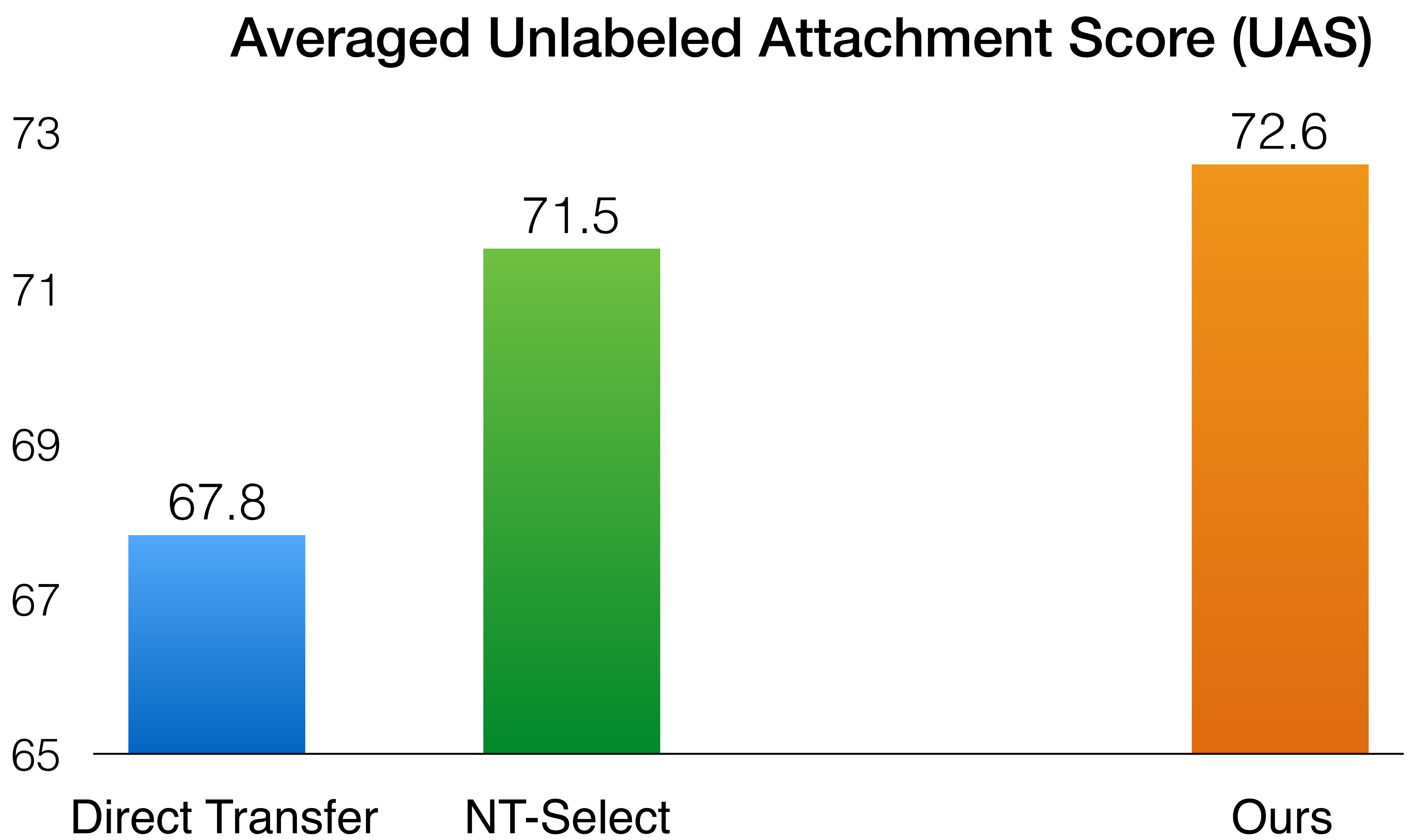


- Setting: no annotations in the target language

Unsupervised Results

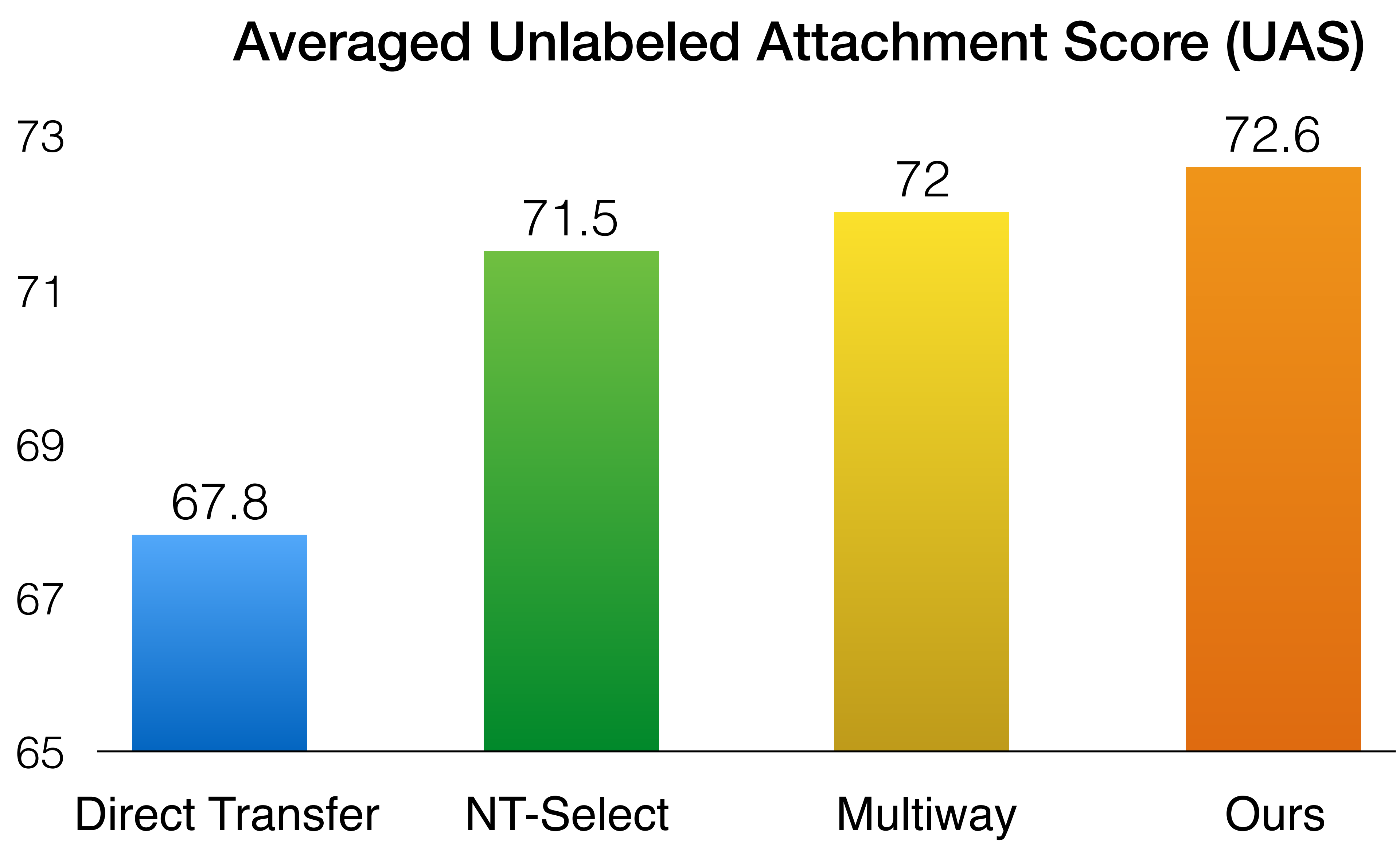


Unsupervised Results



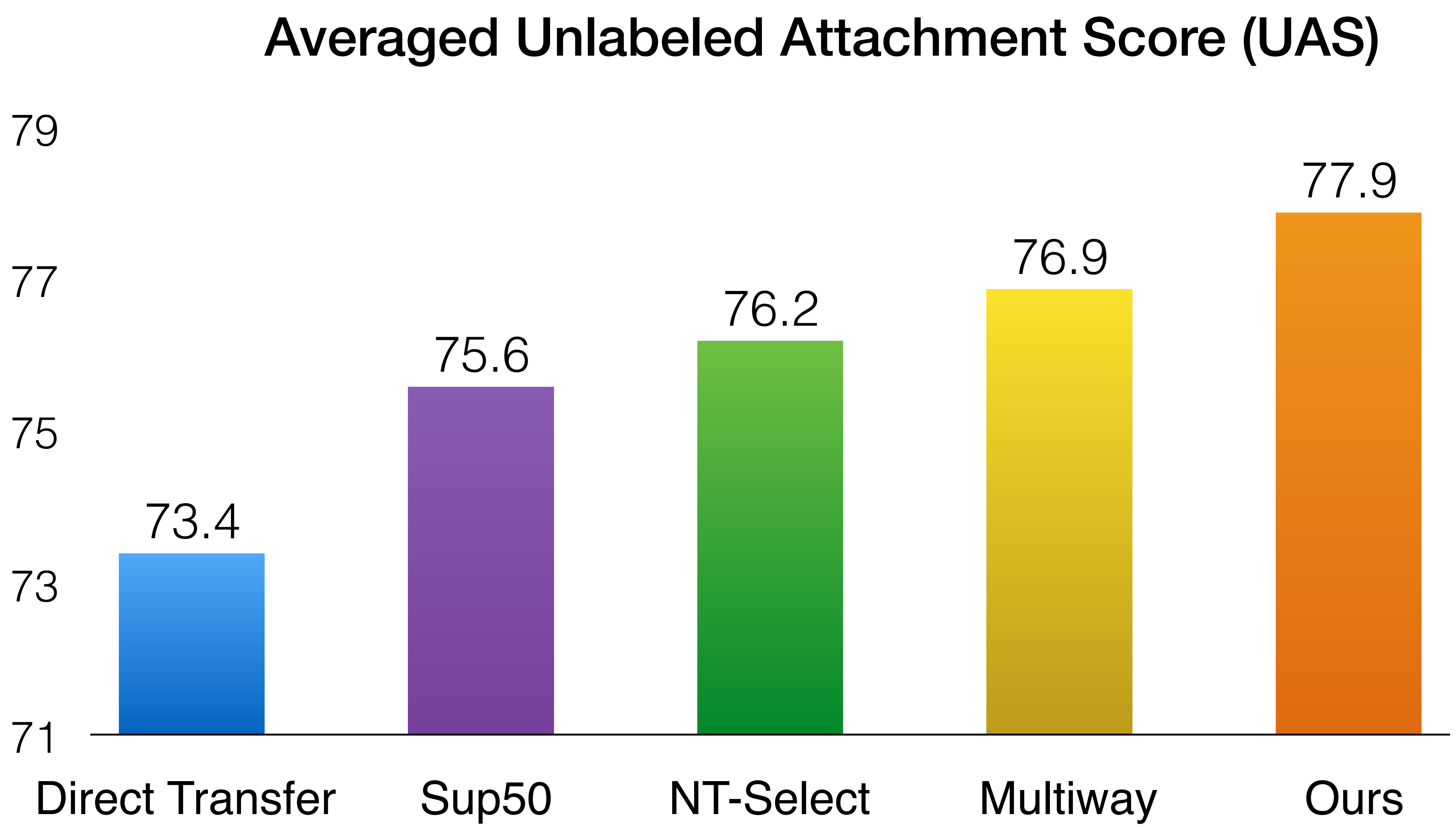
- NT-Select: our model without the tensor component, corresponding to prior feature-based method (Täckström et al., 2013)

Unsupervised Results



- Multiway: traditional multiway tensor without hierarchical structure

Semi-supervised Results



- Setting: 50 annotated sentences in the target language
- Sup50: trained only on the 50 sentences in the target language

Summary

- *Modeling:* we present a hierarchical tensor that effectively uses linguistic prior knowledge
- *Performance:* our model outperforms state-of-the-art approach and traditional tensors
- *Limitation:* our model heavily relies on non-lexical transfer via universal POS tags

Next part: lexical-level multilingual transfer

Our Approach

Multilingual Transfer:

- Hierarchical tensors for dependency parsing

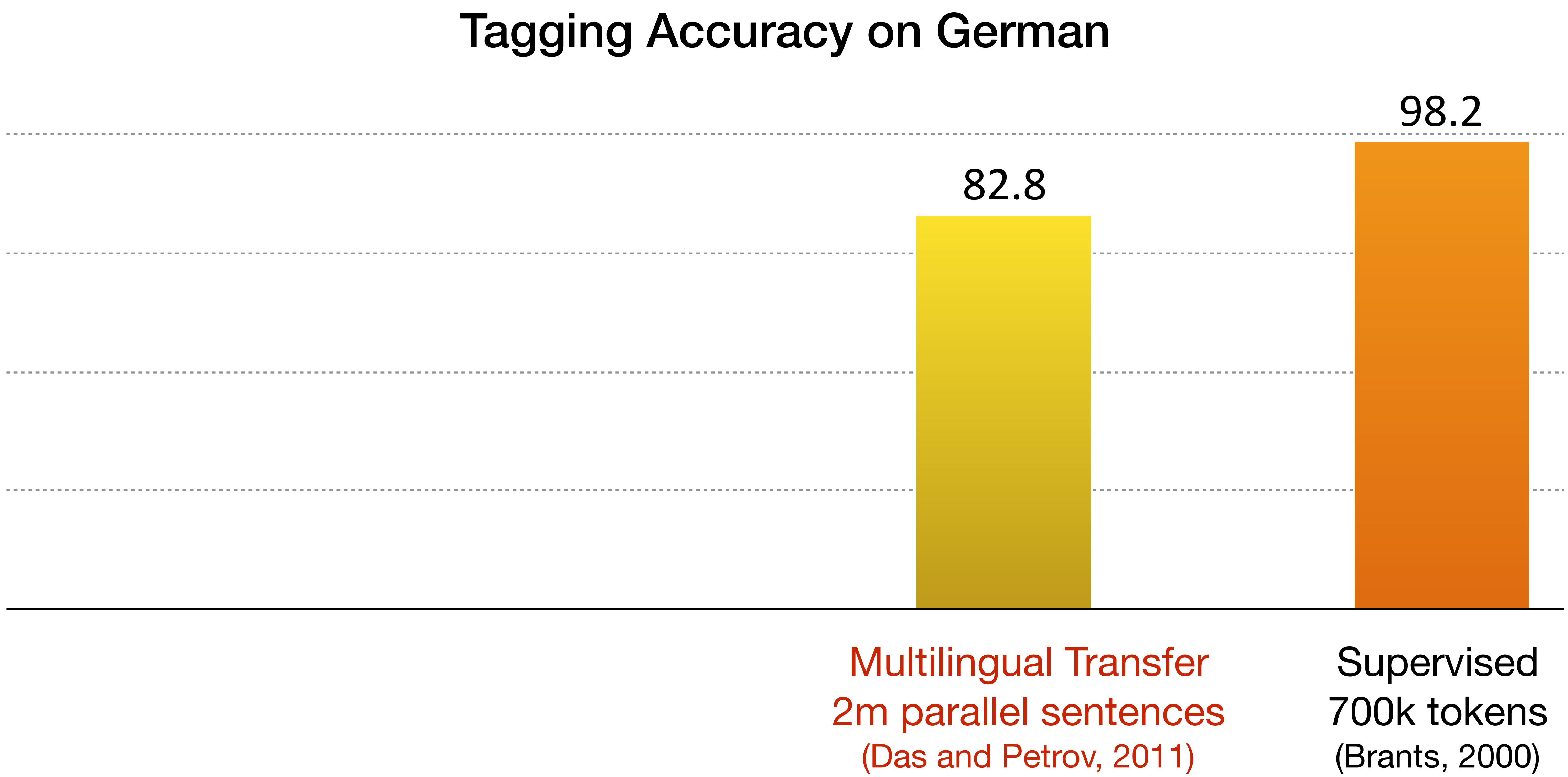
- Multilingual embeddings for POS tagging
  - *Effective multilingual transfer with ten translation pairs*

Monolingual Transfer:

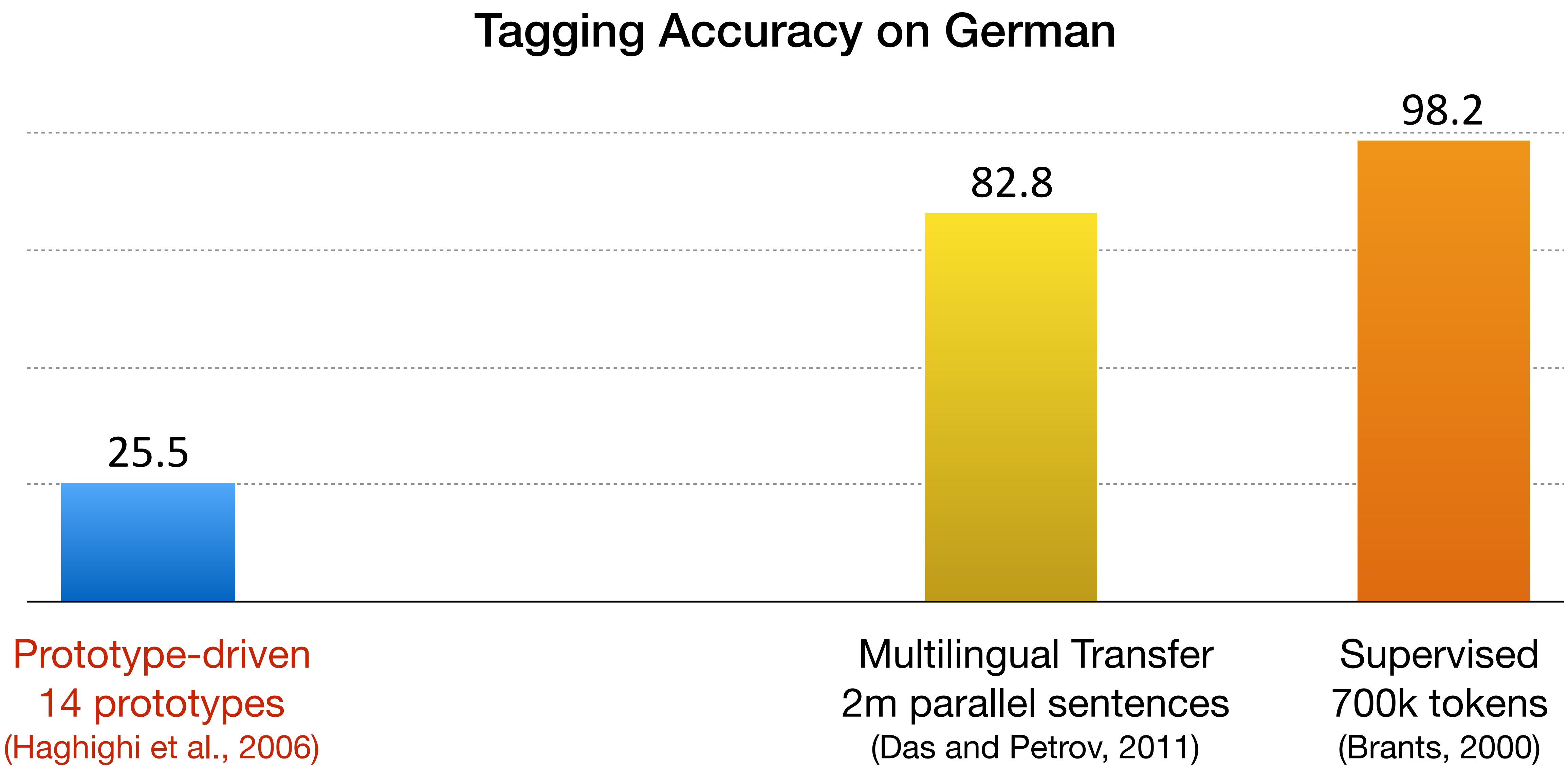
- Adversarial networks for aspect transfer

