Reminders

HW10 ON NEURAL MACHINE TRANSLATION OR MILESTONE 2 IS DUE ON WEDNESDAY.

QUIZ ON CHAPTER 18 AND 20 (IE AND SRL) IS DUE TONIGHT AT MIDNIGHT.

Review: Machine Translation

Machine Translation

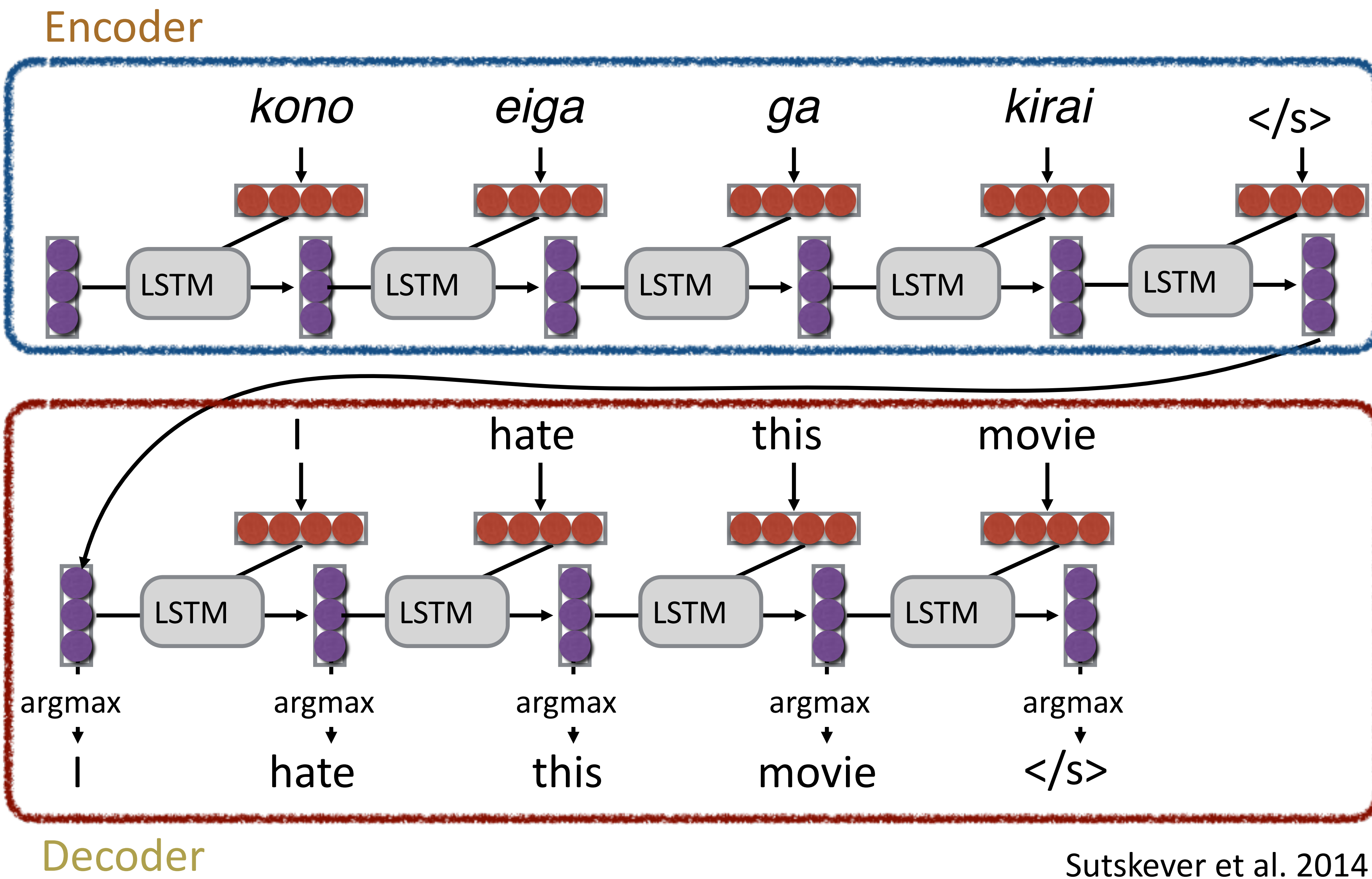
Translation from one language to another

I'm giving a talk at University of Pennsylvania

↓

ペンシルベニア大学で講演をしています。

Review: Encoder-Decoder MT



Sutskever et al. 2014

Review: Encoder-Decoder MT

MT is the task of automatically translating sentences from one language into another.

We use bilingual parallel texts to train MT systems – pairs of **source-target** sentences that are translations of each other.

To extend LMs and autoregressive generation to MT, we will:

1. Add an end-of-sentence marker to each source sentence. Concatenate the target sentence to it.
2. Train an RNN LM based on this combined data.
3. To translate, simply treat the input sentence as a prefix, create a hidden state representation for it **(encoding step)**.
4. Use the hidden state produced by the encoder to then start generating (**decoding step)**

Evaluating MT Quality

Evaluating MT Quality

Why do we want to do it?

- Want to rank systems
- Want to evaluate incremental changes
- What to make scientific claims

How not to do it

- “Back translation”
- The vodka is not good

Human Evaluation of MT v. Automatic Evaluation

Human evaluation is

- Ultimately what we're interested in, but
- Very time consuming
- Not re-usable

Automatic evaluation is

- Cheap and reusable, but
- Not necessarily reliable

Manual Evaluation

Source: Estos tejidos están analizados, transformados y congelados antes de ser almacenados en Hema-Québec, que gestiona también el único banco público de sangre del cordón umbilical en Quebec.

Reference: These tissues are analyzed, processed and frozen before being stored at Héma-Québec, which manages also the only bank of placental blood in Quebec.