Multilingual Transfer of POS Tagging

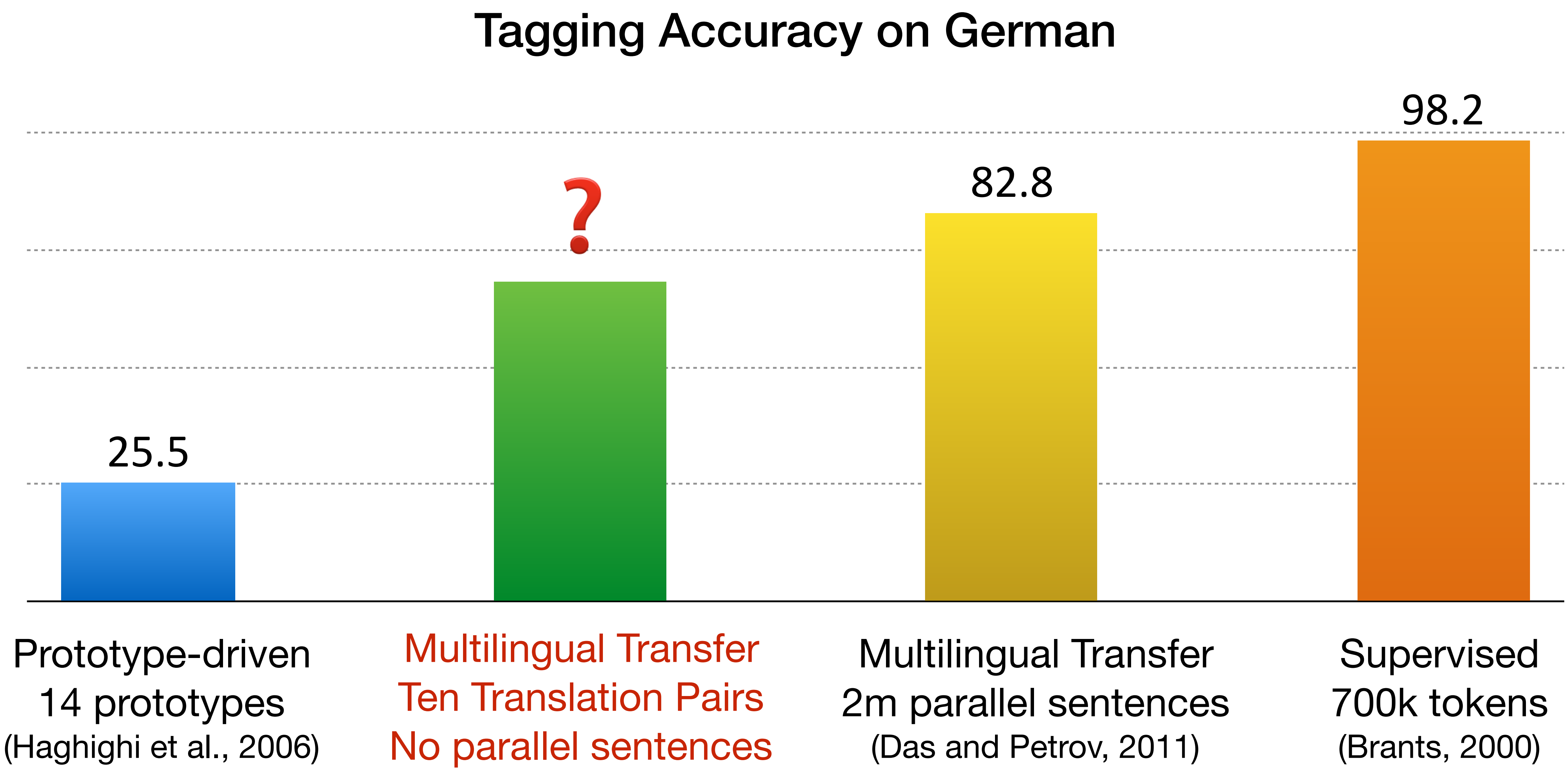
Multilingual Transfer of POS Tagging



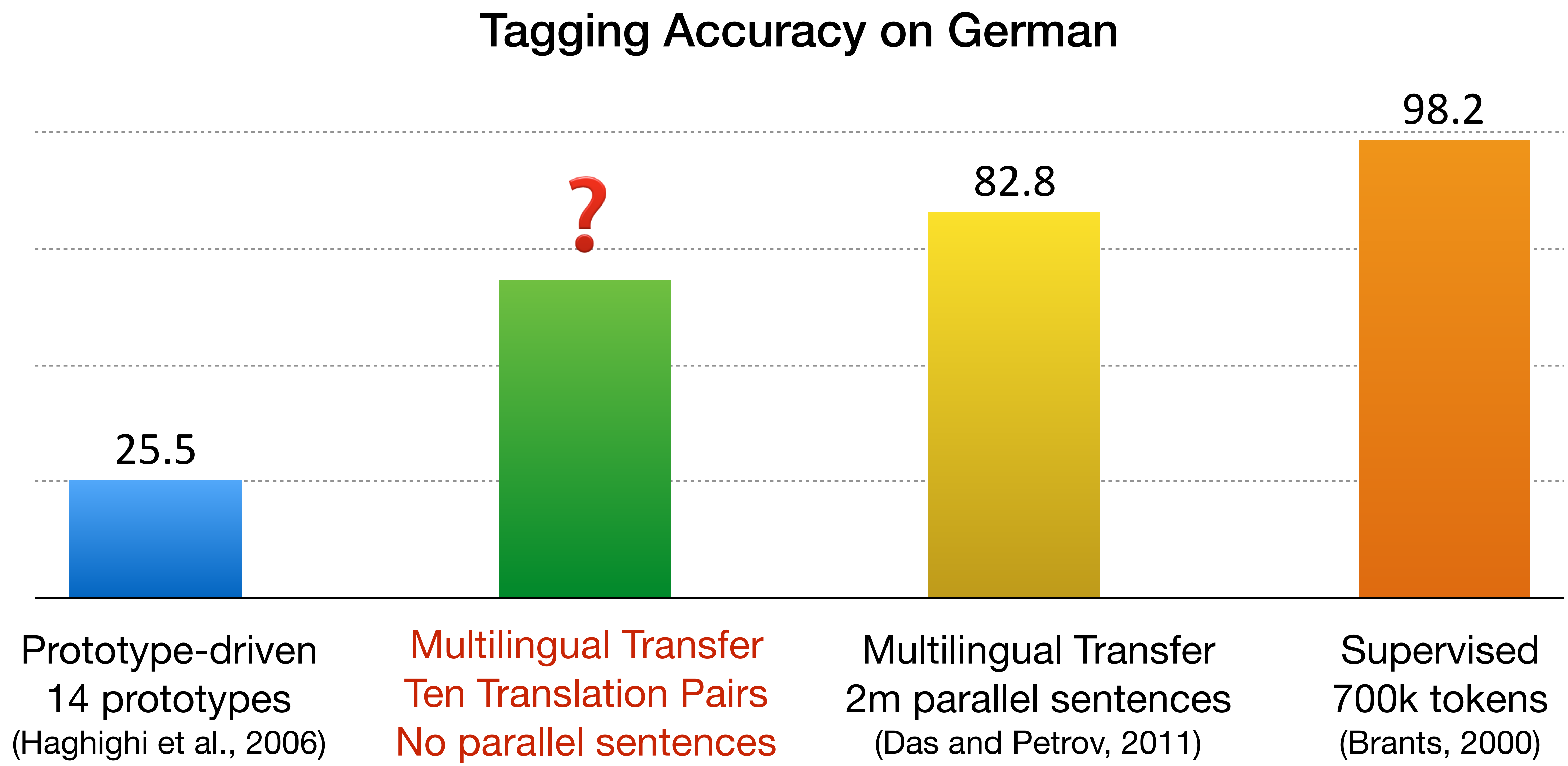
Multilingual Transfer of POS Tagging



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

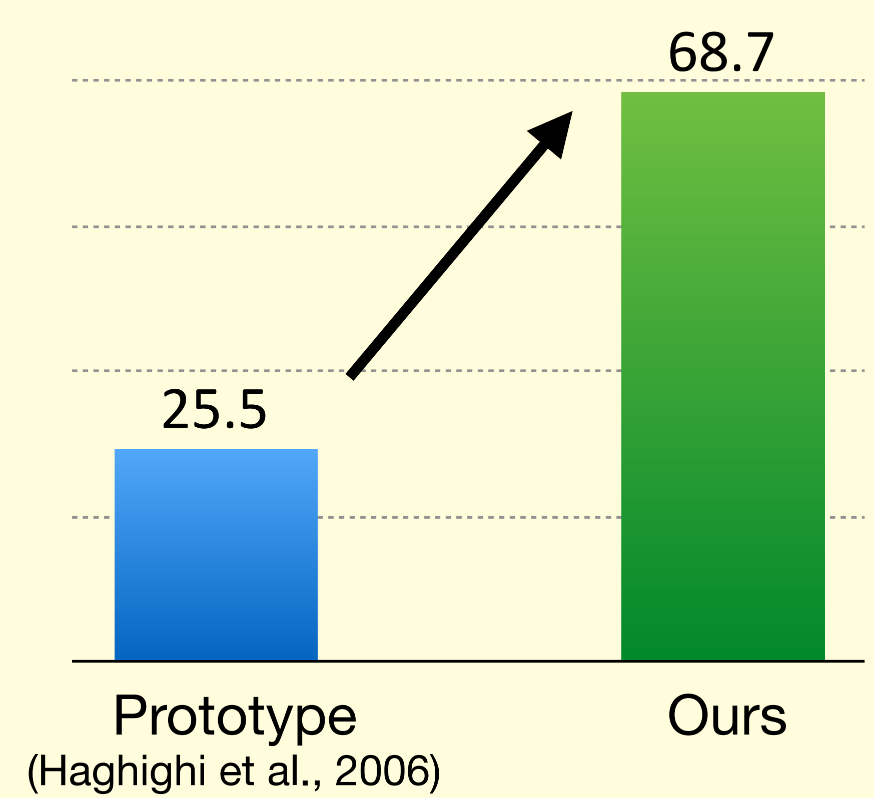
- Task: multilingual transfer of part-of-speech (POS) tagging
- Data:

	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



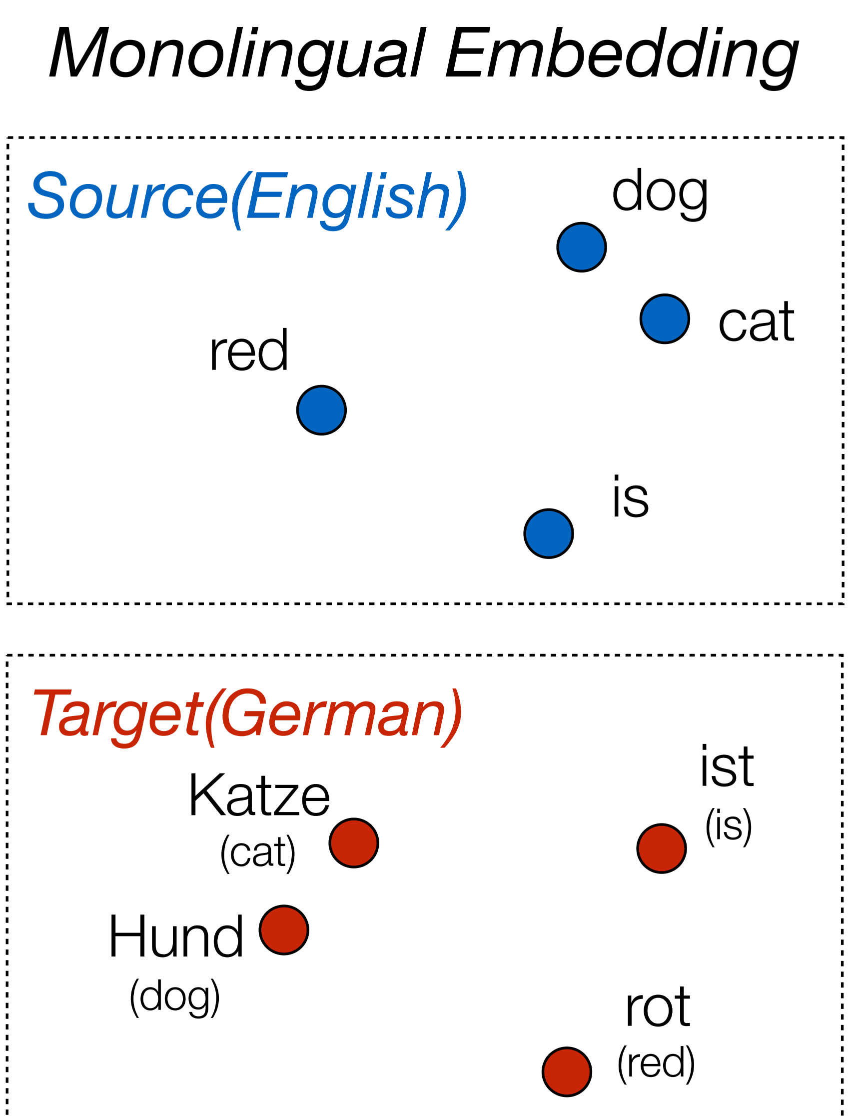
Our Two-step Method

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

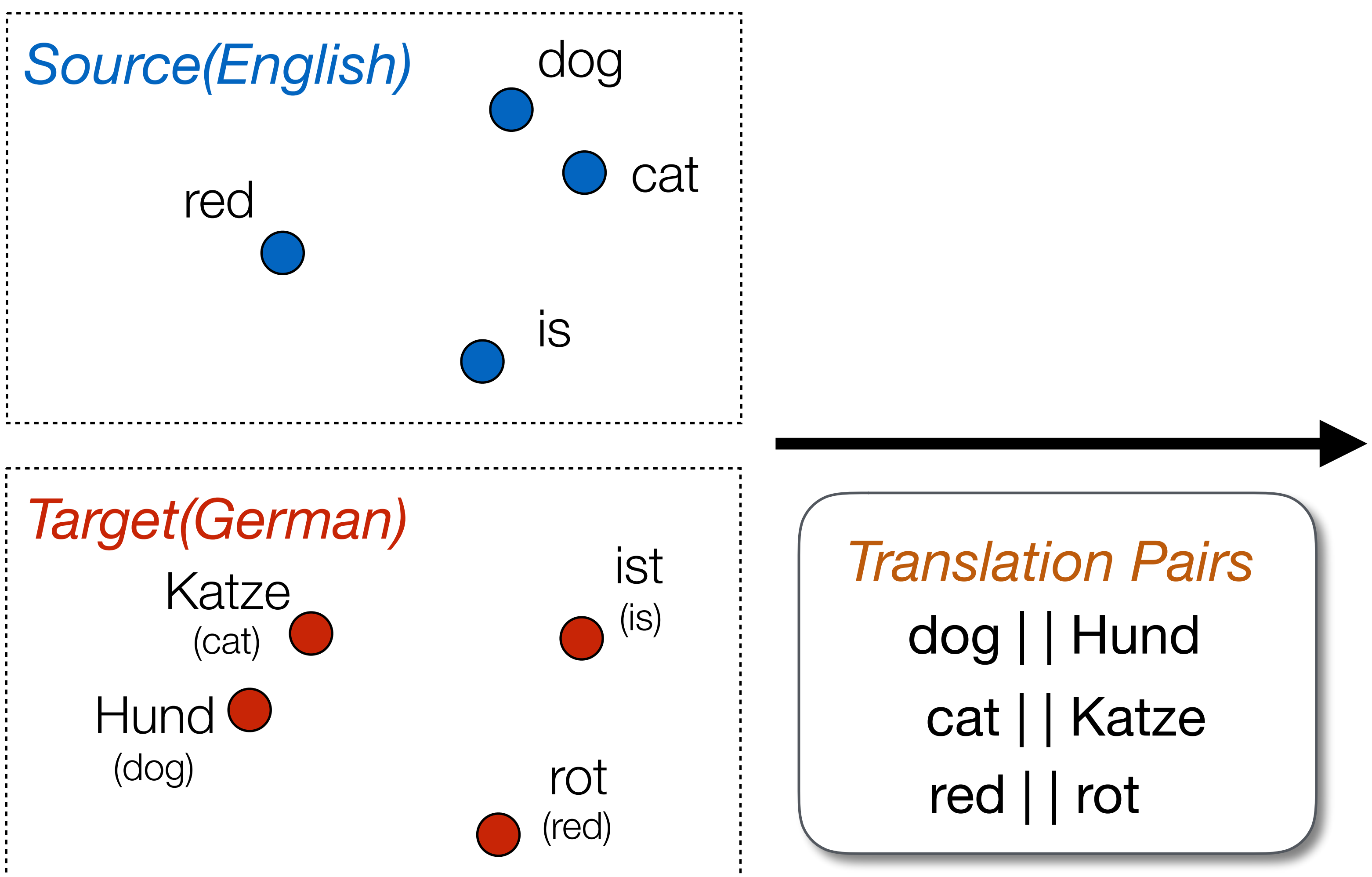
- Goal: find a linear transformation from target to source embedding space
- Objective: minimize the distance between translation pairs



Coarse Mapping between Embeddings

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

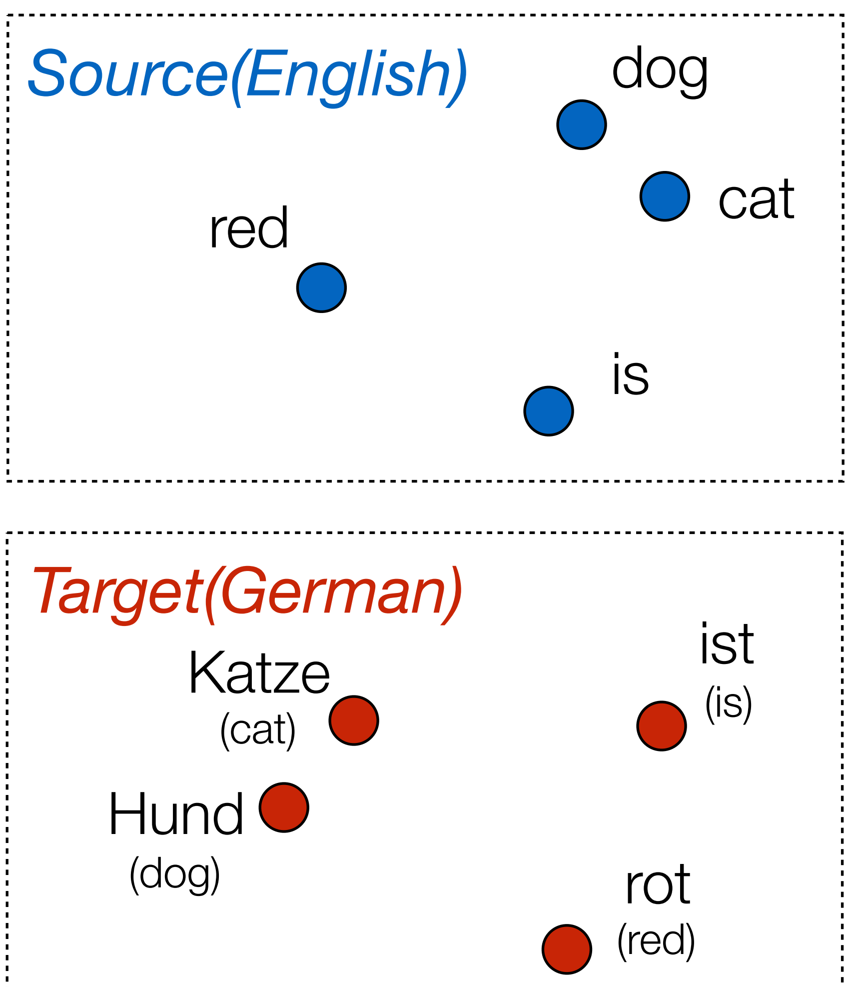
Monolingual Embedding



Coarse Mapping between Embeddings

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

Monolingual Embedding



Translation Pairs

dog || Hund

cat || Katze

red || rot

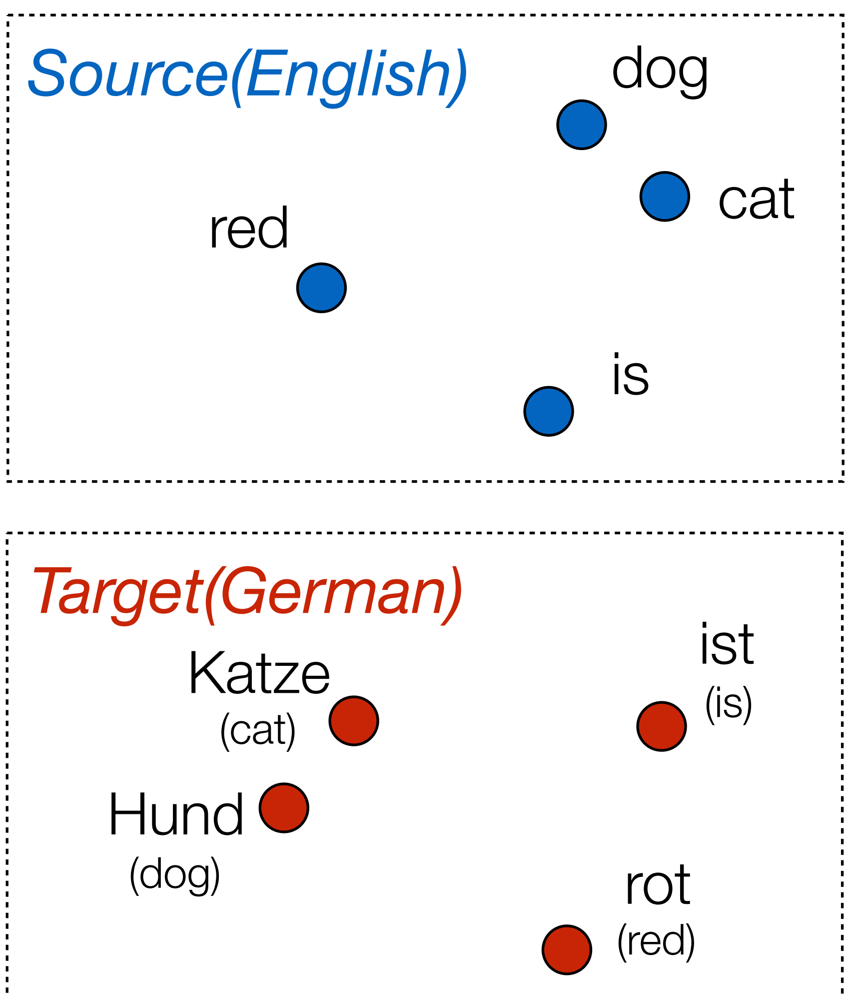
Too many degrees of freedom

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

Monolingual Embedding



Translation Pairs

dog || Hund

cat || Katze

red || rot

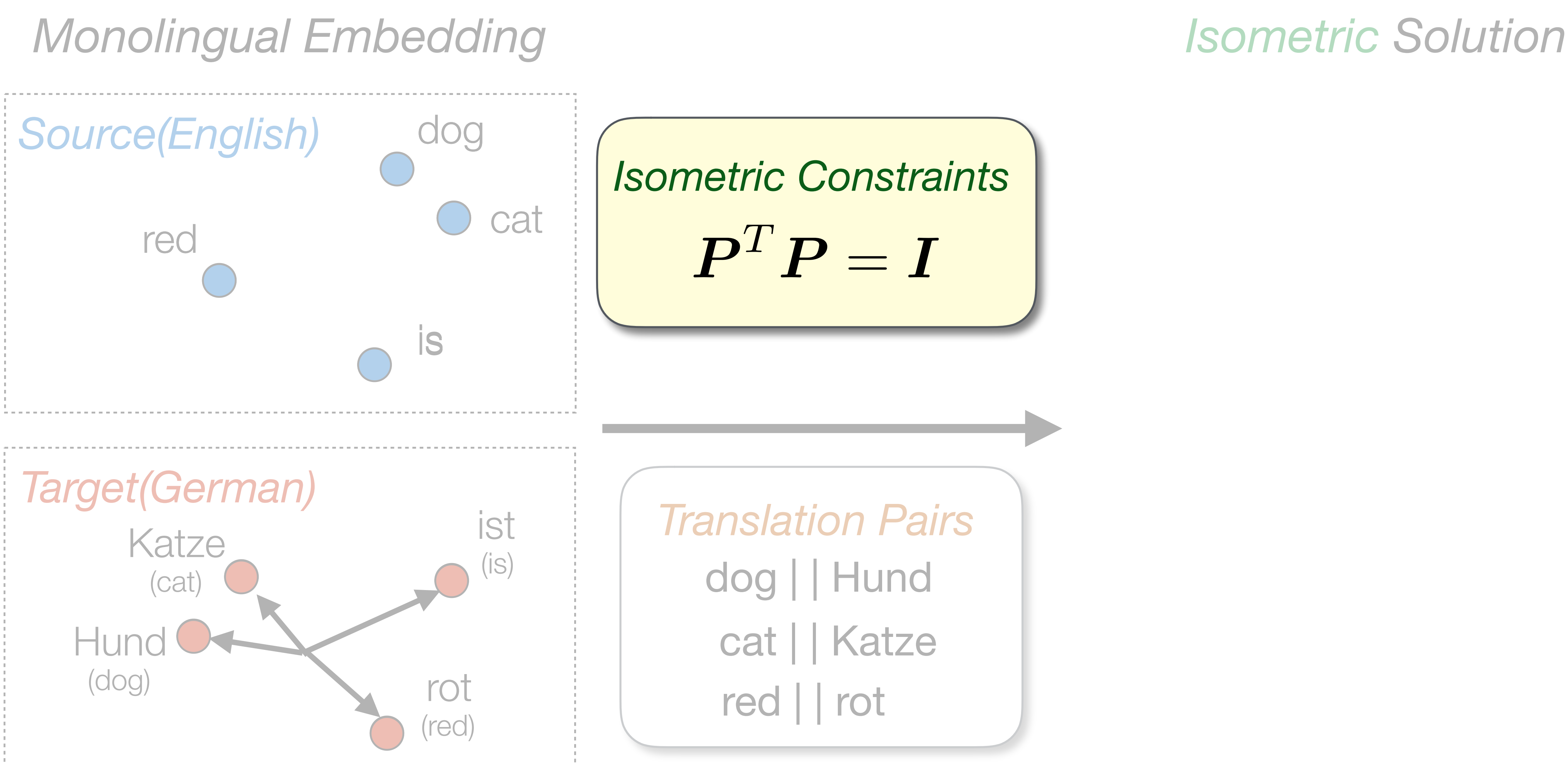
Too many degrees of freedom

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

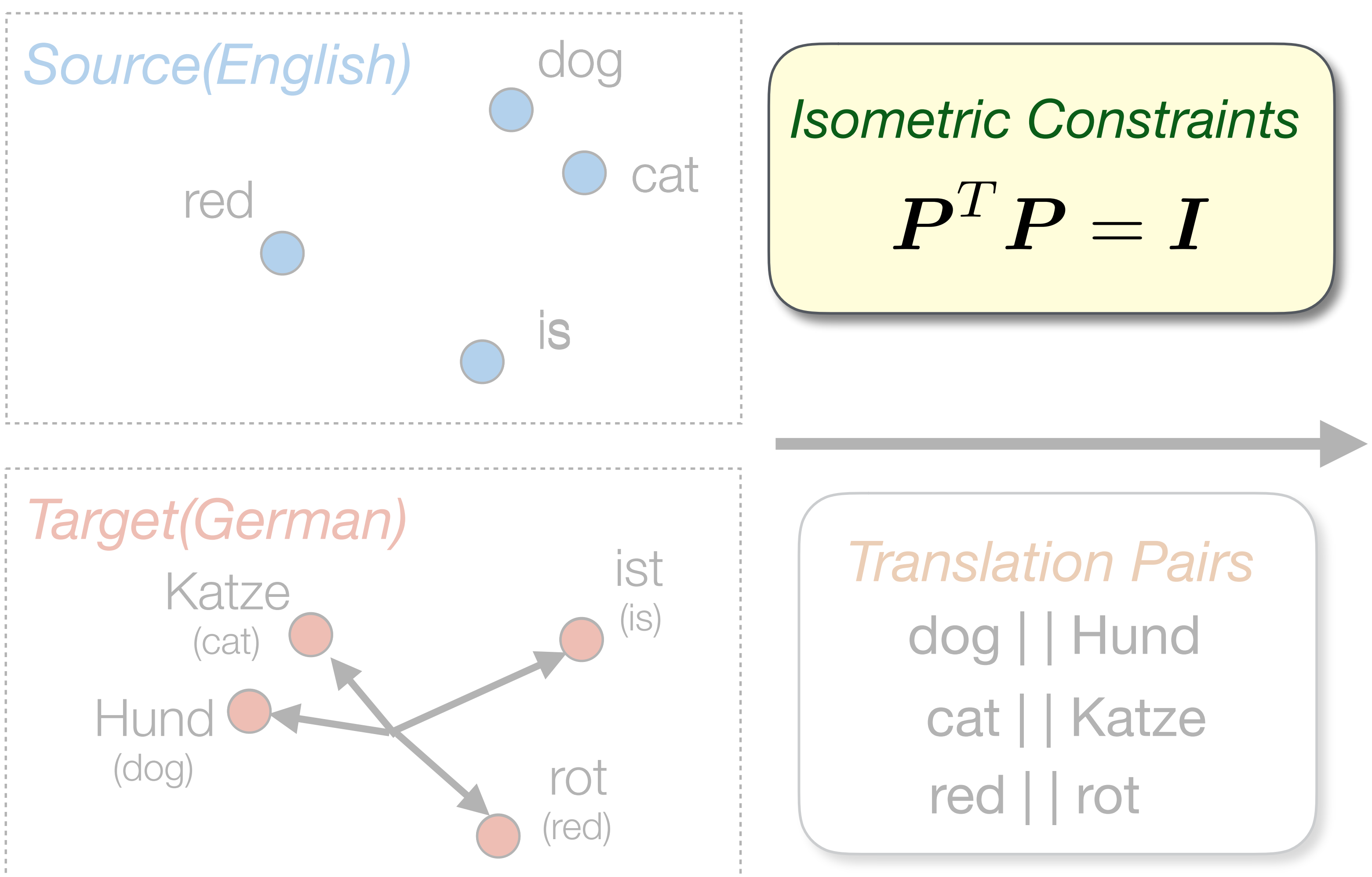


Our Solution: Isometric Constraints

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

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

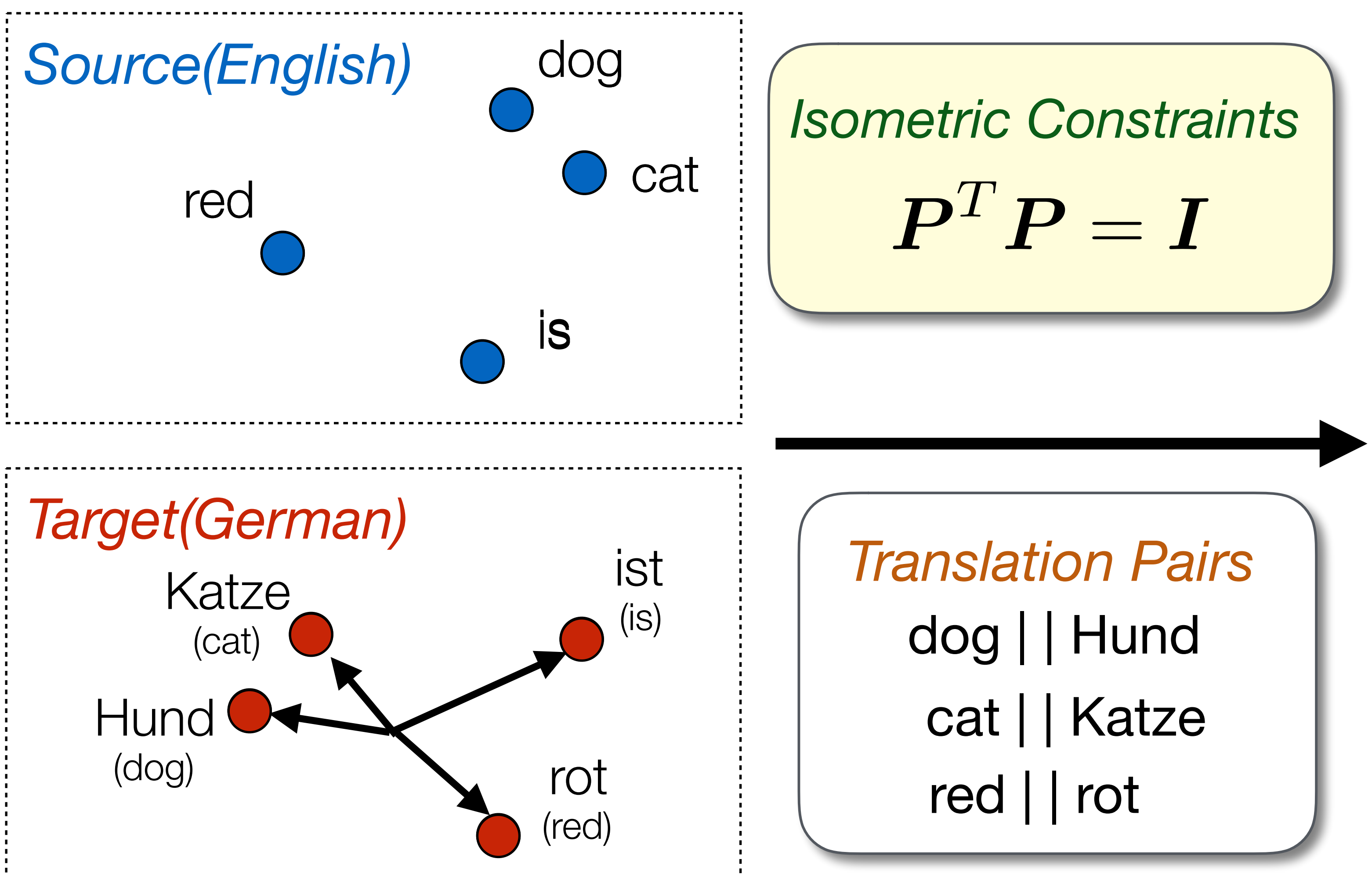
Monolingual Embedding *Isometric Solution*



Our Solution: Isometric Constraints

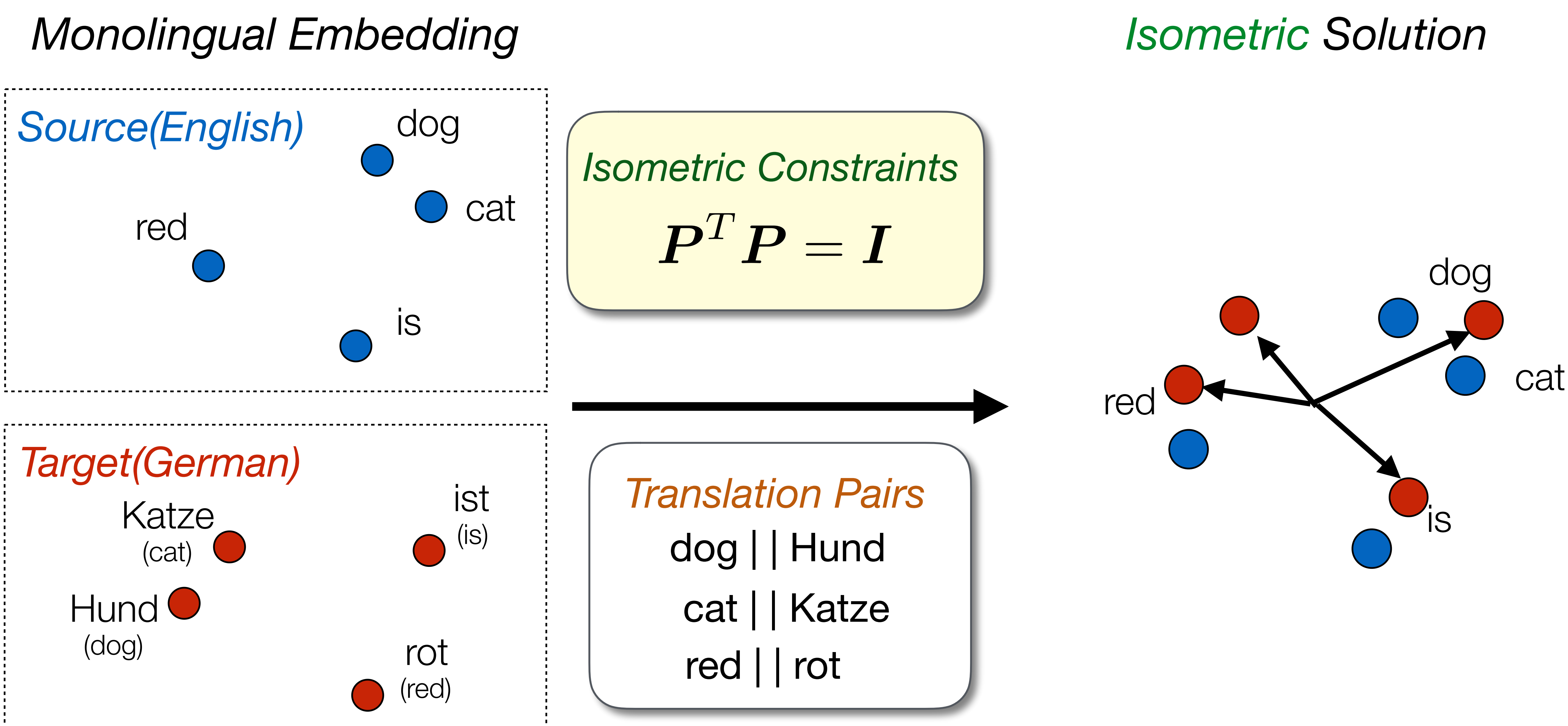
- Transformation P is an isometric (orthonormal) matrix
- Transformation preserves angles and lengths (cosine similarity) of word vectors, thus preserving semantic relations

Monolingual Embedding *Isometric Solution*



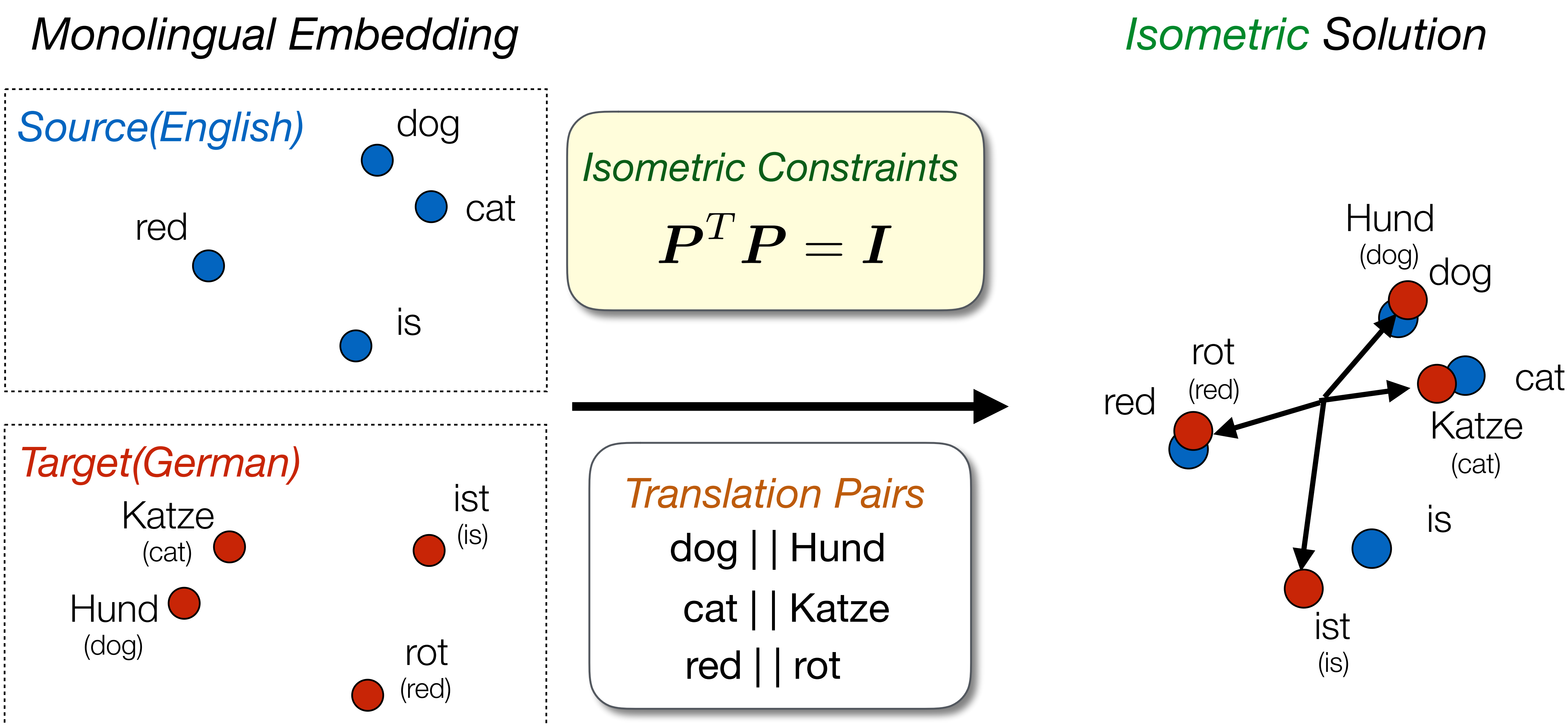
Our Solution: Isometric Constraints

- 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)

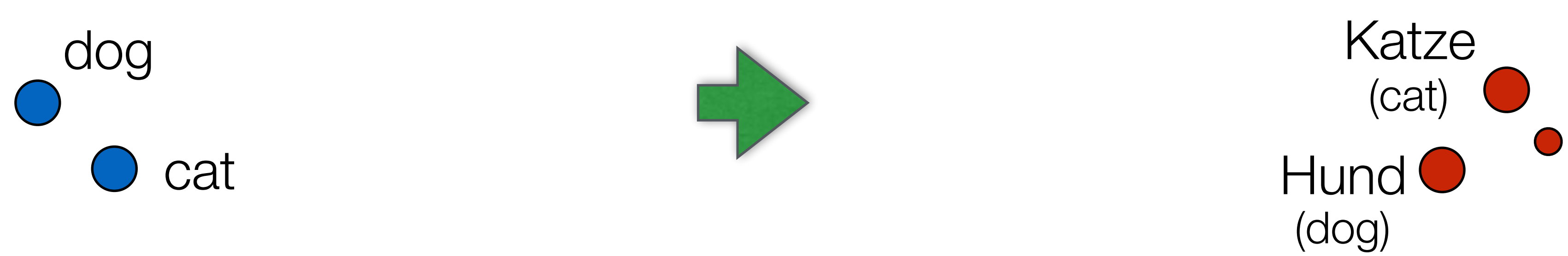


Validation of Isometric Constraints

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

English: nearest neighbor

German: k-th (k≤2) nearest neighbor?

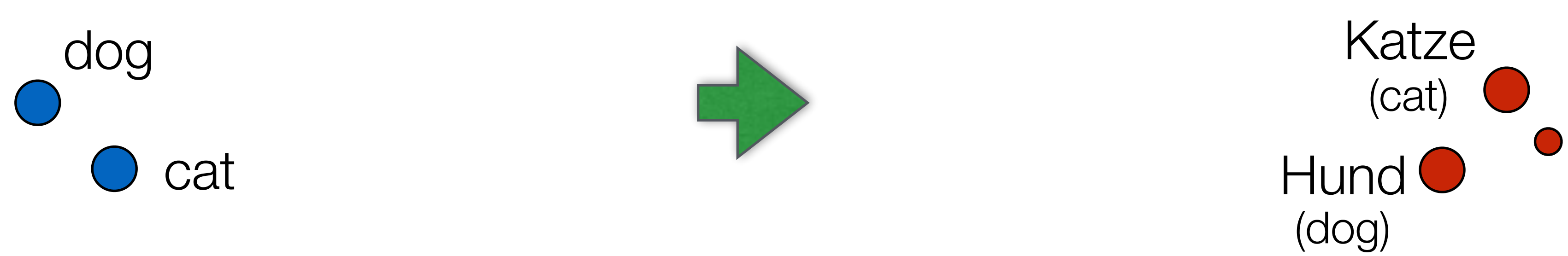


✦ For 50% of word pairs, $k \leq 2$

Validation of Isometric Constraints

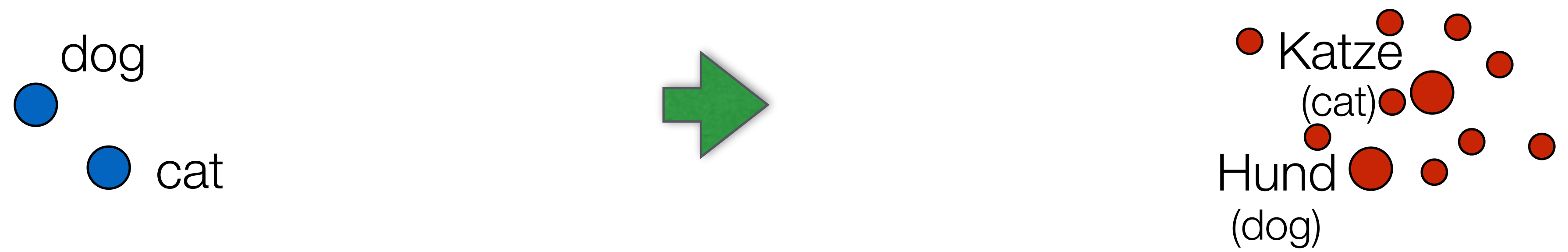
- Validation for $\cos\langle\text{cat, dog}\rangle \approx \cos\langle\text{Katze, Hund}\rangle$
- Verify whether nearest neighbors are preserved after translations

English: nearest neighbor *German: k-th (k≤2) nearest neighbor?*



✦ For 50% of word pairs, $k \leq 2$

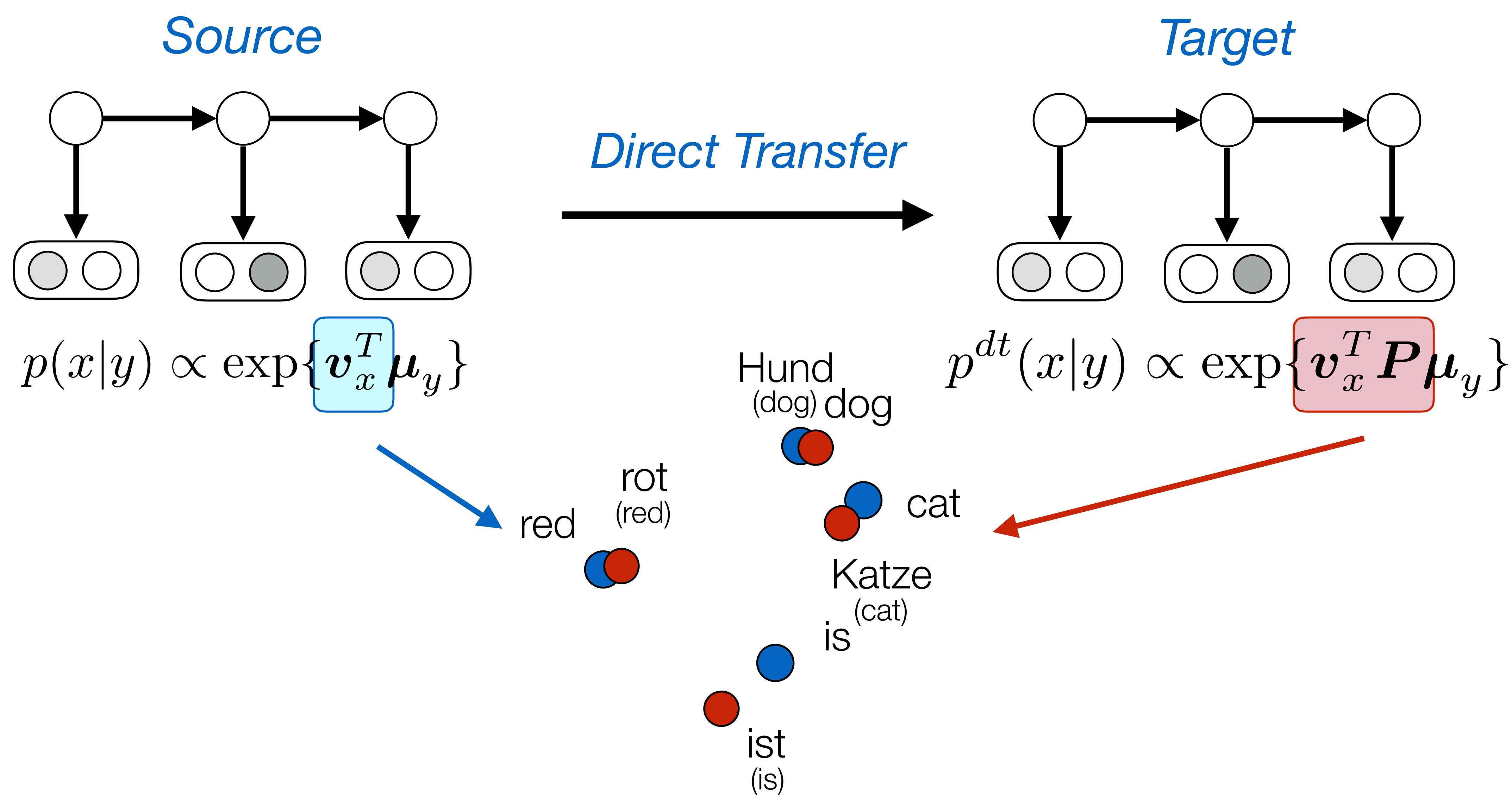
English: nearest neighbor *German: k-th (k≤10) nearest neighbor?*