Translation	Rank
These weavings are analyzed, transformed and frozen before being stored in Hema-Quebec, that negotiates also the public only bank of blood of the umbilical cord in Quebec.	○ 1 Best ○ 2 ○ 3 ○ 4 ● 5 Worst
These tissues analysed, processed and before frozen of stored in Hema-Québec, which also operates the only public bank umbilical cord blood in Quebec.	○ 1 Best ○ 2 ● 3 ○ 4 ○ 5 Worst
These tissues are analyzed, processed and frozen before being stored in Hema-Québec, which also manages the only public bank umbilical cord blood in Quebec.	○ 1 Best ● 2 ○ 3 ○ 4 ○ 5 Worst
These tissues are analyzed, processed and frozen before being stored in Hema-Quebec, which also operates the only public bank of umbilical cord blood in Quebec.	● 1 Best ○ 2 ○ 3 ○ 4 ○ 5 Worst
These fabrics are analyzed, are transformed and are frozen before being stored in Hema-Québec, who manages also the only public bank of blood of the umbilical cord in Quebec.	○ 1 Best ○ 2 ○ 3 ● 4 ○ 5 Worst

Goals for Automatic Evaluation

No cost evaluation for incremental changes

Ability to rank systems

Ability to identify which sentences we're doing poorly on, and categorize errors

Correlation with human judgments

Interpretability of the score

Methodology

- Comparison against reference translations
- Intuition: closer we get to human translations, the better we're doing
- Could use WER like in speech recognition?

Word Error Rate

Levenshtein Distance (also known as "edit distance")

Minimum number of insertions, substitutions, and deletions needed to transform one string into another

Useful measure in speech recognition

- This shows how easy it is to recognize speech
- This shows how easy it is to wreck a nice beach

Problems with WER

Unlike speech recognition we don't have the assumption of **exact match against the reference or linearity**

In MT there can be many possible (and equally valid) ways of translating a sentence, and phrases can be rearranged.

1

Compare against lots of test sentences

2

Use multiple reference translations for each test sentence

3

Look for phrase / n-gram matches, allow movement

Solutions

BLEU

Bi**L**ingual **E**valuation **U**nderstudy

Uses multiple reference translations

Look for n-grams that occur anywhere in the sentence

Ref 1	Orejuela appeared calm as he was led to the American plane which will take him to Miami, Florida.
Ref 2	Orejuela appeared calm while being escorted to the plane that would take him to Miami, Florida.
Ref 3	Orejuela appeared calm as he was being led to the American plane that was to carry him to Miami in Florida.
Ref 4	Orejuela seemed quite calm as he was being led to the American plane that would take him to Miami in Florida.

Multiple references

$$p_n = \frac{\sum_{S \in C} \sum_{ngram \in S} Count_{matched}(ngram)}{\sum_{S \in C} \sum_{ngram \in S} Count(ngram)}$$

n-gram precision

BLEU MODIFIES THIS PRECISION TO ELIMINATE REPETITIONS THAT OCCUR ACROSS SENTENCES.

Ref 1	Orejuela appeared calm as he was led to the American plane which will take him **to Miami**, Florida.
Ref 2	Orejuela appeared calm while being escorted to the plane that would take him **to Miami**, Florida.
Ref 3	Orejuela appeared calm as he was being led to the American plane that was to carry him **to Miami** in Florida.
Ref 4	Orejuela seemed quite calm as he was being led to the American plane that would take him in Florida. **to Miami**

Multiple references

"to Miami" can only be counted as correct once

Ref 1	Orejuela appeared calm as he was led to the American plane which will take him to Miami, Florida.
Ref 2	Orejuela appeared calm while being escorted to the plane that would take him to Miami, Florida.
Ref 3	Orejuela appeared calm as he was being led to the American plane that was to carry him to Miami in Florida.
Ref 4	Orejuela seemed quite calm as he was being led to the American plane that would take him to Miami in Florida.

Hyp	appeared calm when he was taken to the American plane, which will to Miami, Florida.

American, Florida, Miami, Orejuela, **appeared**, as, being, calm, carry, escorted, he, him, in, led, **plane**, quite, seemed, take, that, the, **to**, **to**, to, **was** , was, **which**, while, **will**, would, **,, .**

1-gram precision = 15/18

Hyp	**appeared calm** when **he was** taken **to the American plane** , **which will to Miami , Florida .**

American plane, **Florida .**, **Miami ,**, Miami in, Orejuela appeared, Orejuela seemed, **appeared calm**, as he, being escorted, being led, calm as, calm while, carry him, escorted to, **he was**, him to, in Florida, led to, plane that, plane which, quite calm, seemed quite, take him, that was, that would, **the American**, the plane, **to Miami**, to carry, **to the**, was being, was led, was to, **which will**, while being, will take, would take, , Florida

2-gram precision = 10/17

Hyp	**appeared calm** when **he was** taken **to the American plane** , **which will to Miami , Florida .**

n-gram precision

Hyp	appeared calm when he was taken to the American plane, which will to Miami, Florida.

1-gram precision = 15/18 = .83

2-gram precision = 10/17 = .59

3-gram precision = 5/16 = .31

4-gram precision = 3/15 = .20

- Geometric average

$$(0.83 * 0.59 * 0.31 * 0.2)^\wedge(1/4) = 0.417$$

or equivalently

$$\exp(\ln .83 + \ln .59 + \ln .31 + \ln .2/4) = 0.417$$

Ref 1	Orejuela appeared calm as he was led to the American plane which will take him to Miami, Florida.
Ref 2	Orejuela appeared calm while being escorted to the plane that would take him to Miami, Florida.
Ref 3	Orejuela appeared calm as he was being led to the American plane that was to carry him to Miami in Florida.
Ref 4	Orejuela seemed quite calm as he was being led to the American plane that would take him to Miami in Florida.

Hyp	to the American plane

Is this better?