✦ For 90% of word pairs, $k \leq 10$

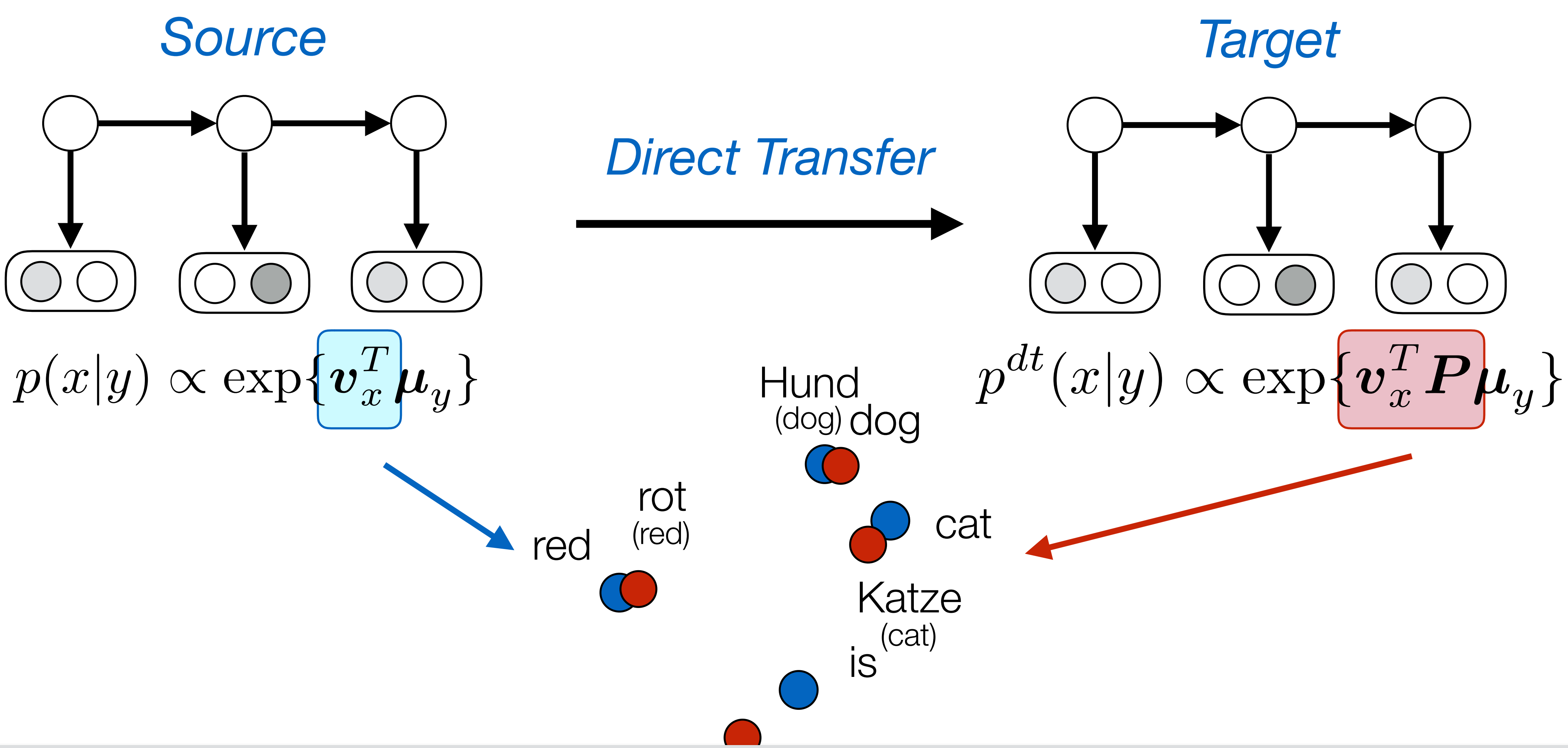
Direct Transfer Model

- 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



Coarse mapping is not accurate

Our Two-step 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

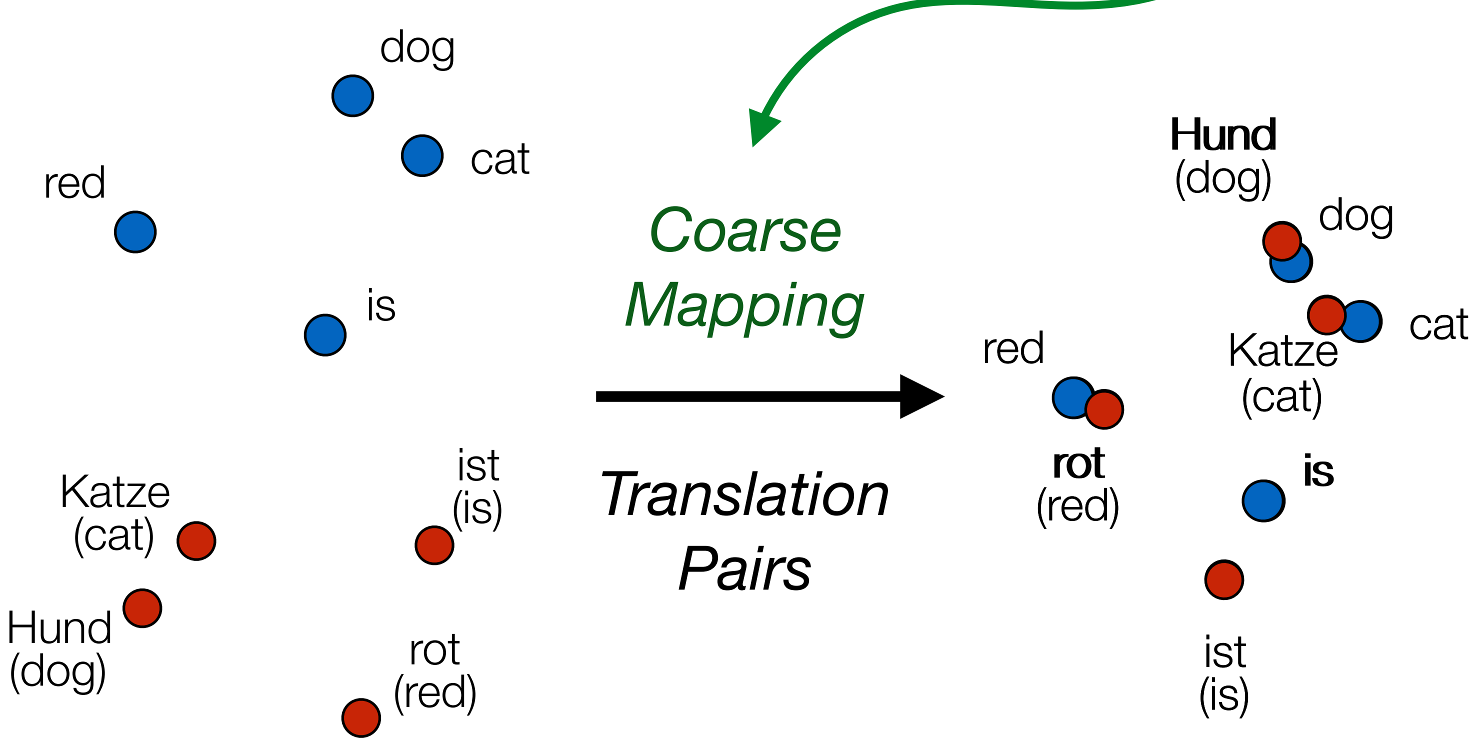
- Use the direct transfer model (based on the coarse mapping) to initialize and regularize the unsupervised tagger on the target language
- Refine mapping via global linear transformation M and local non-linear adjustment $\theta_{x,y}$

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

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 $\theta_{x,y}$

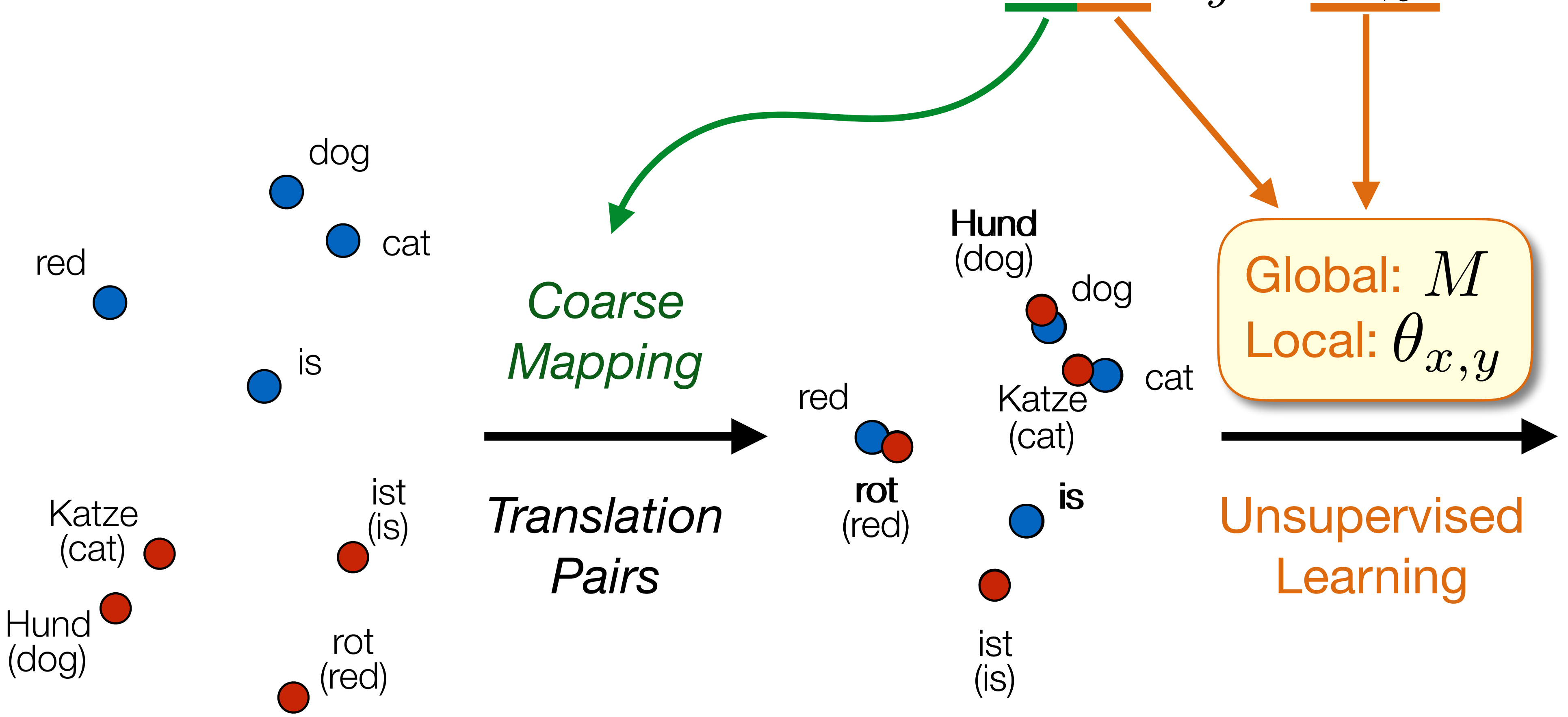
$$p(x|y) \propto \exp\{\boldsymbol{v}_x^T \underline{\boldsymbol{P}} \boldsymbol{M} \boldsymbol{\mu}_y + \theta_{x,y}\}$$



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 $\theta_{x,y}$

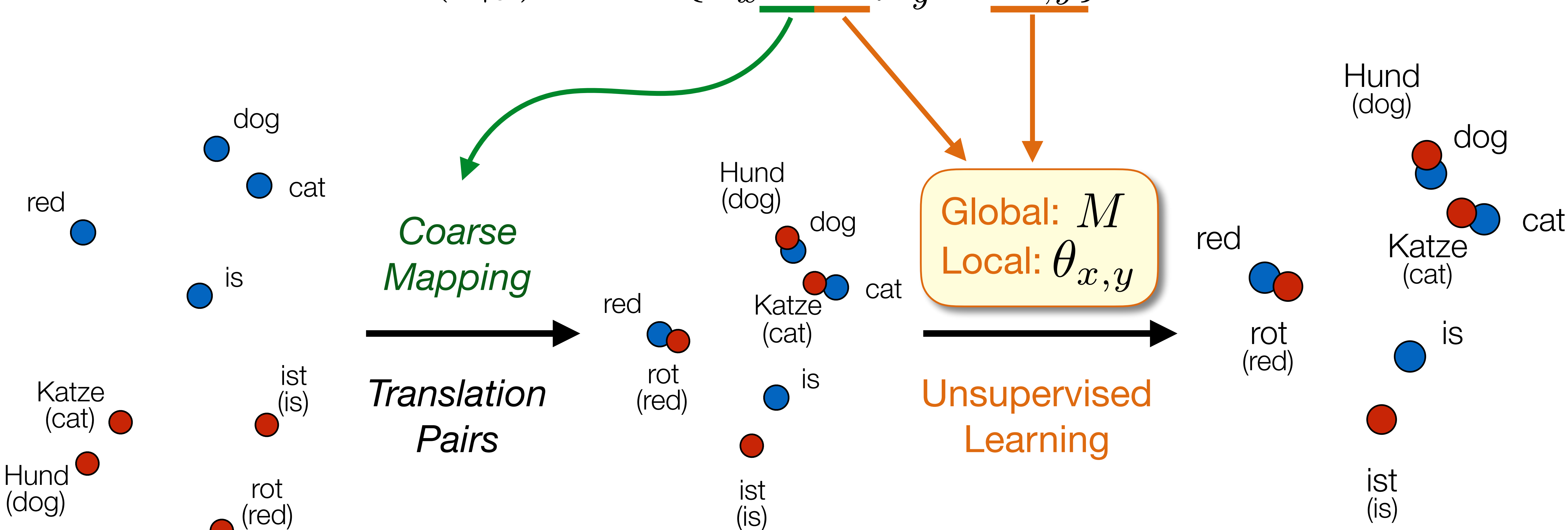
$$p(x|y) \propto \exp\{\boldsymbol{v}_x^T \boldsymbol{P}\boldsymbol{M}\boldsymbol{\mu}_y + \theta_{x,y}\}$$



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 $\theta_{x,y}$

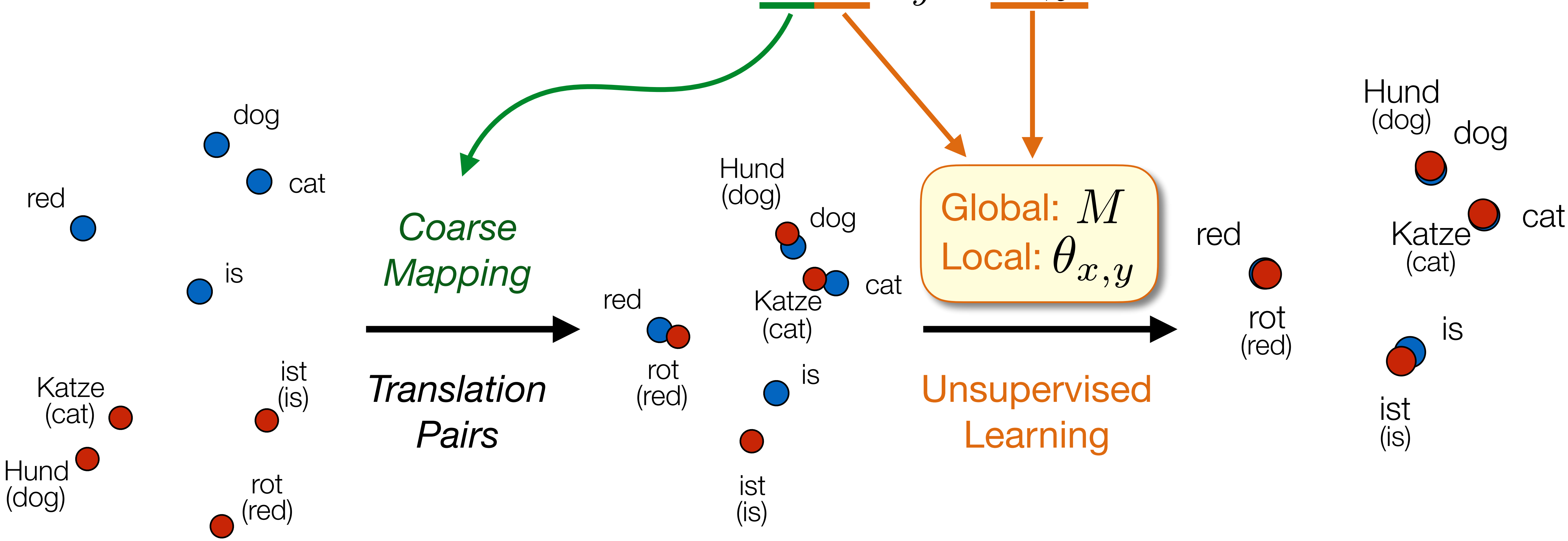
$$p(x|y) \propto \exp\{\boldsymbol{v}_x^T \boldsymbol{P}\boldsymbol{M}\boldsymbol{\mu}_y + \theta_{x,y}\}$$



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 $\theta_{x,y}$

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



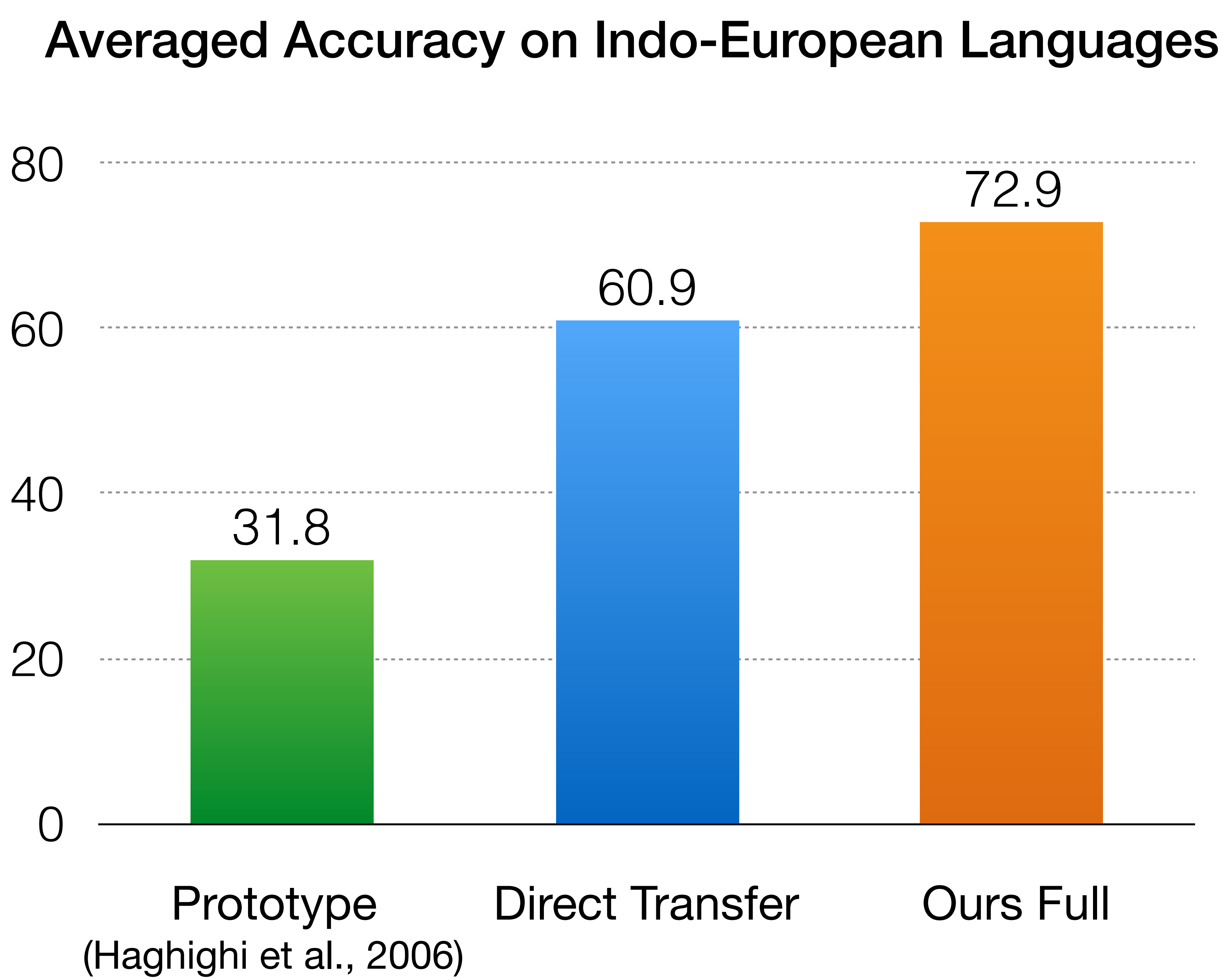
Experimental Setup

- Datasets: Universal Dependency Treebank v1.2
  - ✦ Source: English
  - ✦ Target (Indo-European): Danish, German, Spanish
  - ✦ Target (non-Indo-European): Finnish, Hungarian, Indonesian

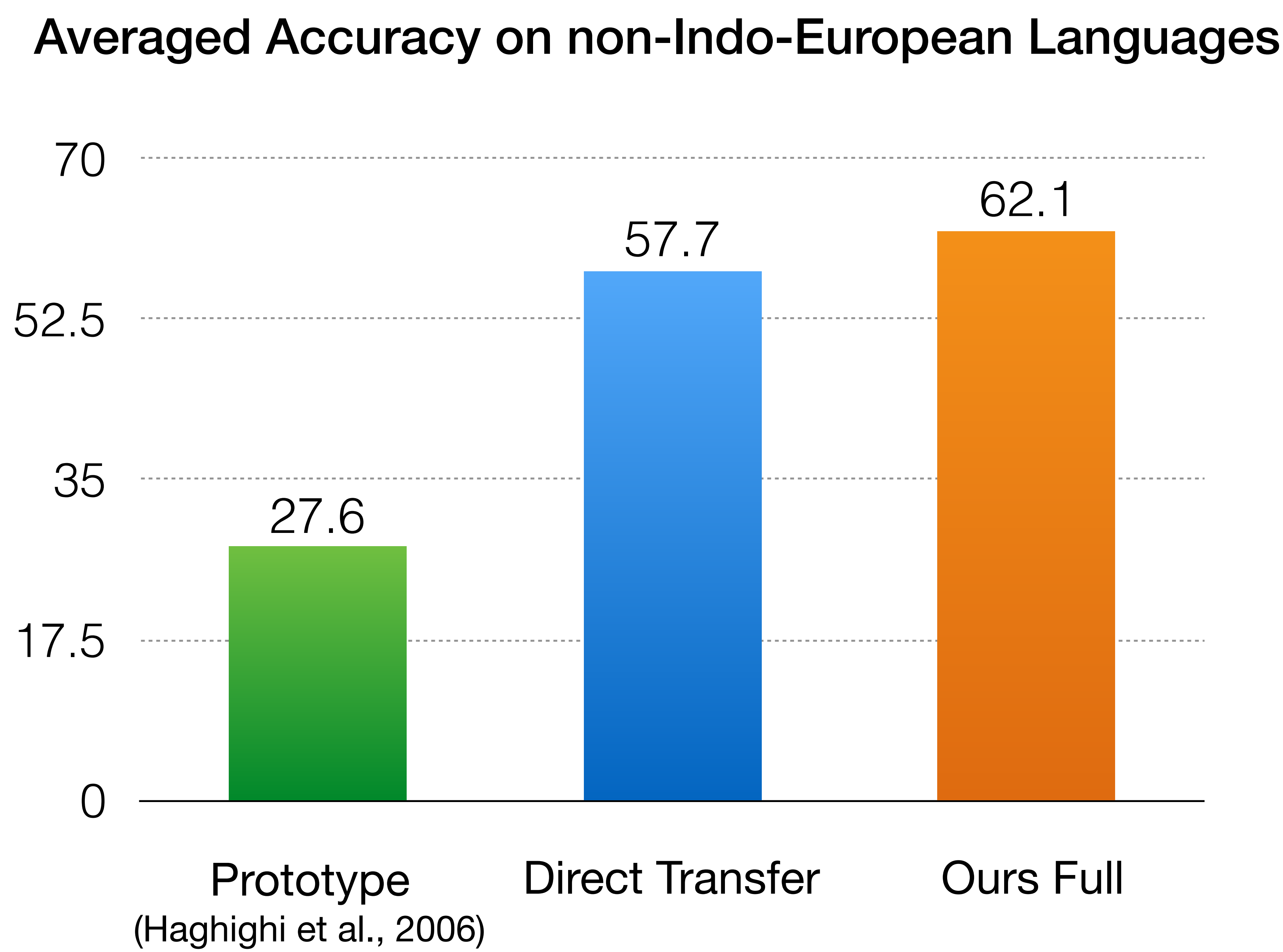
- Universal tagset: 14 tags (noun, verb, adjective etc.)