Hyp	to the American plane

1-gram precision = 4/4 = 1.0

2-gram precision = 3/3 = 1.0

3-gram precision = 2/2 = 1.0

4-gram precision = 1/1 = 1.0

exp(ln 1 + ln 1 + ln 1 + ln 1) = 1

Brevity Penalty

c is the length of the corpus of hypothesis translations

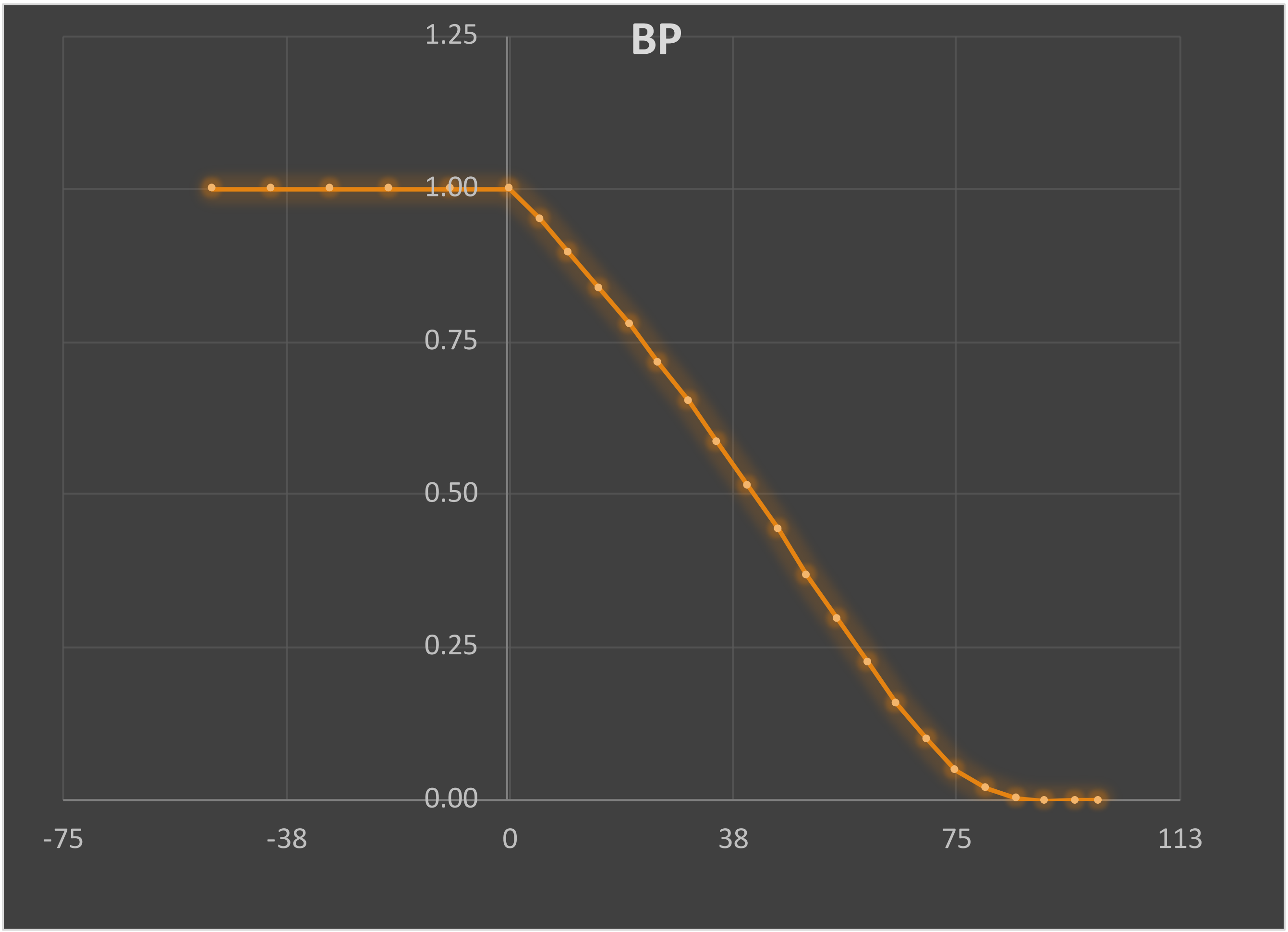
r is the effective reference corpus length

The effective reference corpus length is the sum of the single reference translation from each set that is closest to the hypothesis translation.

$$\mathrm{BP} = \begin{cases} 1 & \text{if} \quad c > r \\ e^{1-r/c} & \text{if} \quad c \leq r \end{cases}$$

Brevity Penalty

MT is Longer MT is Shorter



Difference with effective reference length (%)

Ref 1	Orejuela appeared calm as he was led to the American plane which will take him to Miami, Florida. $r = 20$
Hyp	appeared calm when he was taken to the American plane, which will to Miami, Florida. $c = 18$

$$BP = \exp(1-(20/18)) = 0.89$$

Ref 1	Orejuela appeared calm as he was led to the American plane which will take him to Miami, Florida. $r = 20$
Hyp	to the American plane $c = 4$

$$BP = \exp(1-(20/4)) = 0.02$$

BLEU

Geometric average of the n-gram precisions

Optionally weight them with w

Multiplied by the brevity penalty

$$\text{Bleu} = \text{BP} * \exp\left(\sum_{n=1}^{N} w_n \log p_n\right)$$

BLEU

Hyp	appeared calm when he was taken to the American plane, which will to Miami, Florida.

$$\exp(1-(20/18)) * \exp((\ln .83 + \ln .59 + \ln .31 + \ln .2)/4) = \mathbf{0.374}$$

Hyp	to the American plane

$$\exp(1-(20/4)) * \exp((\ln 1 + \ln 1 + \ln 1 + \ln 1)/4) = \mathbf{0.018}$$

Problems with BLEU

Synonyms and paraphrases are only handled if they are in the set of multiple reference translations

The scores for **words are equally weighted** so missing out on content-bearing material brings no additional penalty.

The brevity penalty is a stop-gap measure to compensate for the fairly serious problem of not being able to calculate **recall**.

WER - word error rate

PI-WER - position independent WER

METEOR - **M**etric for **E**valuation of **T**ranslation with **E**xplicit **OR**dering

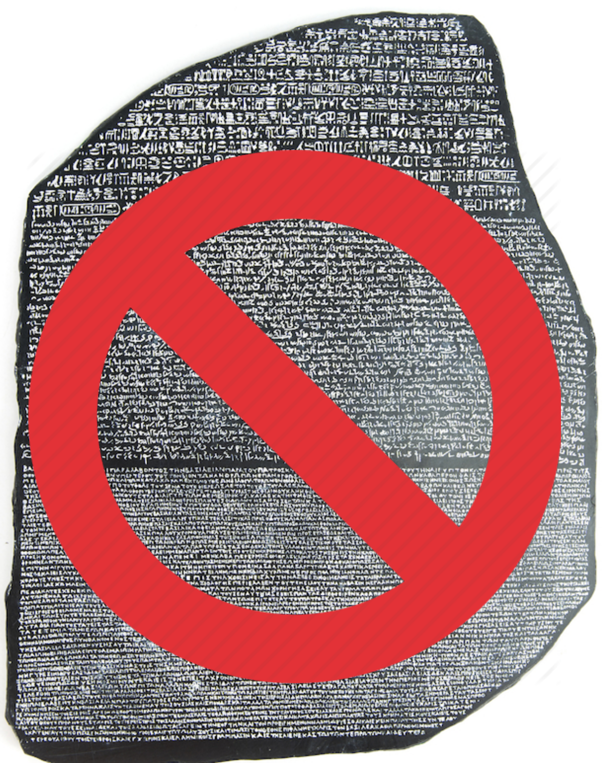
TERp - **T**ranslation **E**dit **R**ate **p**lus