- Word embeddings: 20-dimension vectors trained on Wiki dumps using word2vec

Indo-European Results

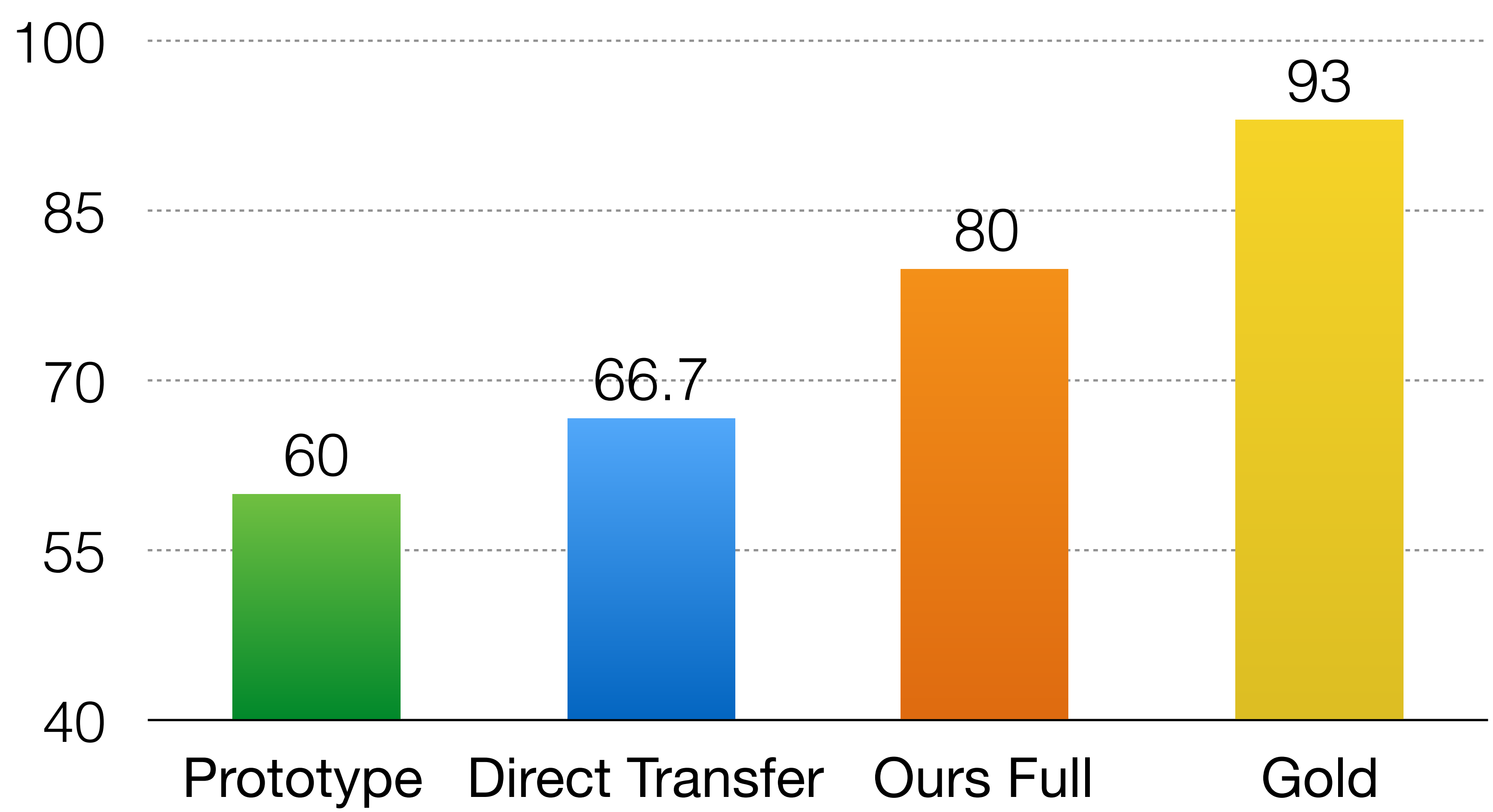


Non-Indo-European Results



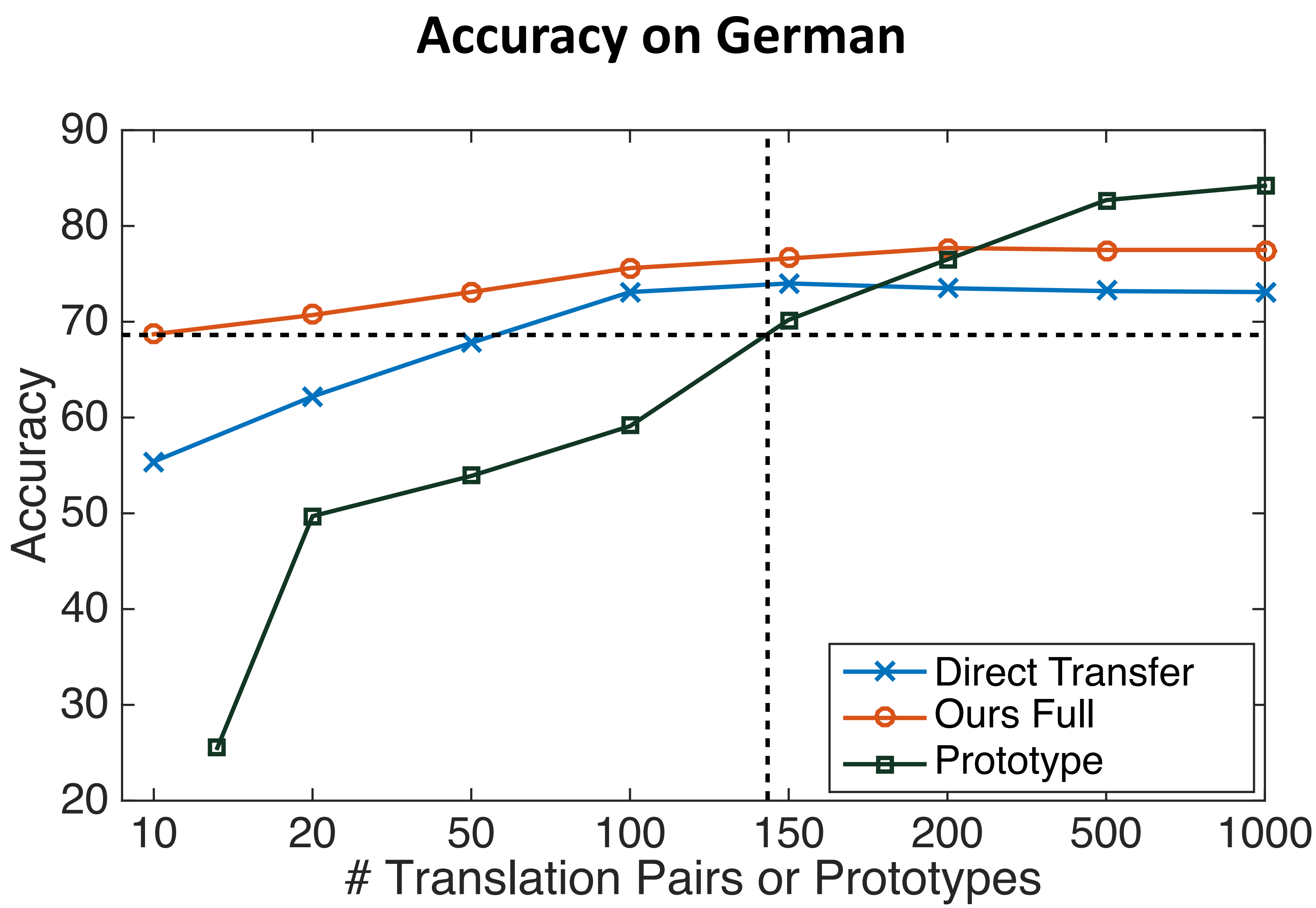
Prediction of Linguistic Typology

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



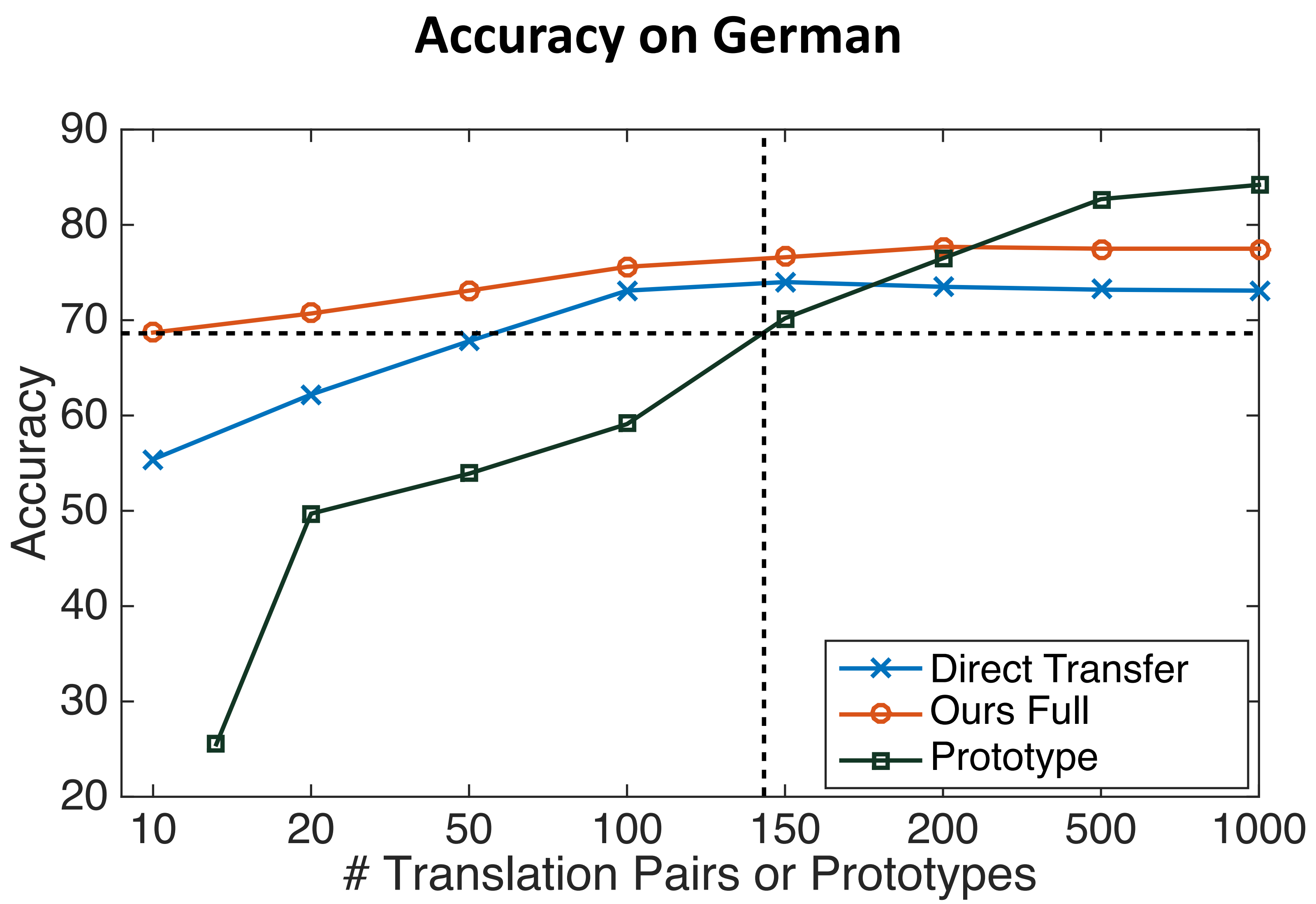
Impact of Amount of Supervision

- Ours Full with 10 pairs = 150 prototypes



Impact of Amount of Supervision

- Ours Full with 10 pairs = 150 prototypes
- Prototype improves with large amount of annotations



Summary

- *Modeling:* ten translation pairs are sufficient to enable multilingual transfer for POS tagging

- *Performance:* our model significantly outperforms the direct transfer and the prototype-driven method

Our Approach

Multilingual Transfer:

- Hierarchical tensors for dependency parsing

- Multilingual embeddings for POS tagging

Monolingual Transfer:

- Adversarial networks for aspect transfer
  - *Joint aspect-driven encoding and domain adversarial training*

Aspect Transfer in Pathology Report

Pathology report:

FINAL DIAGNOSIS: BREAST (LEFT) … **INVASIVE DUCTAL CARCINOMA (IDC) Tumor size: num x num x num cm Grade: 3. Lymphatic vessel invasion (LVI): Not identified.** Blood vessel invasion: Suspicious. Margin of invasive carcinoma …

Diagnosis results:

IDC: Positive **LVI: Negative**

Transfer:

Source: IDC → Target: LVI

Challenge

Same report; Different key sentences

Source Aspect: IDC

Target Aspect: LVI

FINAL DIAGNOSIS: BREAST (LEFT) … **INVASIVE DUCTAL CARCINOMA (IDC) Tumor size: num x num x num cm Grade: 3. Lymphatic vessel invasion (LVI): Not identified.** Blood vessel invasion: Suspicious. Margin of invasive carcinoma …

- Traditional methods will fail because they always induce the same representation for the same input

Available Supervision

	Source	Target
Labeled Data	✓	✗
Unlabeled Data	✓	✓
Relevance Rules	✓	✓

- Relevance rules: common names of aspects
  - ALH: Atypical Lobular Hyperplasia, ALH
  - IDC: Invasive Ductal Carcinoma, IDC

Transfer Assumption: Aspects Are Related

- Different aspects share the same label set: positive/negative

IDC: Positive LVI: Negative

Transfer Assumption: Aspects Are Related

- Different aspects share the same label set: positive/negative

IDC: Positive LVI: Negative

- Common words are directly transferrable

Invasive Carcinoma is present Label: Positive	→	Lymphatic vessel invasion: present Label: Positive

Transfer Assumption: Aspects Are Related

- Different aspects share the same label set: positive/negative

IDC: Positive   LVI: Negative

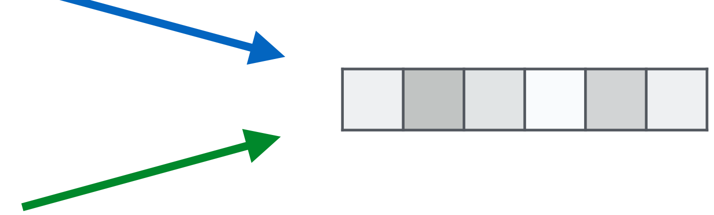
- Common words are directly transferrable

Invasive Carcinoma is present Label: Positive	→	Lymphatic vessel invasion: present Label: Positive

- Aspect-specific words are not directly transferrable
  - Goal: map them to invariant representations

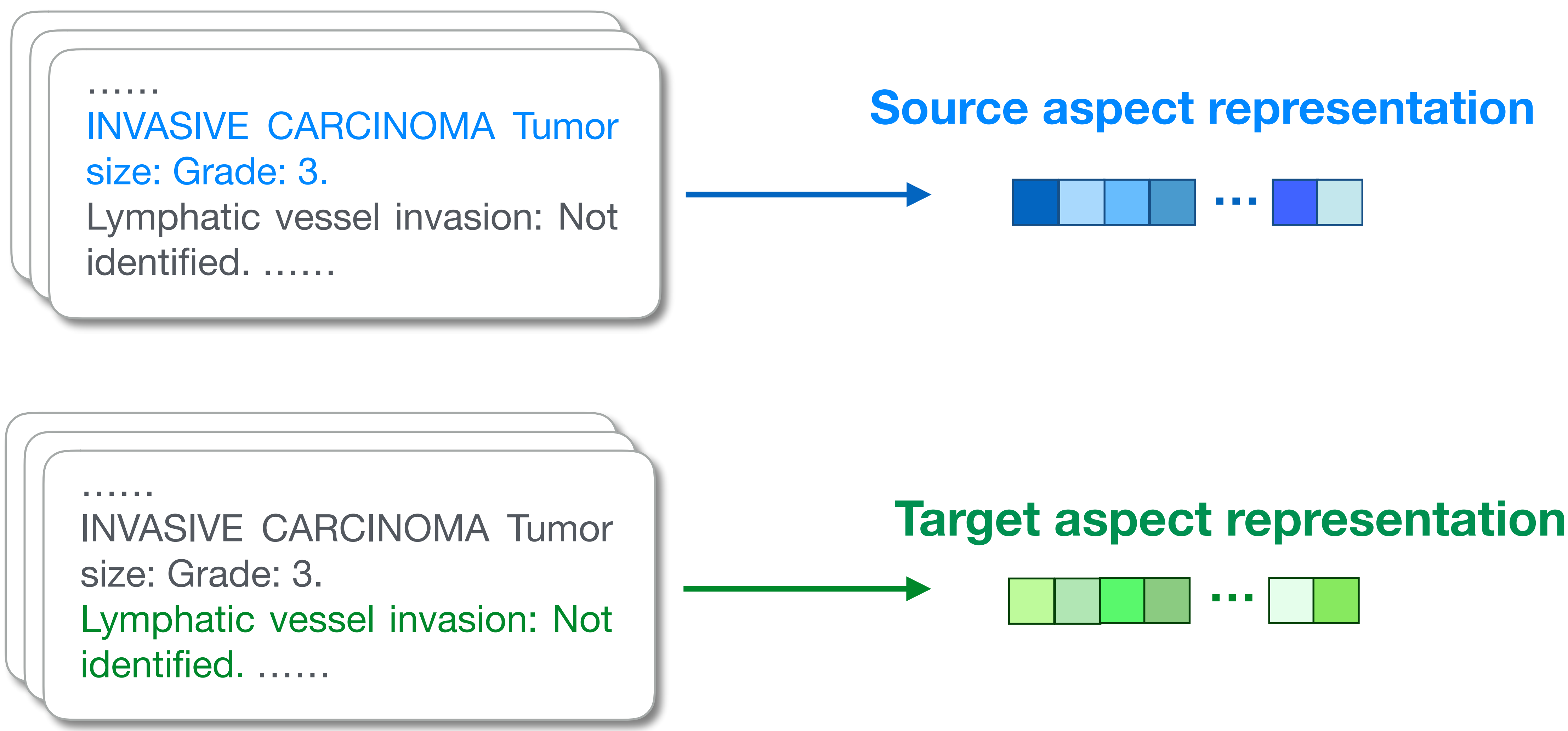
Invasive Ductal Carcinoma

Lymphatic Vessel Invasion



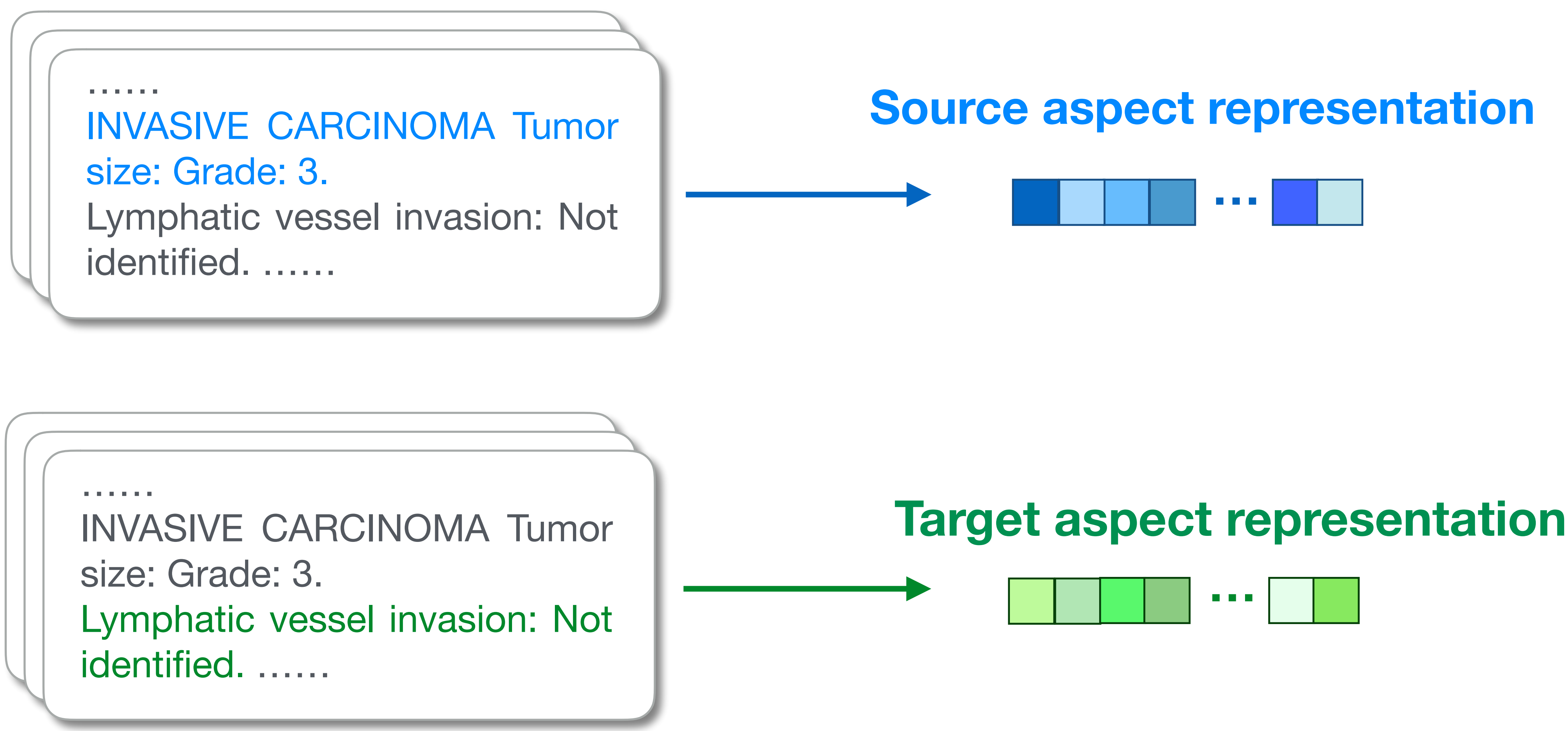
Key Idea: Aspect-driven Encoding

- Leverage relevance rules to learn to identify key sentences
- Learn differential representations for different aspects from the same input



Key Idea: Aspect-driven Encoding

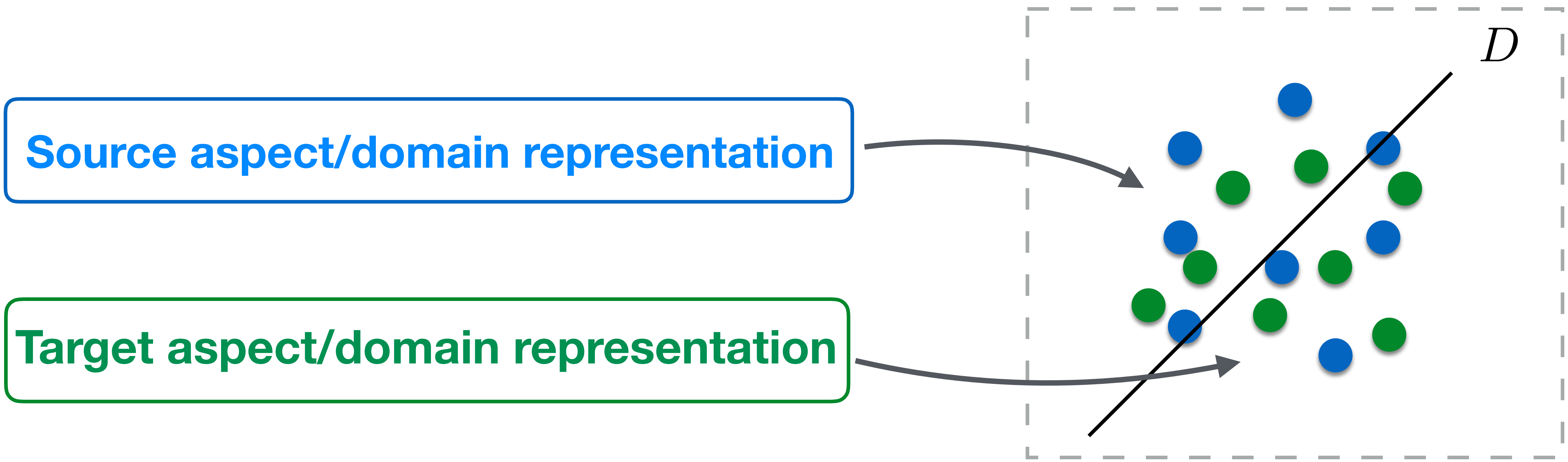
- Leverage relevance rules to learn to identify key sentences
- Learn differential representations for different aspects from the same input



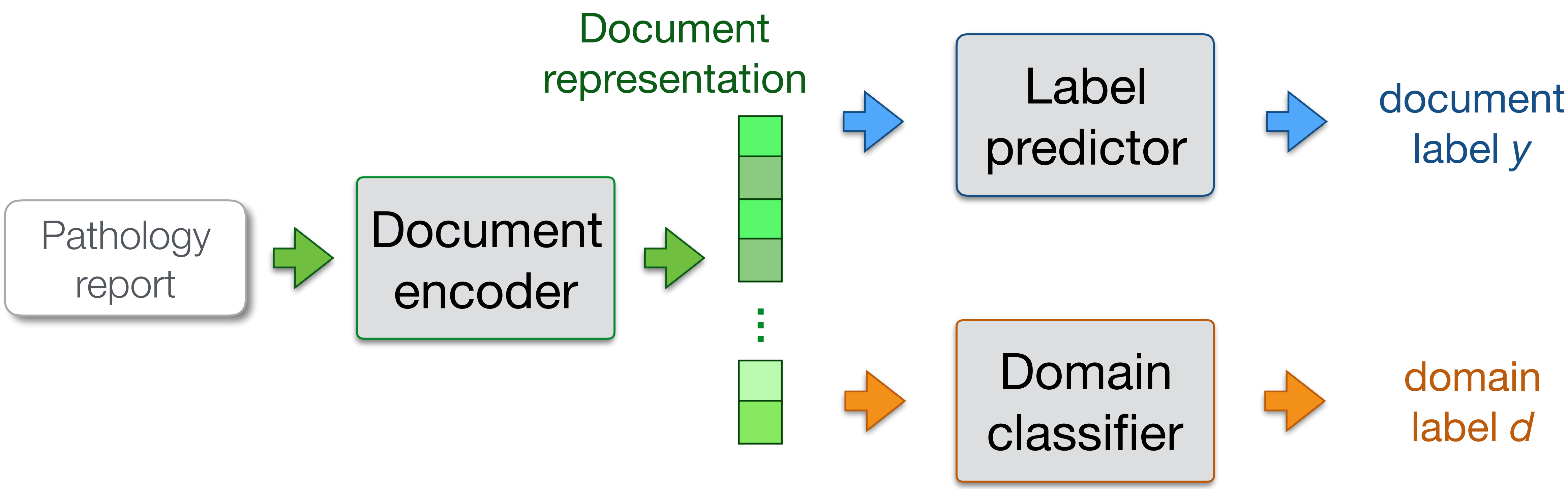
Reduce aspect transfer to standard domain adaptation

Key Idea: Domain-Adversarial

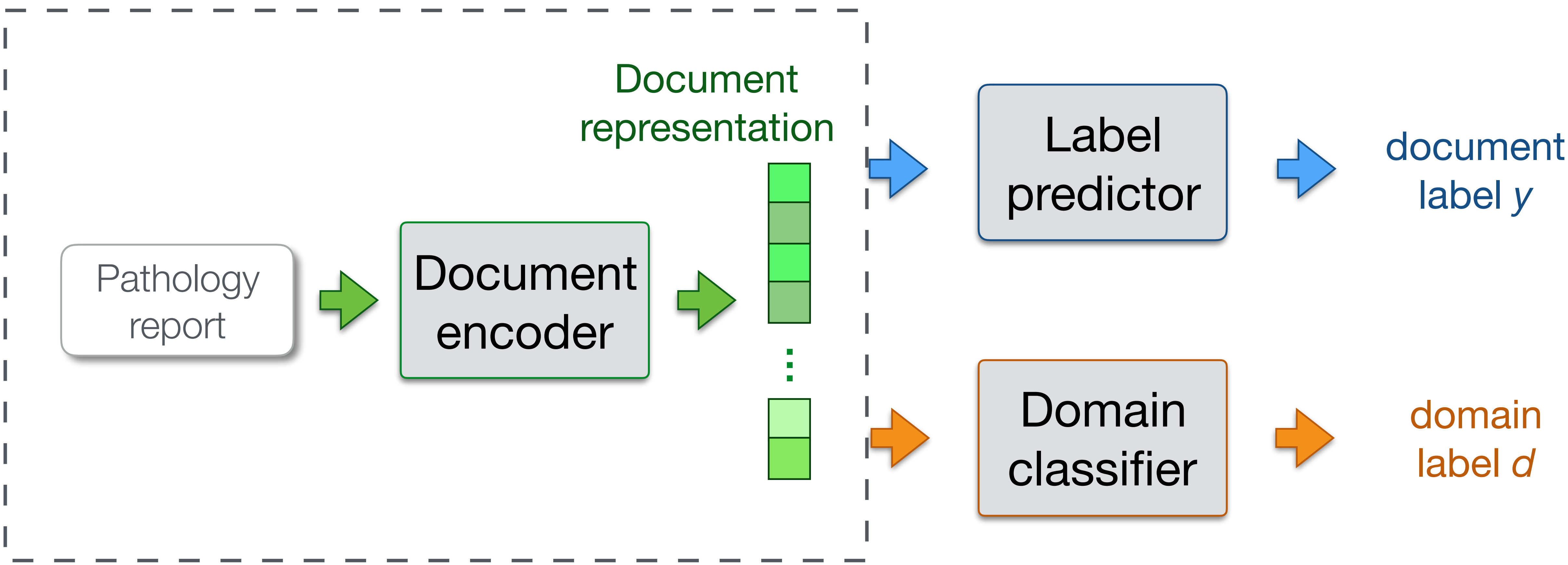
- Jointly train a domain classifier
- Use domain-adversarial training for learning invariant representations
    - Objective: Not separable by the domain classifier



Overall Framework: Three Components

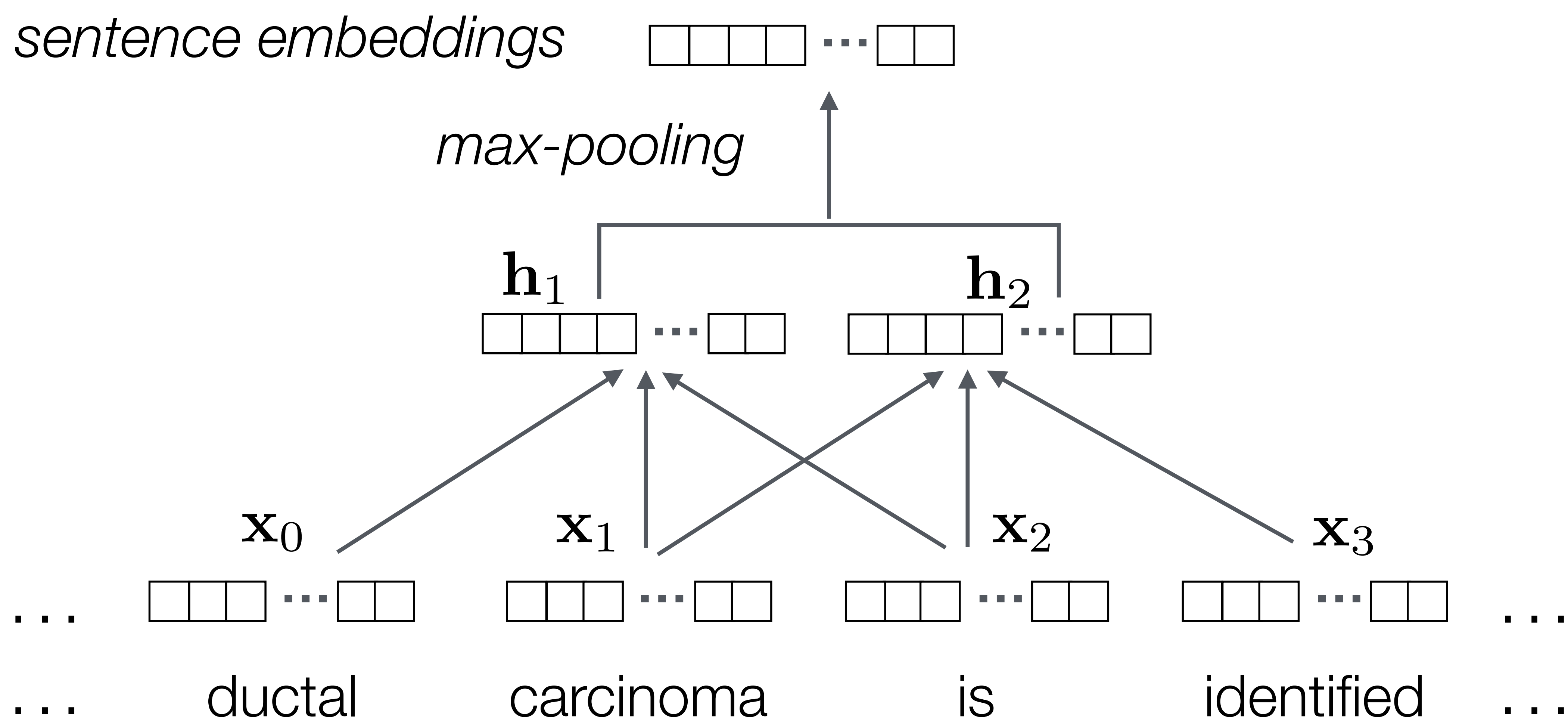


Overall Framework: Three Components



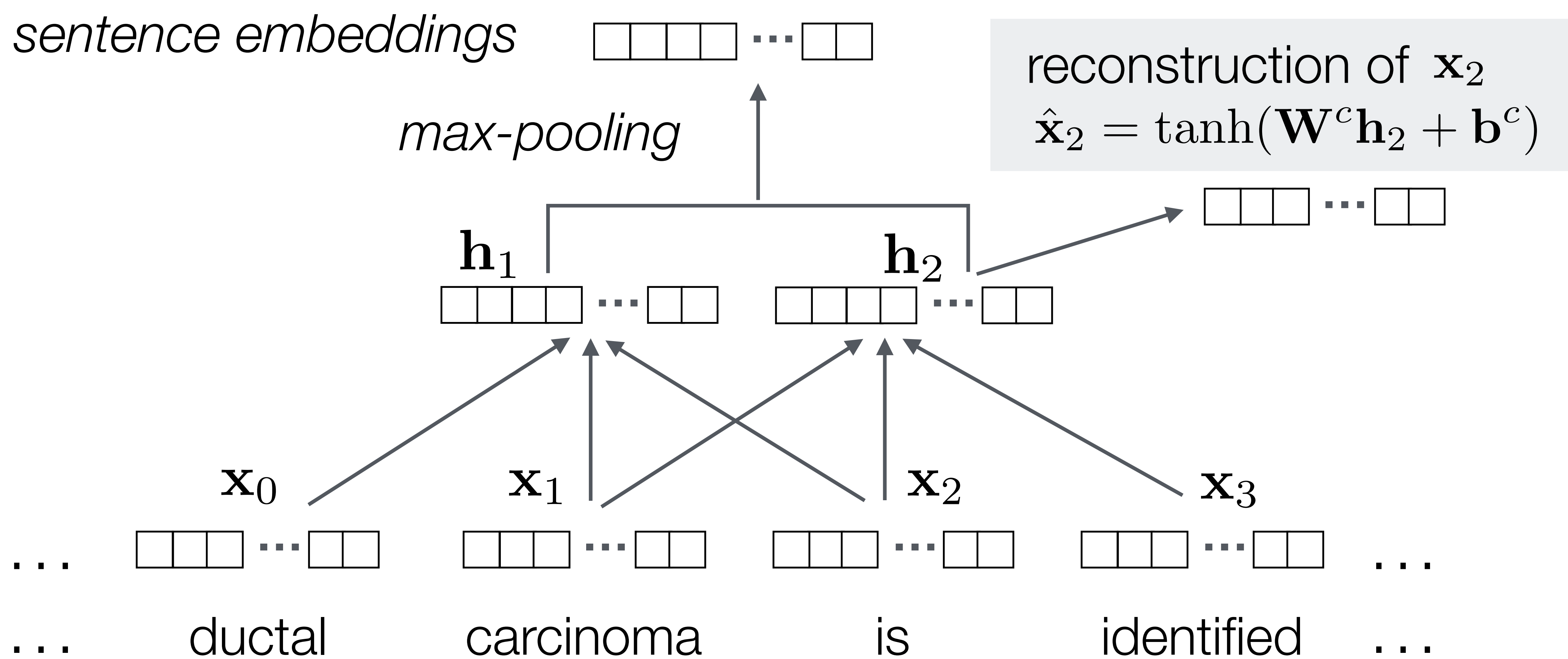
Sentence Embedding

- Apply a CNN to each sentence



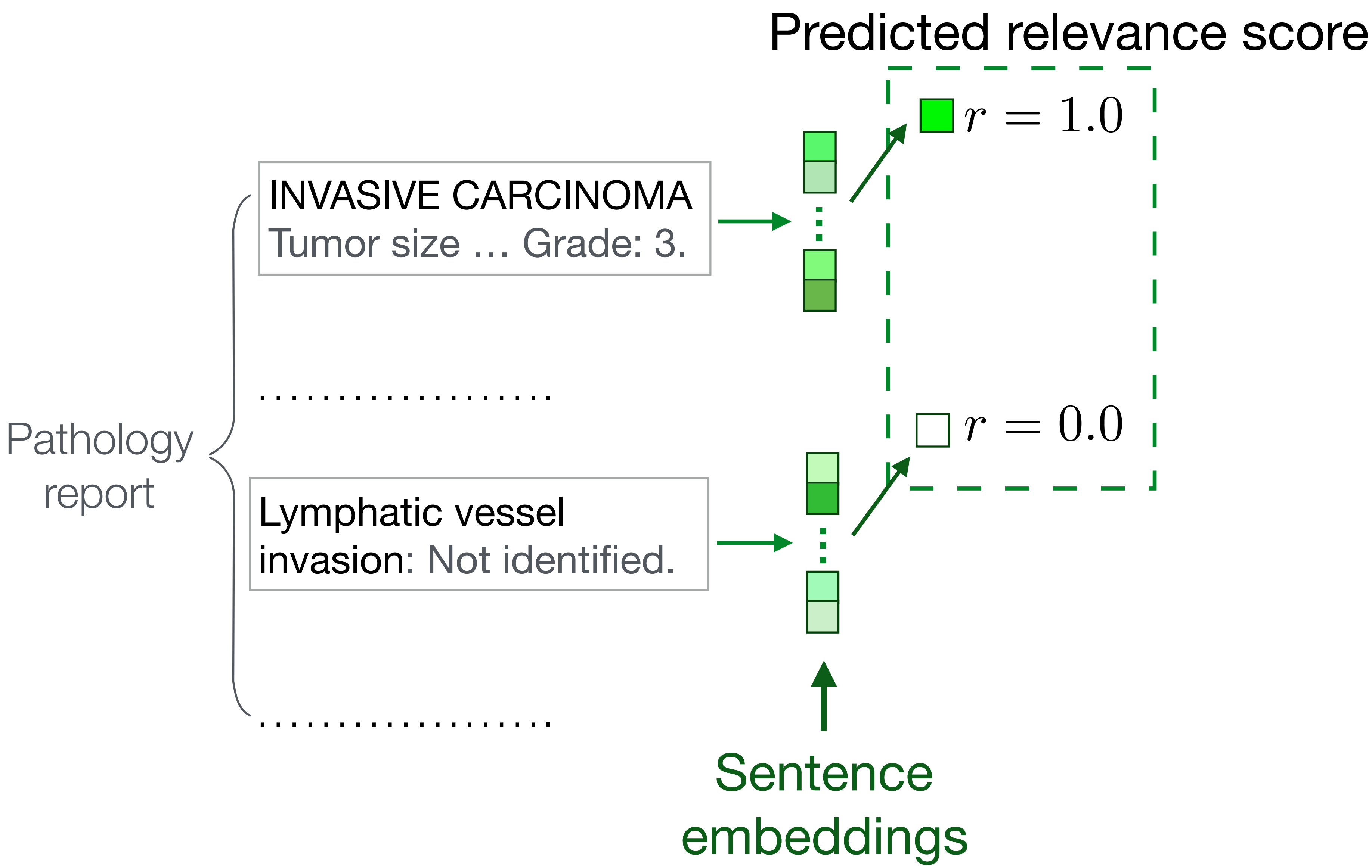
Sentence Embedding

- Apply a CNN to each sentence
- Improve adversarial training by reconstruction



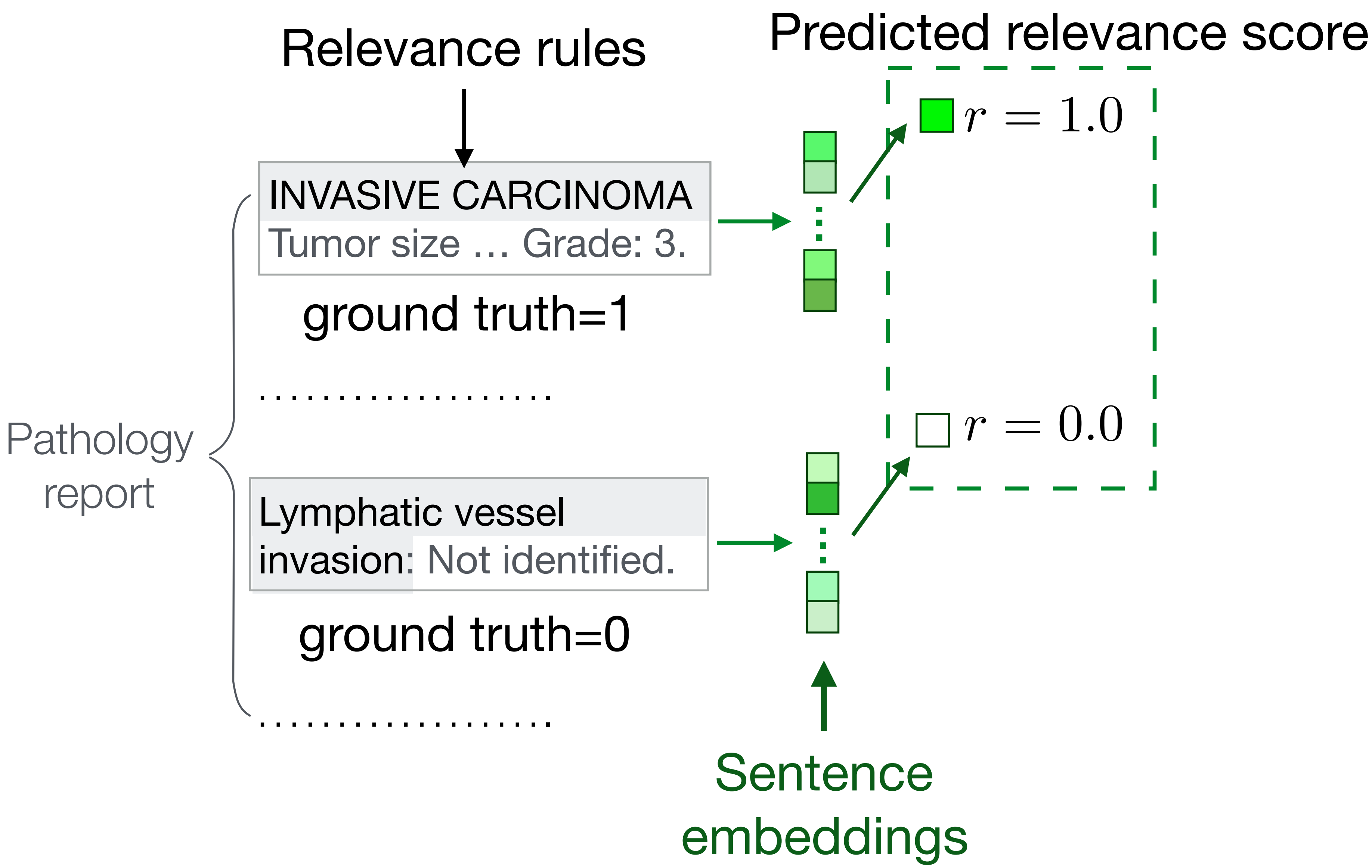
Aspect-relevance Prediction

- Predict relevance score based on sentence embeddings
- Train on relevance rules (e.g., names of IDC, LVI)



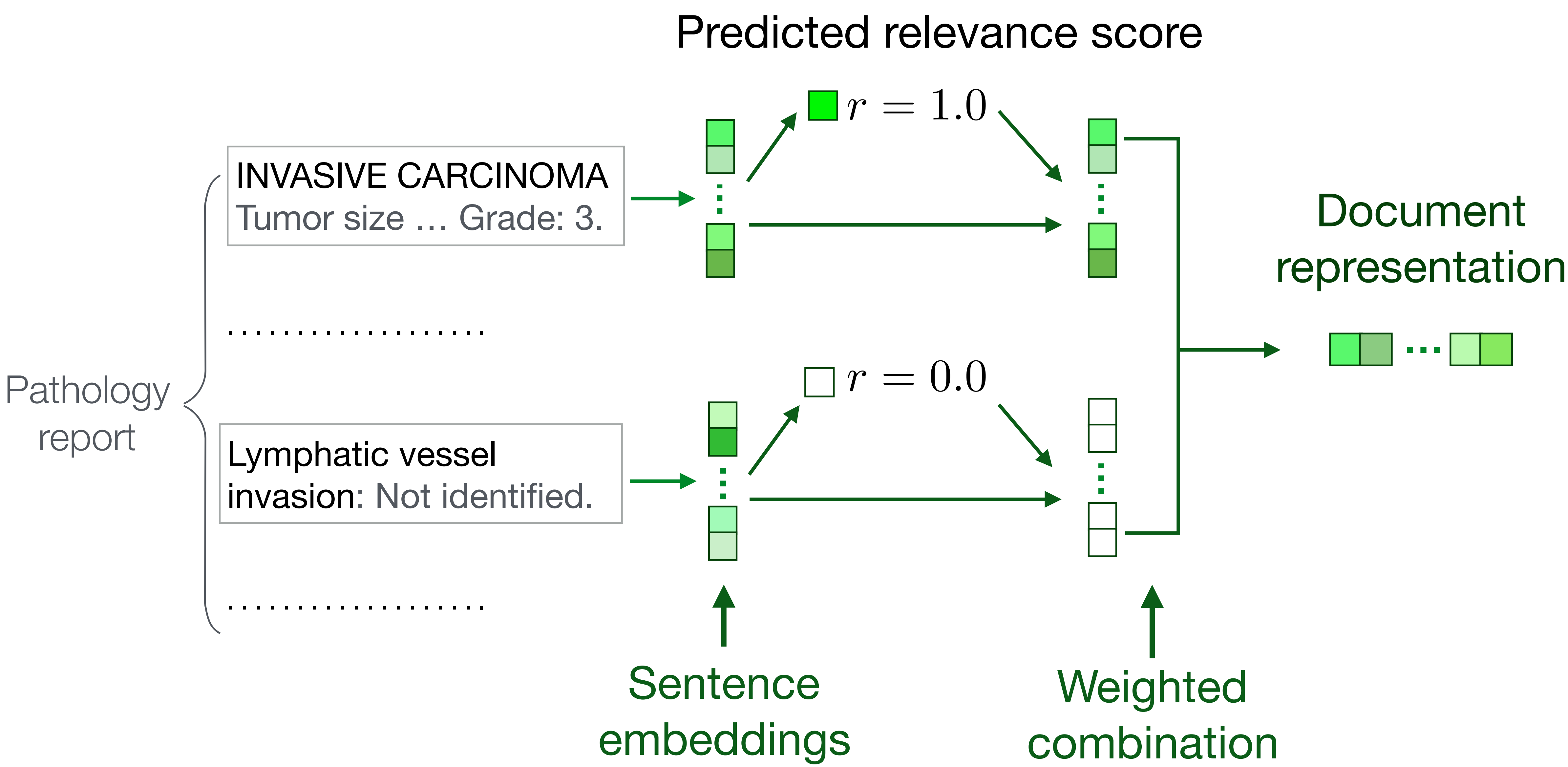
Aspect-relevance Prediction

- Predict relevance score based on sentence embeddings
- Train on relevance rules (e.g., names of IDC, LVI)