More Metrics

Cross-lingual Word Representations

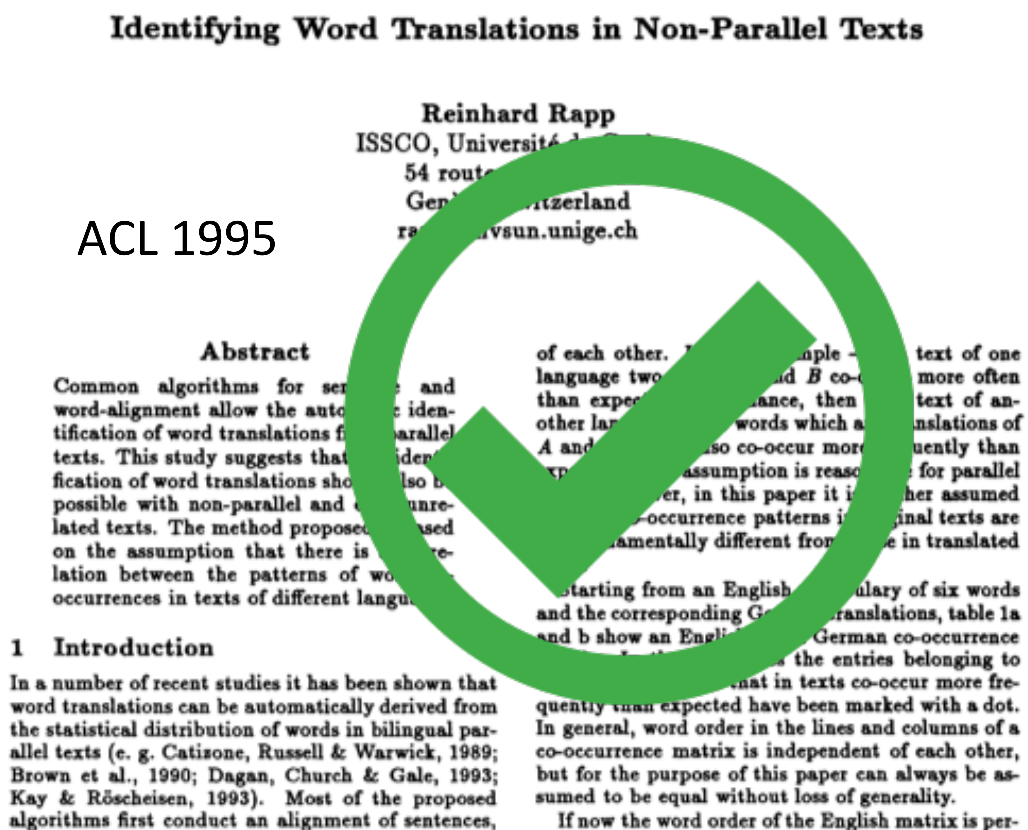
Goal

Learn the translations of individual words without large bilingual parallel corpora



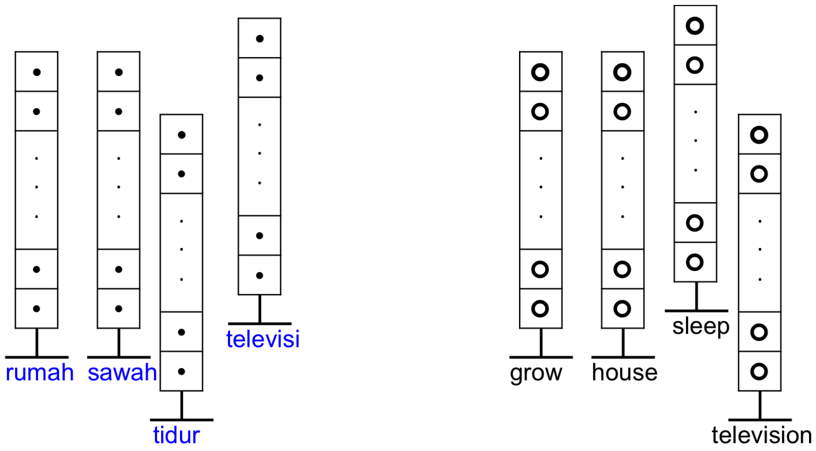
Egypt 196BC

ACL 1995

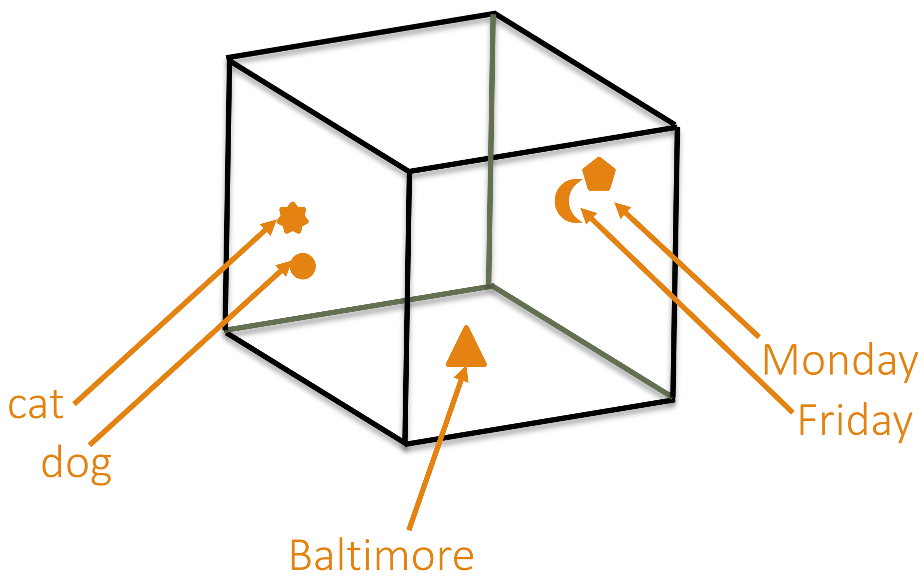


Translations from monolingual texts

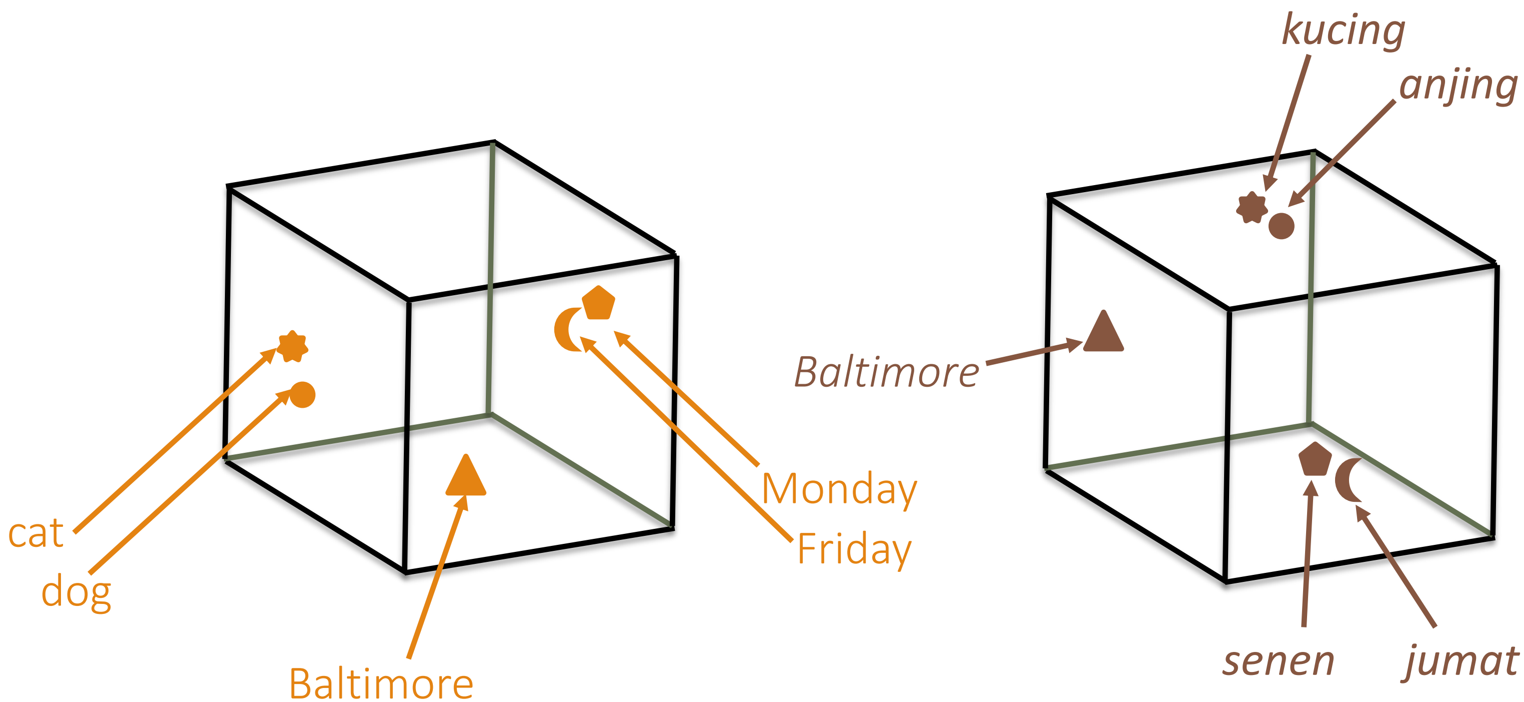
Word embeddings have been shown to be useful for many natural language processing tasks. Can we use these vector space models to learn translations for rare words?