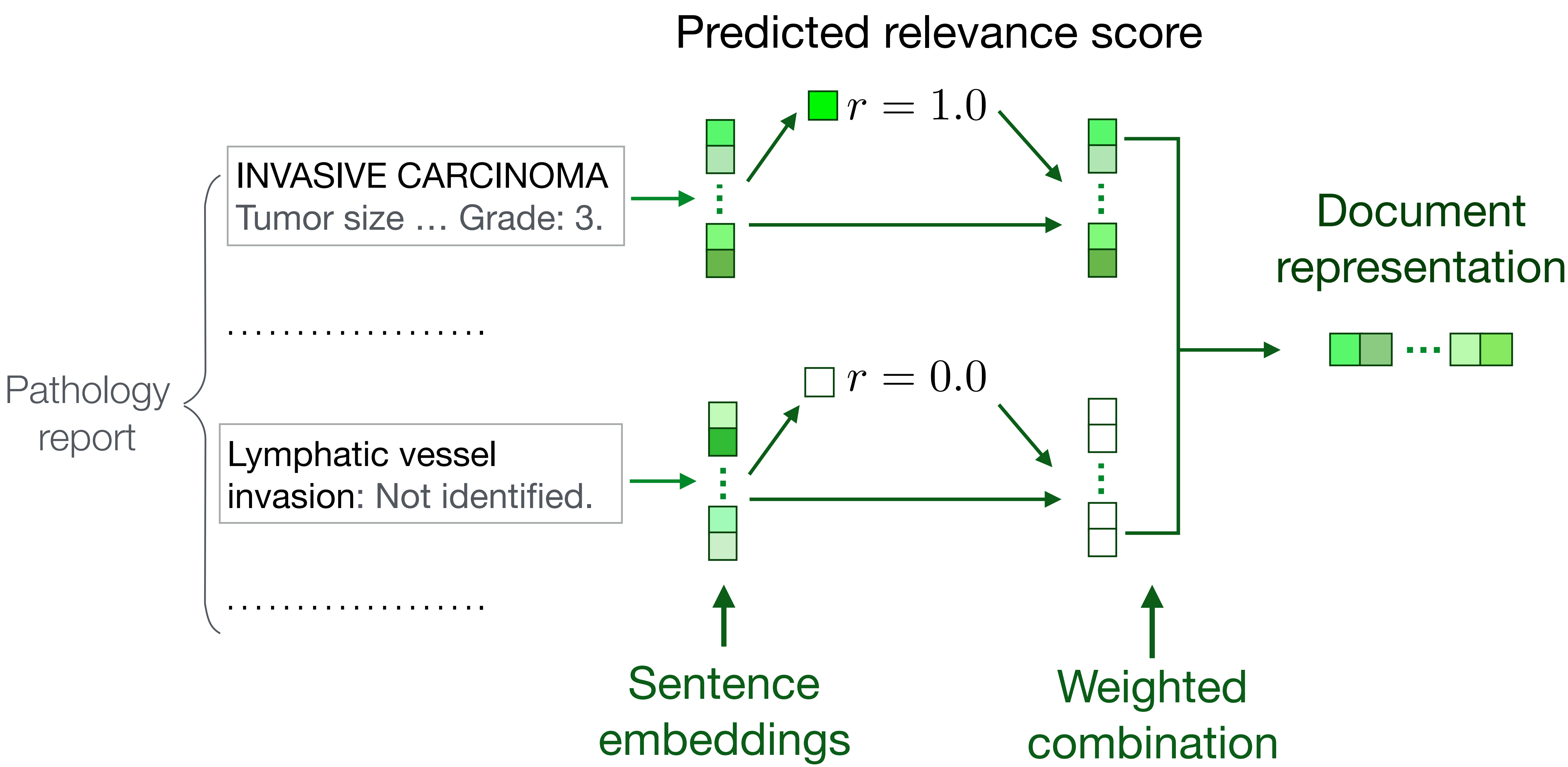
Aspect-driven Document Encoding

- Combine sentence vectors based on relevance weights



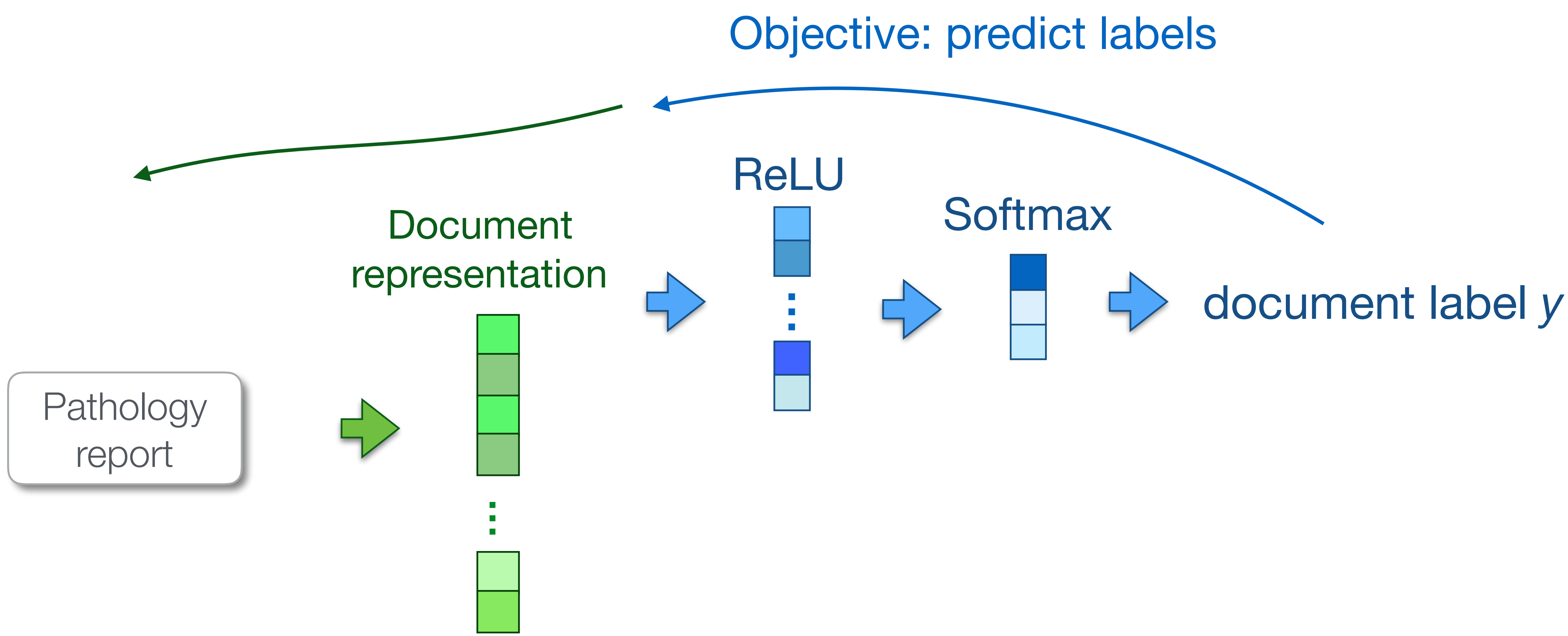
Aspect-driven Document Encoding

- Combine sentence vectors based on relevance weights
- Add a transformation layer at the end



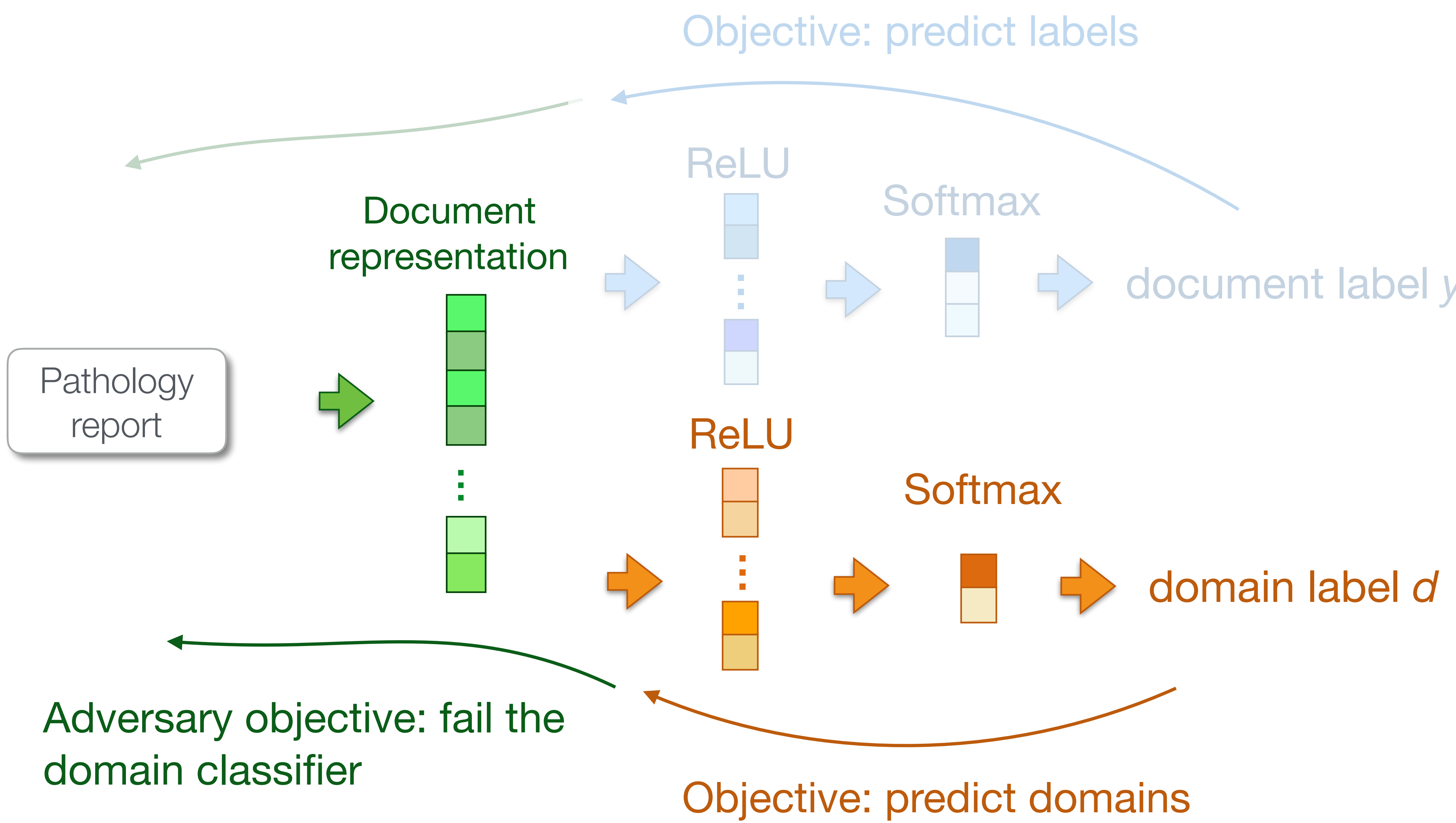
Document Label Predictor

- Share for both source and target aspects
- Train on labeled data in the source aspect



Domain Classifier and Adversary

- Learn domain-invariant representations
- Train on both labeled and unlabeled data



Pathology Dataset

- Aspect-transfer on breast cancer pathology reports from hospitals such as MGH

Source: IDC → Target: LCIS

FINAL DIAGNOSIS: BREAST (LEFT) … **INVASIVE DUCTAL CARCINOMA Grade: 3. Lobular Carcinoma In-situ: Not identified.** Blood vessel invasion: Suspicious. …

Pathology Dataset

- Aspect-transfer on breast cancer pathology reports from hospitals such as MGH

Source: IDC → Target: LCIS

FINAL DIAGNOSIS: BREAST (LEFT) … **INVASIVE DUCTAL CARCINOMA Grade: 3.** **Lobular Carcinoma In-situ: Not identified.** Blood vessel invasion: Suspicious. …

- Statistics and relevance rules:

Aspects	#Labeled	#Unlabeled	Relevance Rules
DCIS	23.8k	96.6k	DCIS, Ductal Carcinoma In-Situ
LCIS	10.7k		LCIS, Lobular Carcinoma In-Situ
IDC	22.9k		IDC, Invasive Ductal Carcinoma
ALH	9.2k		ALH, Atypical Lobular Hyperplasia

✦ 500 reports for testing

Review Dataset

- Domain transfer for sentiment analysis: positive or negative
- Common words (e.g. excellent) are directly transferrable, but domain-specific words are not

Source: Hotel (TripAdvisor)	→	Target: Restaurant (Yelp)
- This place was **excellent**! - In the second bedroom it literally **rained water from above** …		- **Excellent** food. - The fries were **undercooked** and **thrown haphazardly** into the sauce holder …

Review Dataset

- Domain transfer for sentiment analysis: positive or negative
- Common words (e.g. excellent) are directly transferrable, but domain-specific words are not

Source: Hotel (TripAdvisor)

- This place was **excellent**!
- In the second bedroom it literally **rained water from above** …

Target: Restaurant (Yelp)

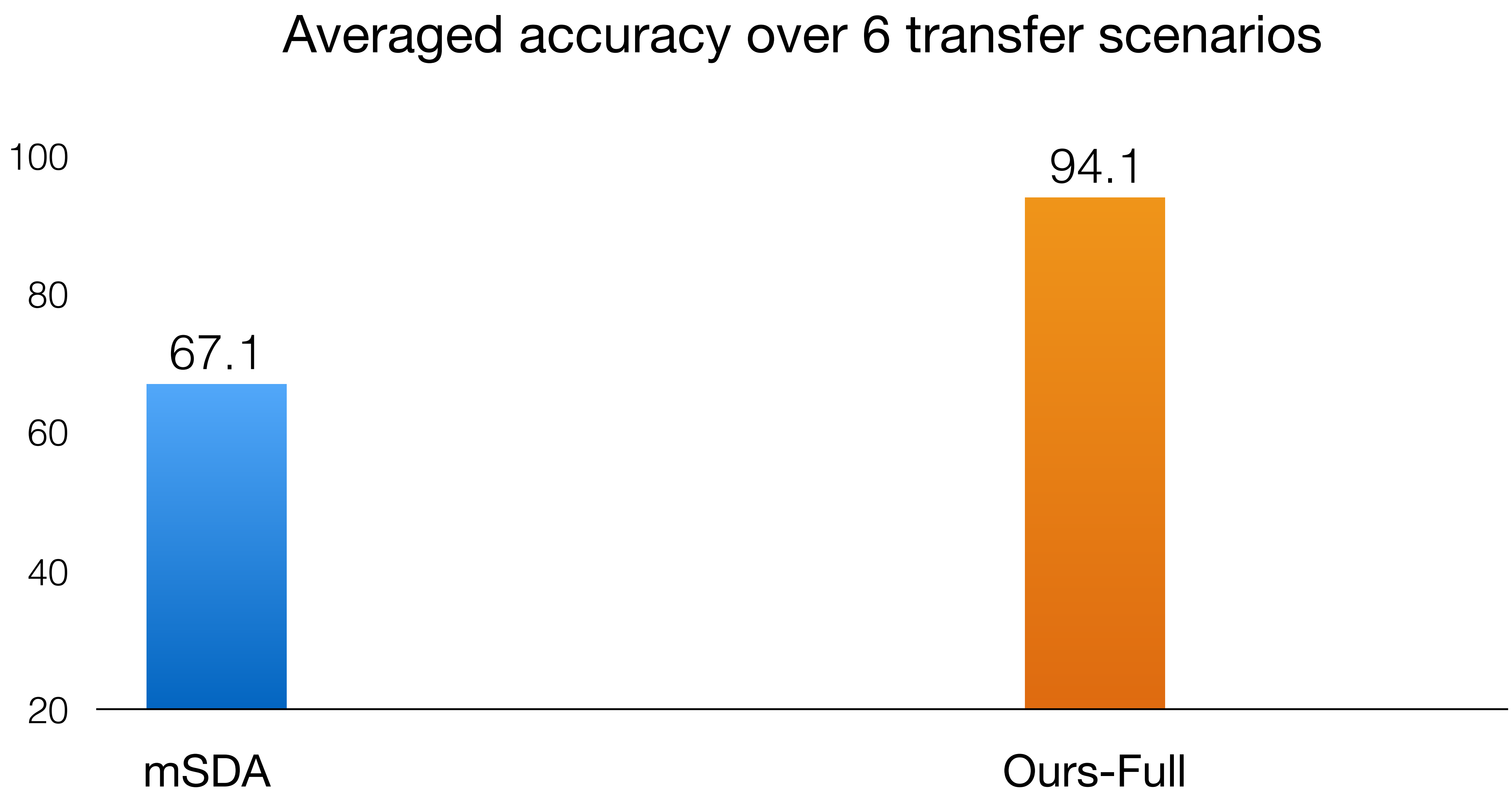
- **Excellent** food.
- The fries were **undercooked** and **thrown haphazardly** into the sauce holder …

- Statistics and relevance rules:

Domains	#Labeled	#Unlabeled	Relevance Rules
Hotel	100k	100k	Five aspects, 290 keywords (Wang et al., 2011)
Restaurant	-	200k	(only one *overall* aspect)

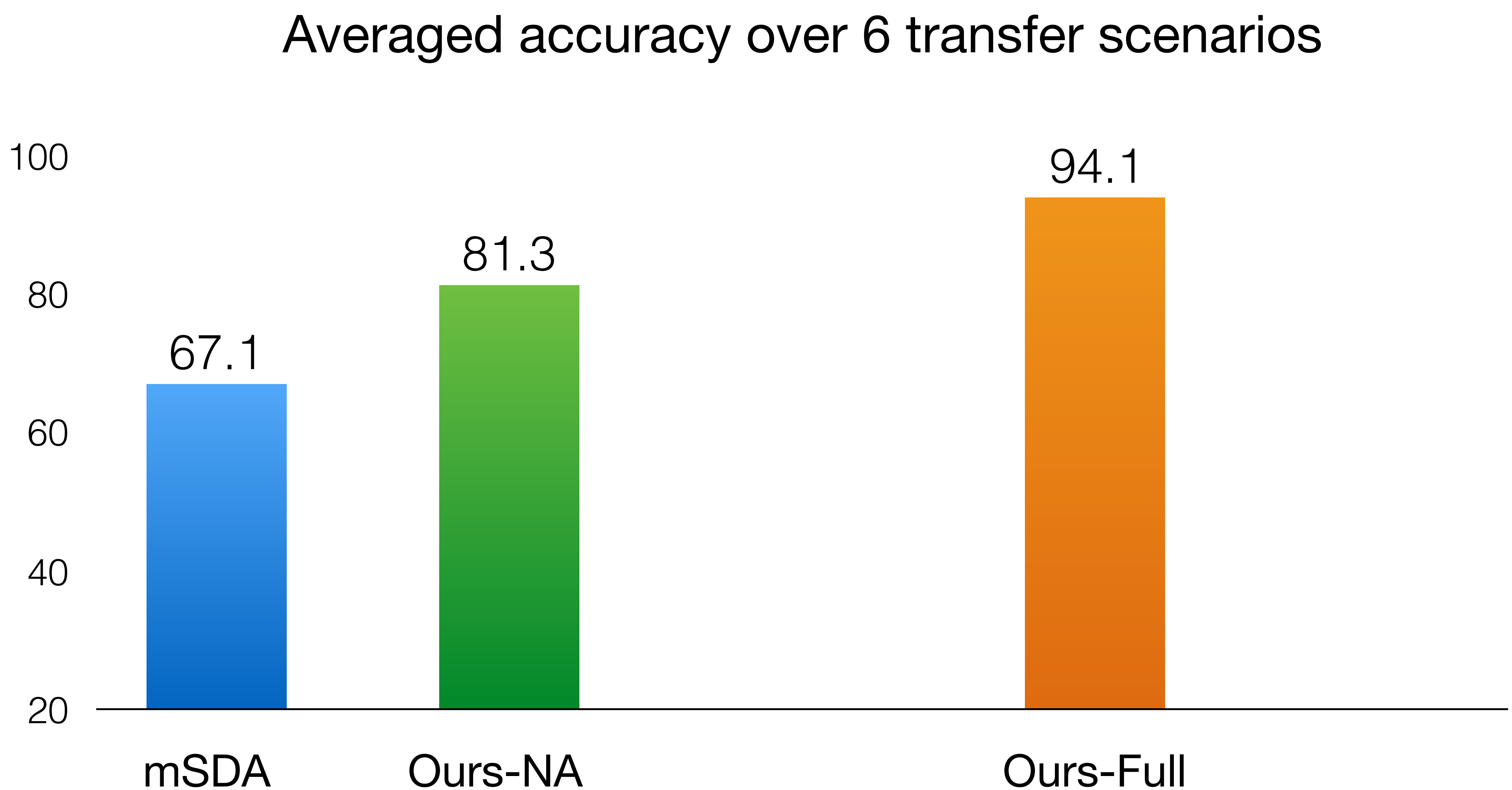
✦ 2k reviews for testing

Results on Pathology Dataset



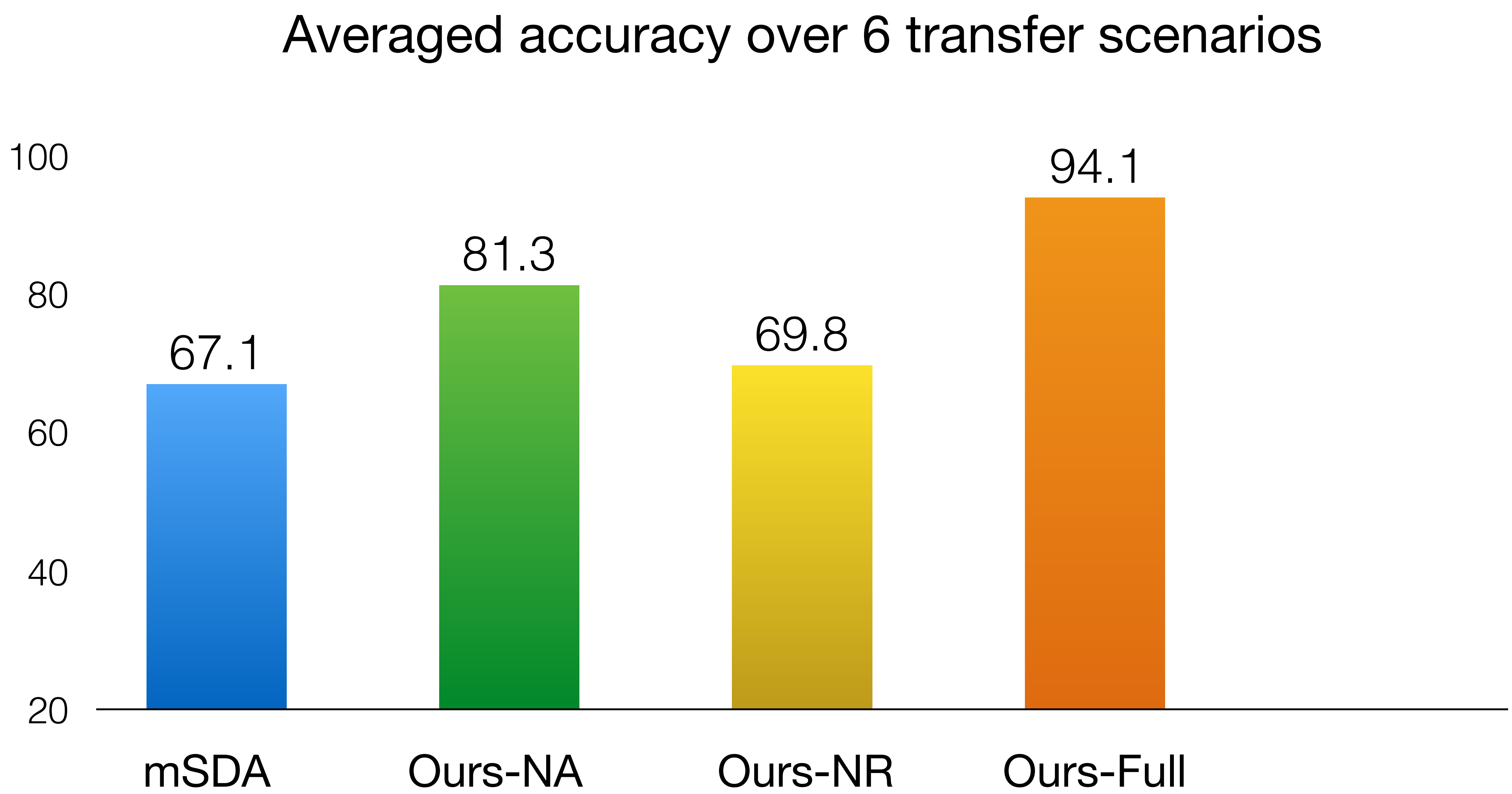
- mSDA: marginalized stacked denoising autoencoder (Chen et al., 2012)