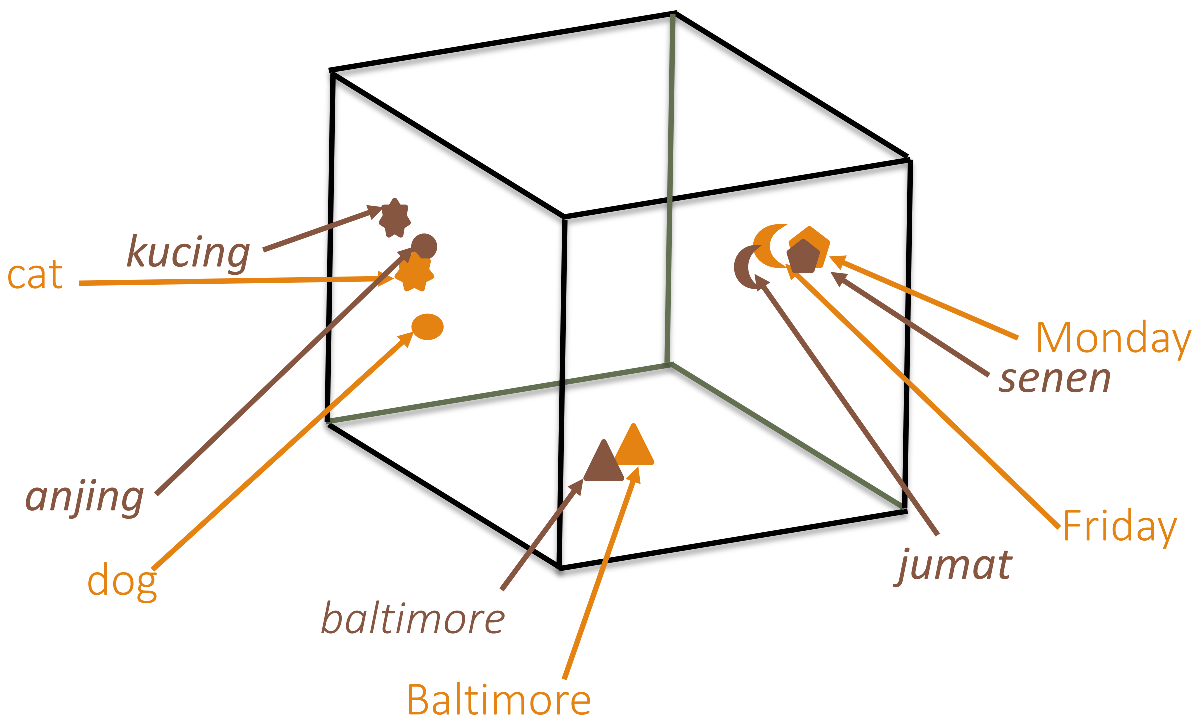
Monolingual Word Embeddings



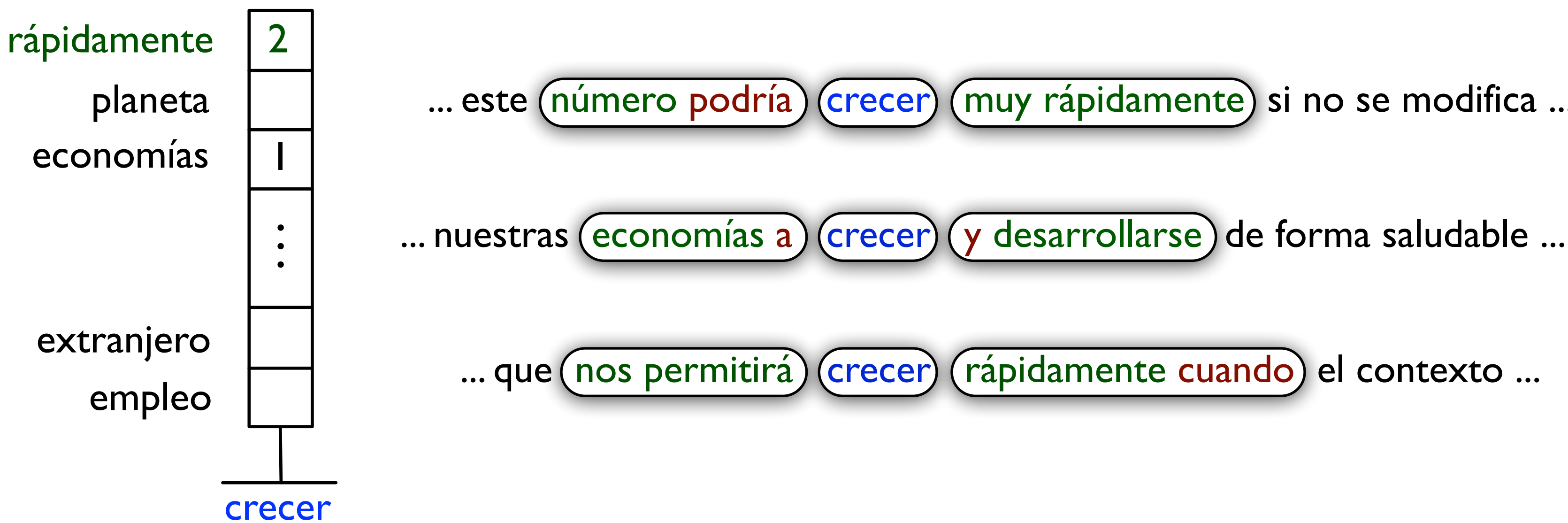
Monolingual Word Embeddings



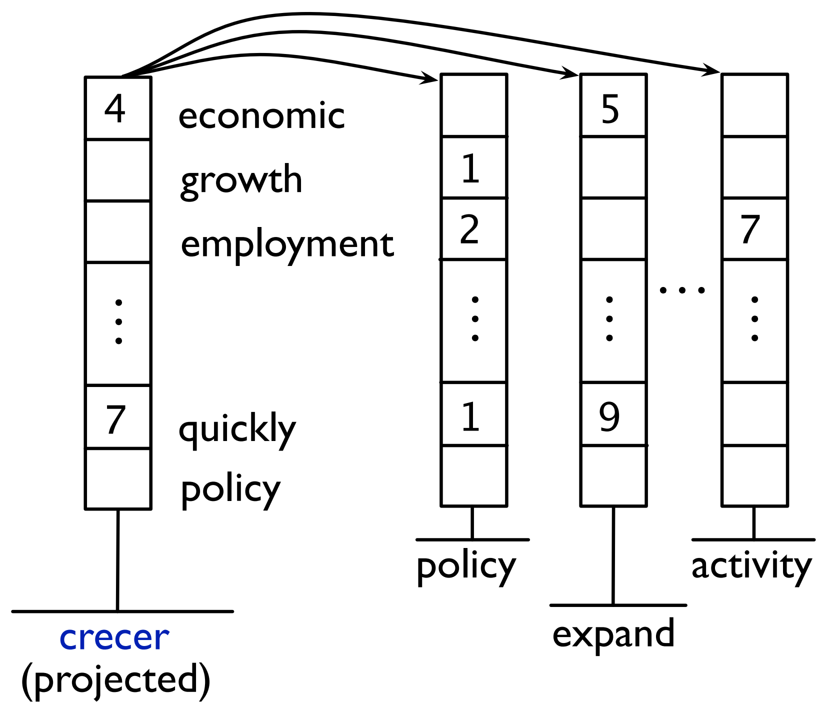
Bilingual Word Embeddings



Projecting Vector Space Models

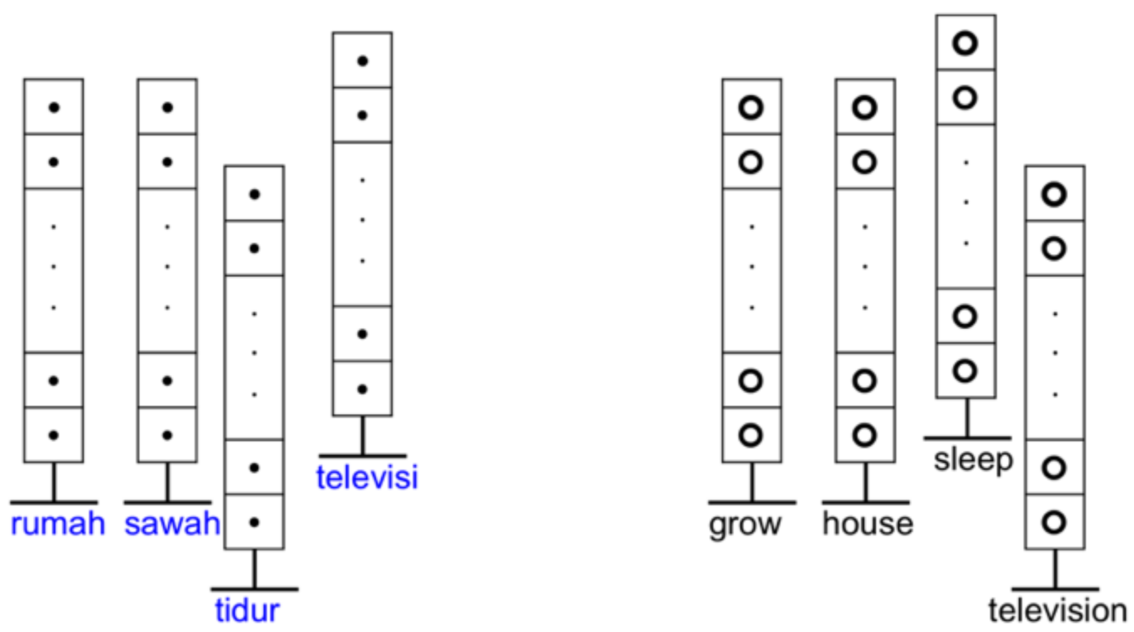


Projecting Vector Space Models



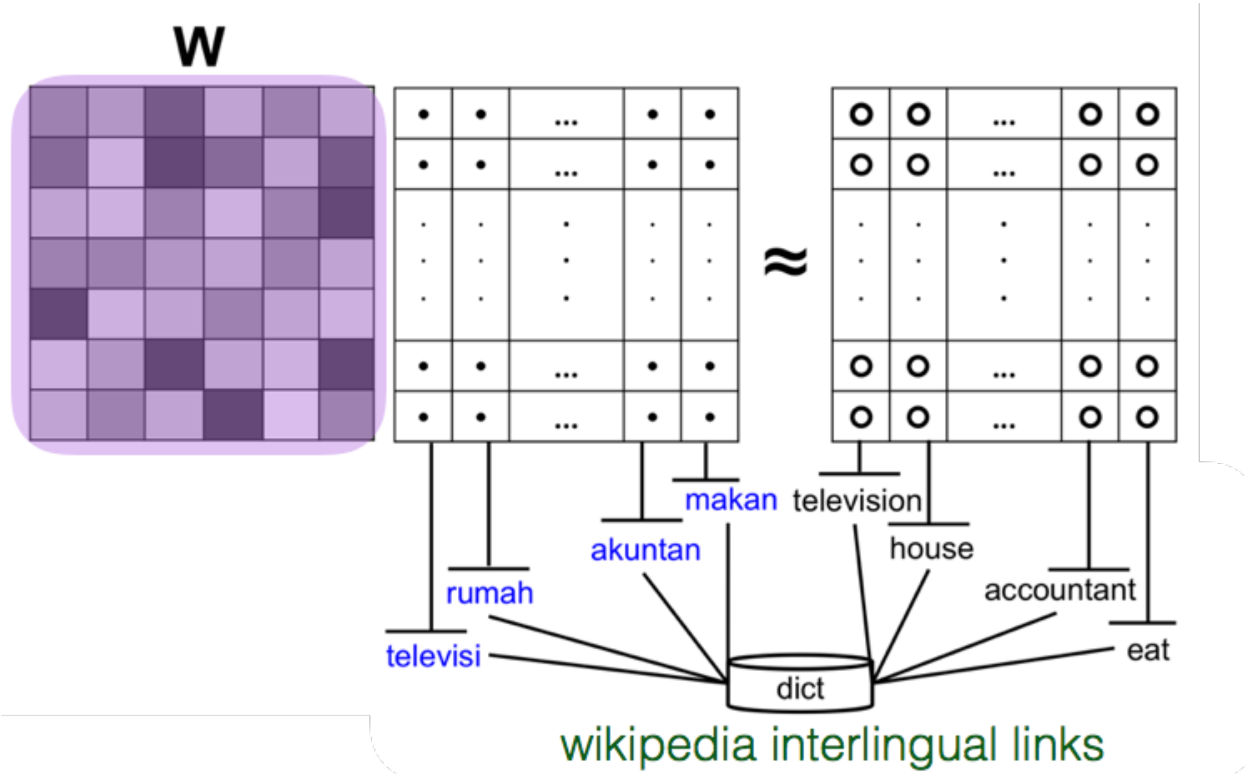
Word Embeddings

Instead of high dimensional vector space models used by Rapp and others in the past, we use low-dimensional word embeddings.