Results on Pathology Dataset



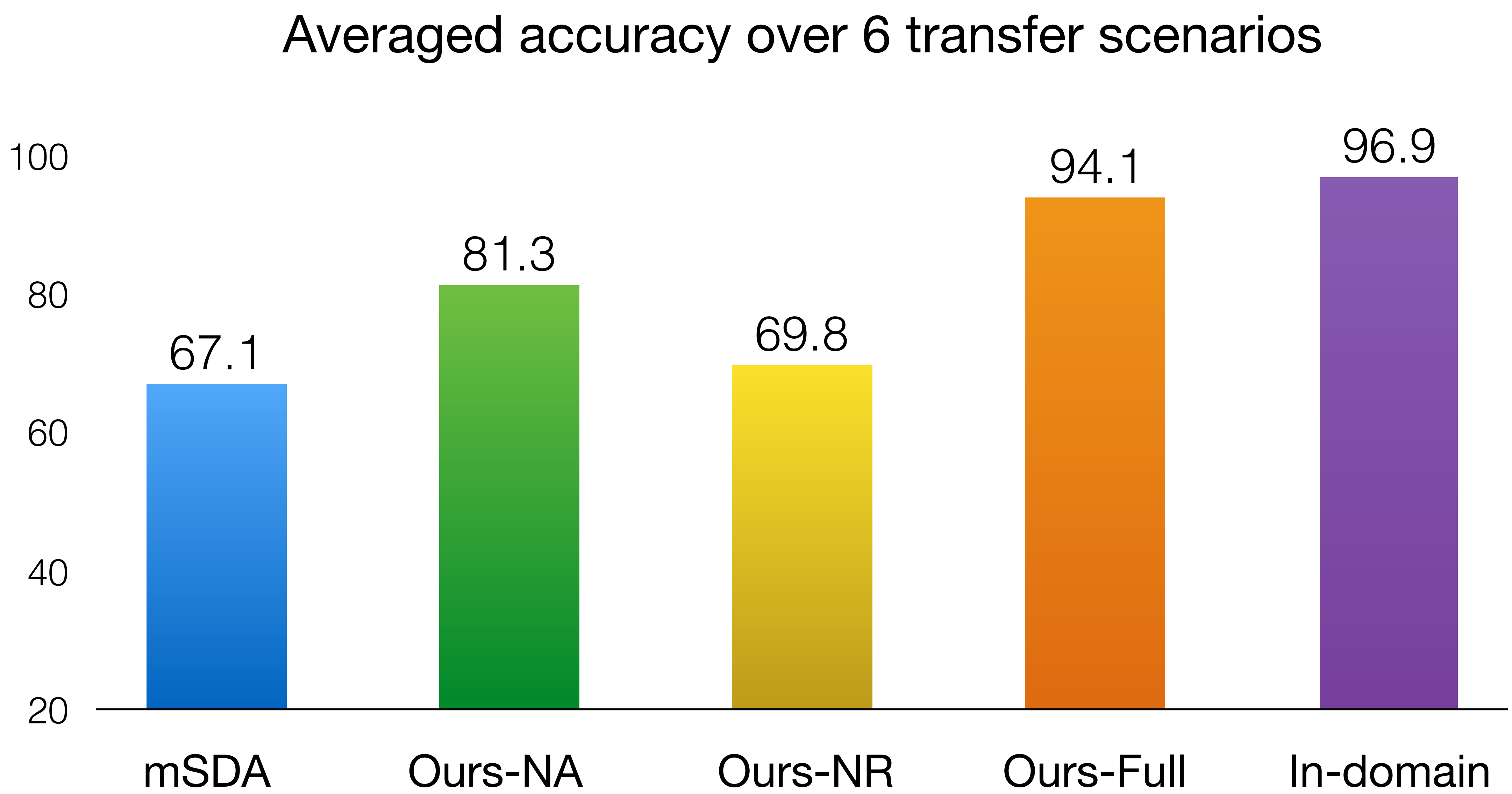
- Ours-NA: our model without adversarial training

Results on Pathology Dataset



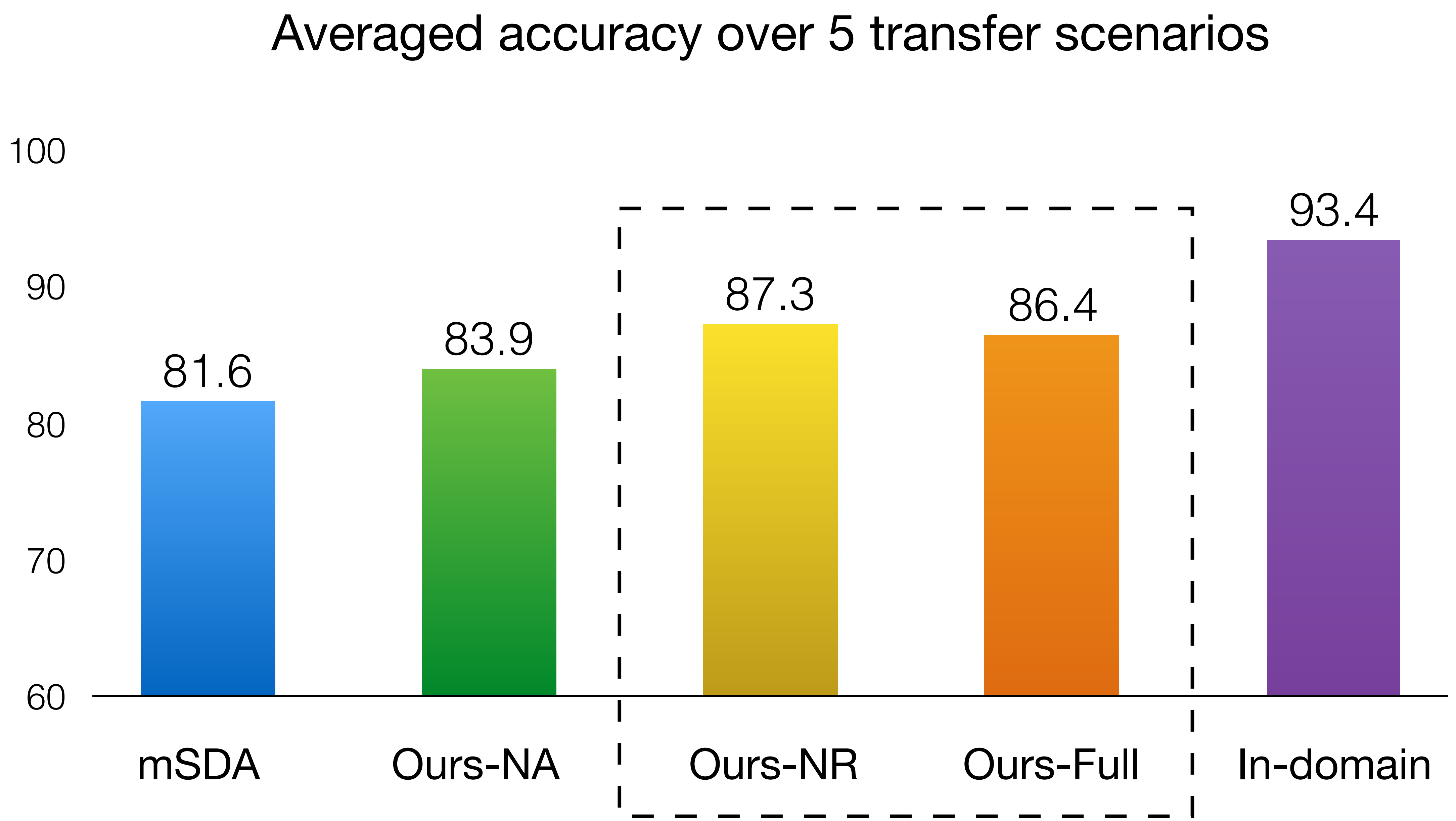
- Ours-NR: our model without aspect-relevance scoring

Results on Pathology Dataset



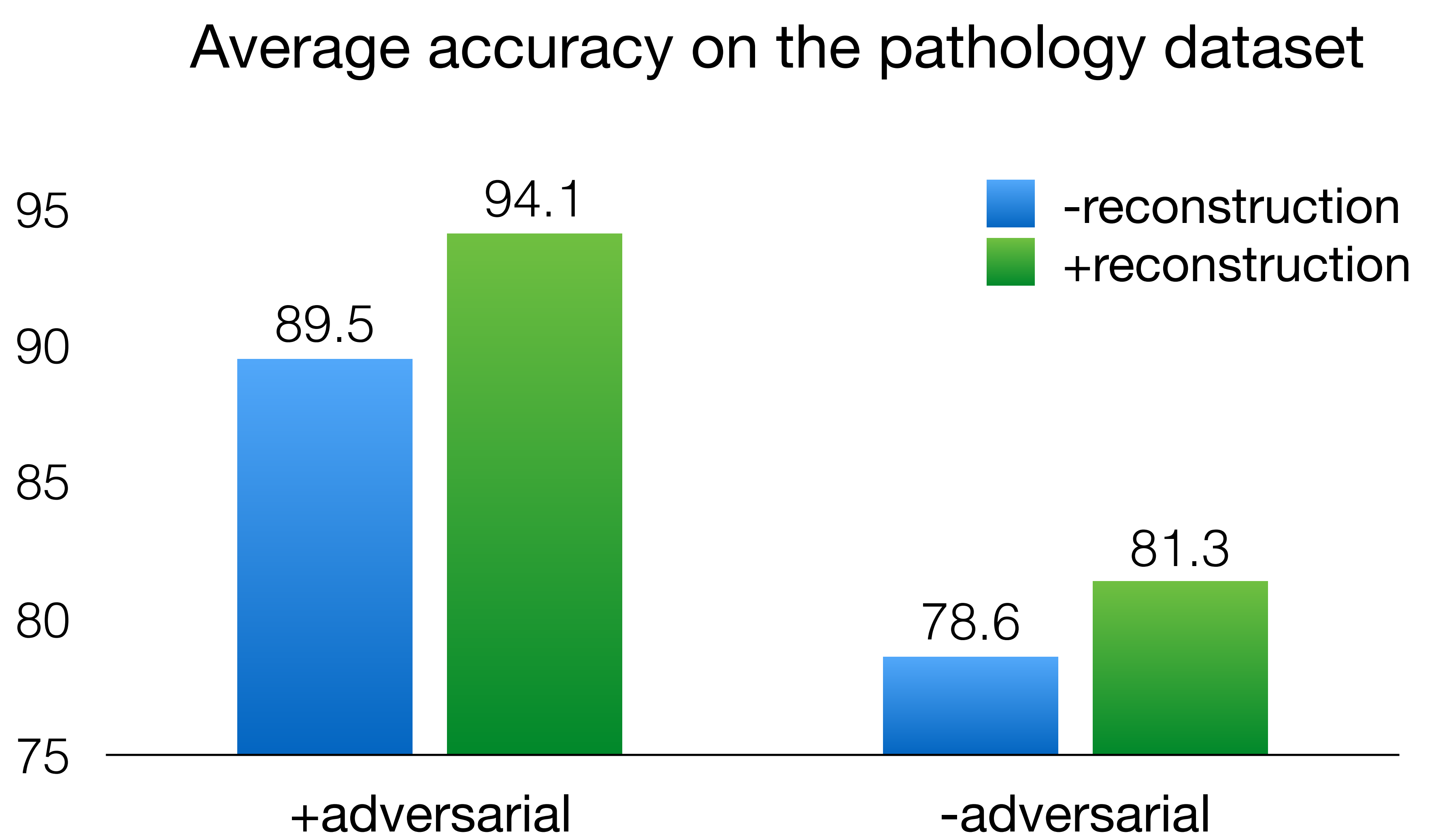
- In-domain: supervised training with in-domain annotations

Results on Review Dataset



- Ours-NR and Ours-Full are the two best performing systems
- Relevance scoring has little impact because aspects are highly correlated

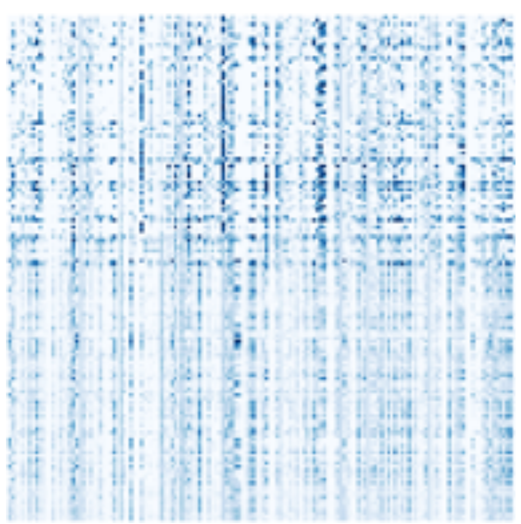
Impact of Reconstruction



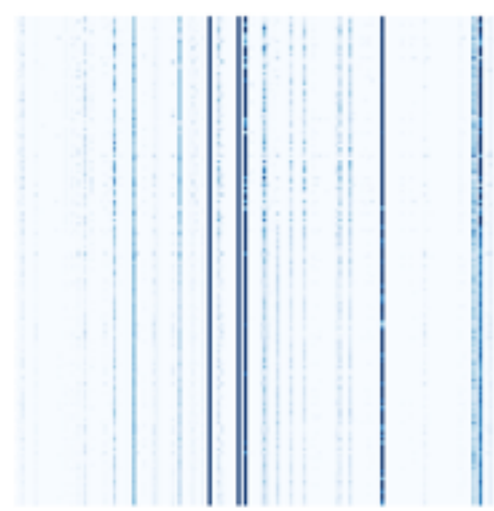
- The same observation on the review dataset

Reason behind Improvement

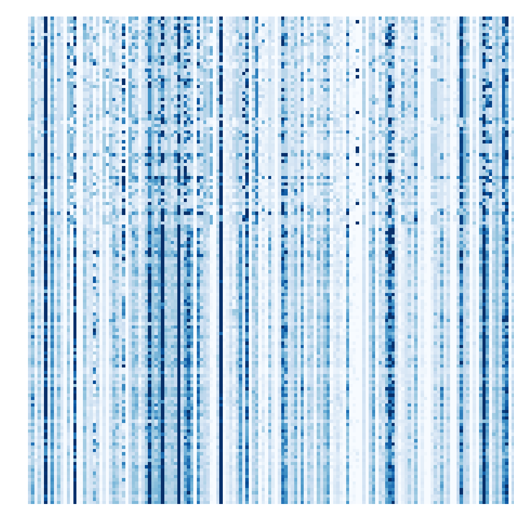
- Heat-map: each row corresponds to a document vector
  - Top: source domain; Bottom: target domain
- Adversarial training removes lots of information



−adversarial, −reconstruction



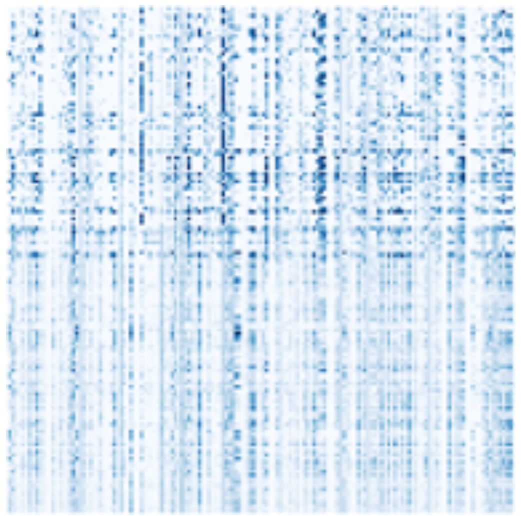
+adversarial, −reconstruction

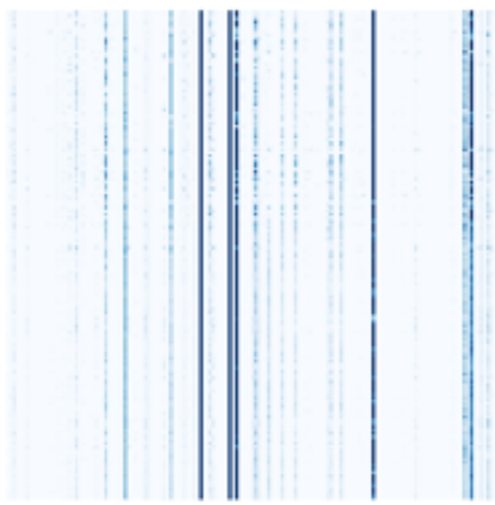


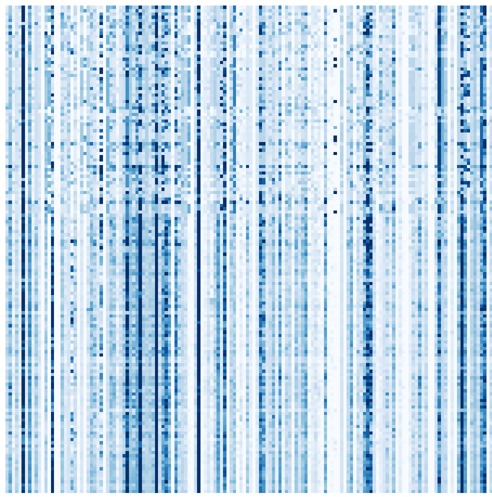
+adversarial, +reconstruction

Reason behind Improvement

- Heat-map: each row corresponds to a document vector
  - Top: source domain; Bottom: target domain
- Adversarial training removes lots of information
- The reconstruction loss improves both the richness and diversity of the learned representations


−adversarial, −reconstruction


+adversarial, −reconstruction


+adversarial, +reconstruction

Case Study of Learned Representations

Restaurant Reviews

- the fries were **undercooked** and **thrown haphazardly** into the sauce holder . the shrimp was **over cooked** and just **deep fried** . … even the water **tasted weird** .

Case Study of Learned Representations

Restaurant Reviews

- the fries were **undercooked** and **thrown haphazardly** into the sauce holder . the shrimp was **over cooked** and just **deep fried** . … even the water **tasted weird** .

Nearest Hotel Reviews by **Ours-Full: learns to map domain-specific words**

- the room was **old** . … we did n't like the night shows at all . …
- however , the decor **was just fair** . … in the second bedroom it literally **rained water from above** .

✦ distance measured by cosine similarity between representations

Case Study of Learned Representations

Restant Reviews

- the fries were **undercooked** and **thrown haphazardly** into the sauce holder . the shrimp was **over cooked** and just **deep fried** . … even the water **tasted weird** .

Nearest Hotel Reviews by **Ours-Full: learns to map domain-specific words**

- the room was **old** . … we did n't like the night shows at all . …
- however , the decor **was just fair** . … in the second bedroom it literally **rained water from above** .

Nearest Hotel Reviews by **Ours-NA: only captures common sentiment phrases**

- rest room in this restaurant is **very dirty** . …
- the only **problem** i had was that … i was very ill with what was suspected to be **food poison**

✦ distance measured by cosine similarity between representations

Summary

- *Modeling:* an aspect-augmented adversarial network for cross-aspect and cross-domain transfer tasks.

- *Performance:* our model significantly improves over the mSDA baseline and our model variants on a pathology and a review dataset

Contributions

Multilingual Transfer:

- Hierarchical tensors for dependency parsing
  - *Prior knowledge incorporation without feature engineering*

- Multilingual embeddings for POS tagging
  - *Effective multilingual transfer with ten translation pairs*

Monolingual Transfer:

- Adversarial networks for aspect transfer
  - *Joint aspect-driven encoding and domain adversarial training*

Thank you!

Contributions

Multilingual Transfer:

- Hierarchical tensors for dependency parsing
  - *Prior knowledge incorporation without feature engineering*

- Multilingual embeddings for POS tagging
  - *Effective multilingual transfer with ten translation pairs*

Monolingual Transfer:

- Adversarial networks for aspect transfer
  - *Joint aspect-driven encoding and domain adversarial training*

Backup Slides

Typological Features

Word ordering: five features, e.g.

Order of Subject and Verb (82A)

Order of Adjective and Noun (87A)

Typological feature templates: eight templates, e.g.

direction, 87A, head POS=NOUN, modifier POS=ADJ

direction, 82A, head POS=VERB, modifier POS=NOUN, label=SUBJ

Feature Weights of Multiway Model

Weights of valid features:

head POS=NOUN, mod POS=ADJ, 87A=ADJ-NOUN 2.24×10^{-3}

Weights of invalid features:

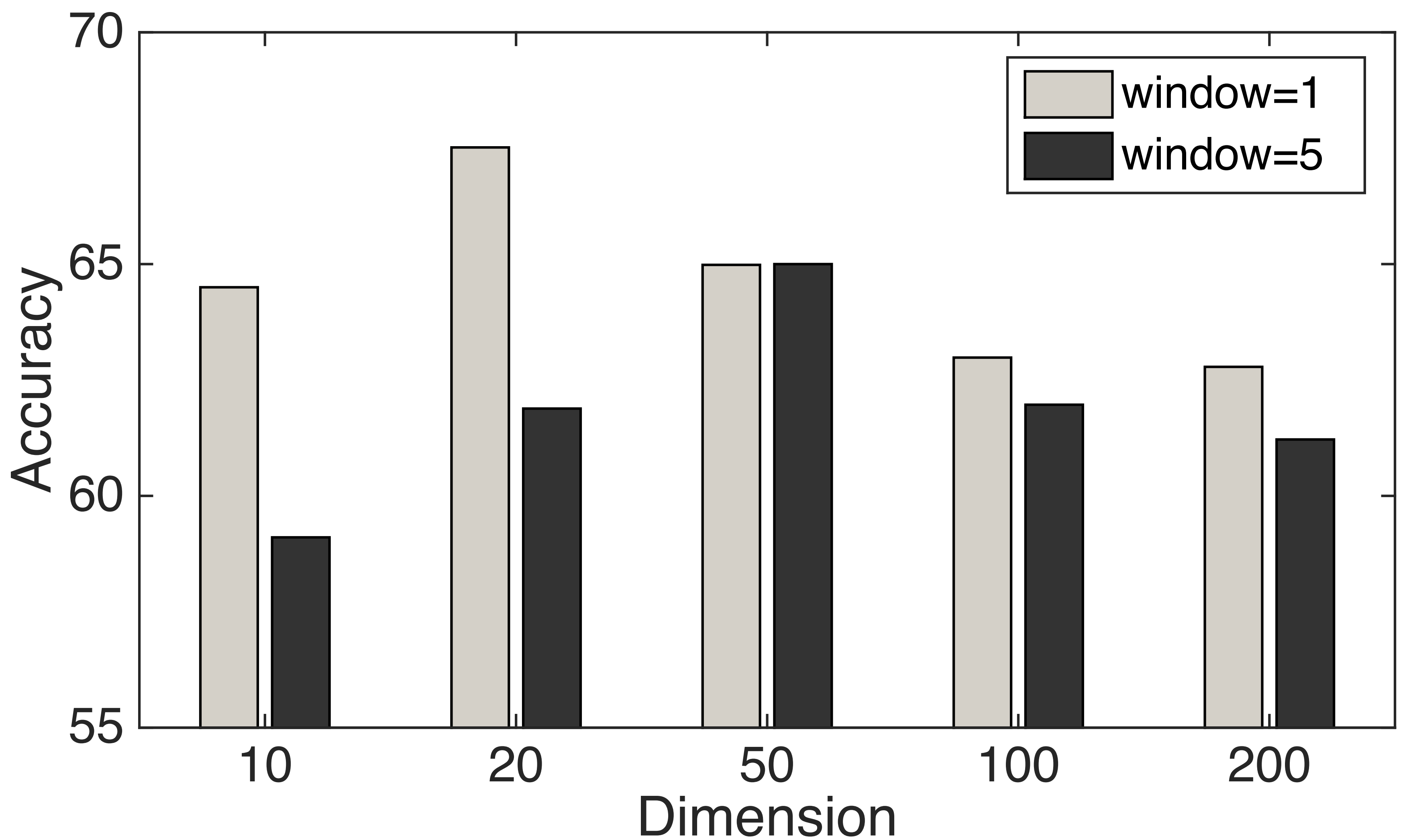
head POS=VERB, mod POS=NOUN, 87A=ADJ-NOUN 8.88×10^{-4}

head POS=NOUN, mod POS=NOUN, 87A=ADJ-NOUN 9.48×10^{-4}

Multiway model assigns non-zero weights to invalid features

Impact of Embedding Dimensions and Window Size

- Train embeddings with different dimensions and context window size
- Small window size favors POS tagging



Impact of Embedding Dimensions and Window Size

- Train embeddings with different dimensions and context window size
- Small window size favors POS tagging
- Performance drops with either smaller or larger dimensions