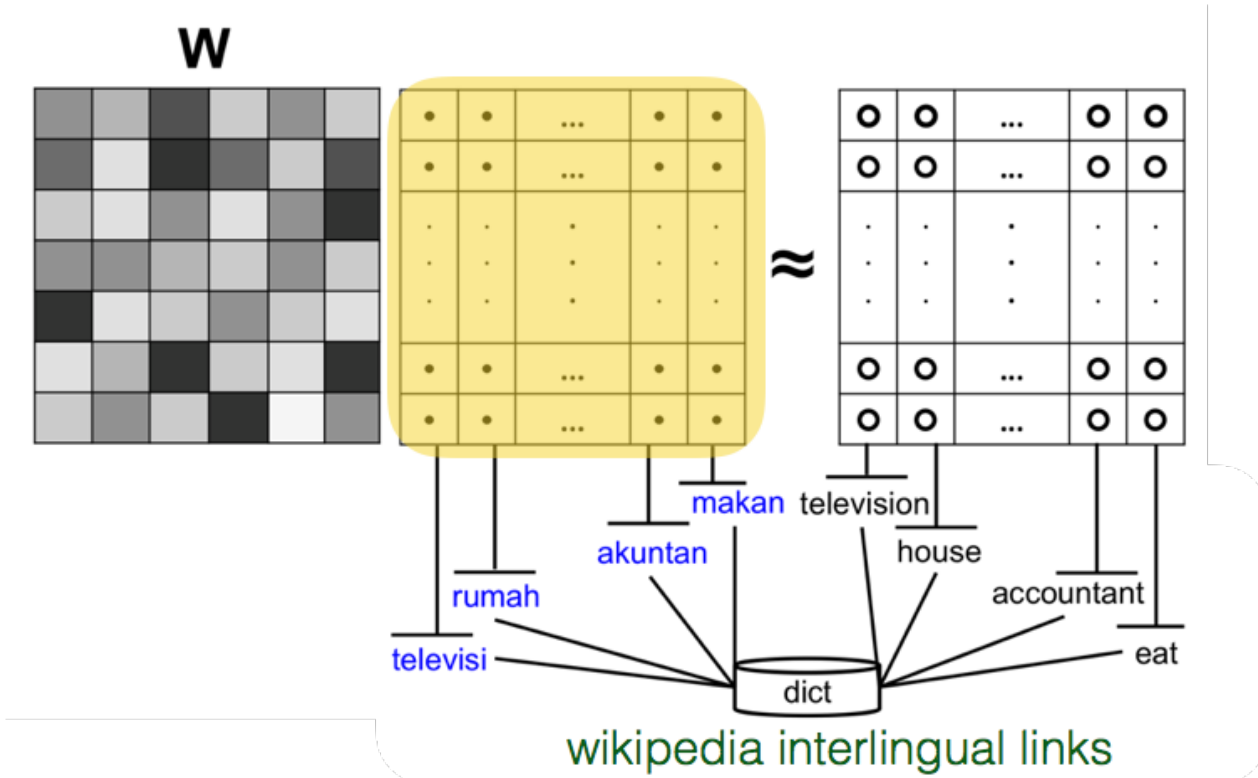
Learning Bilingual Embeddings

mapping function W



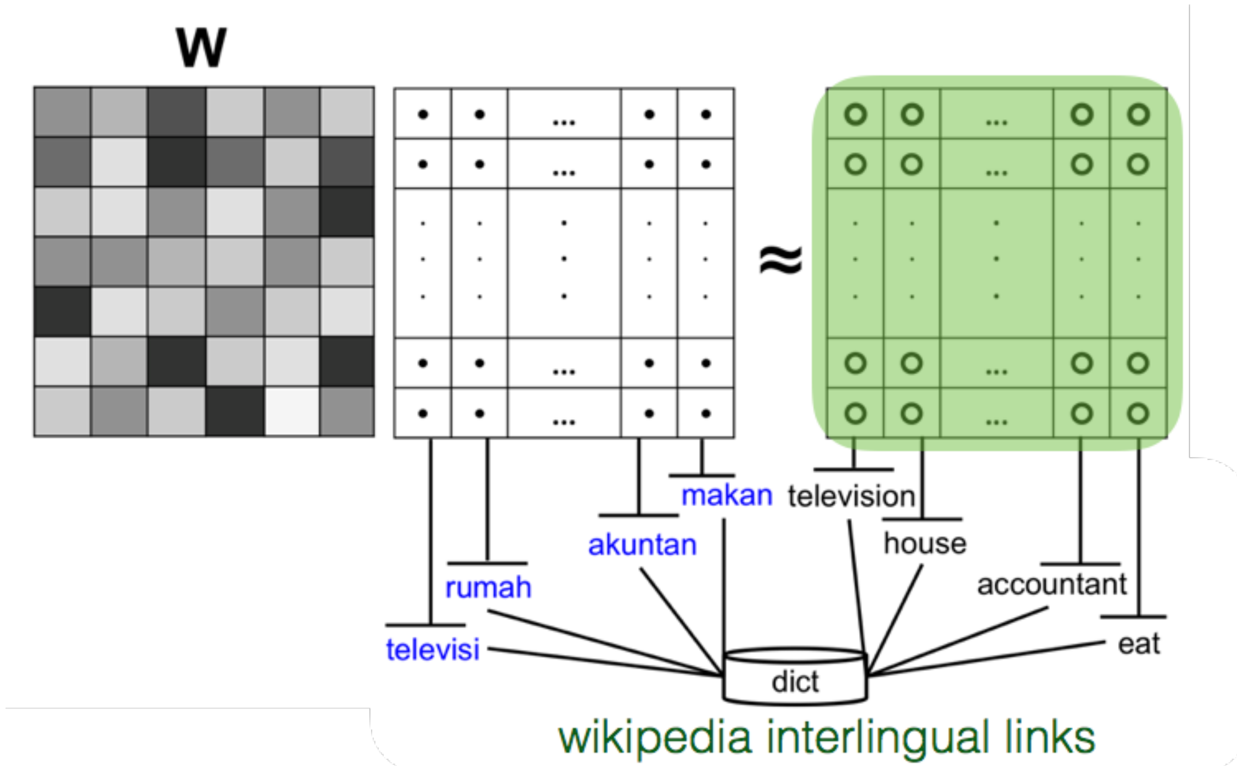
Learning Bilingual Embeddings

matrix of source language embeddings



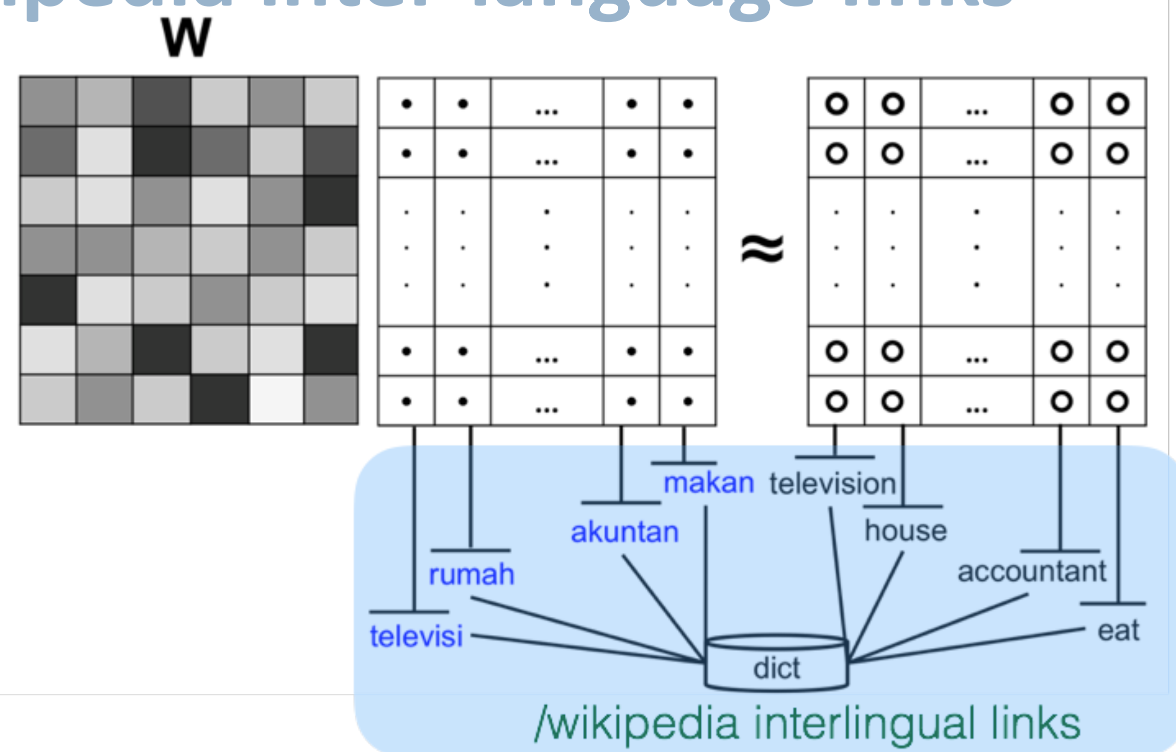
Learning Bilingual Embeddings

matrix of target language embeddings



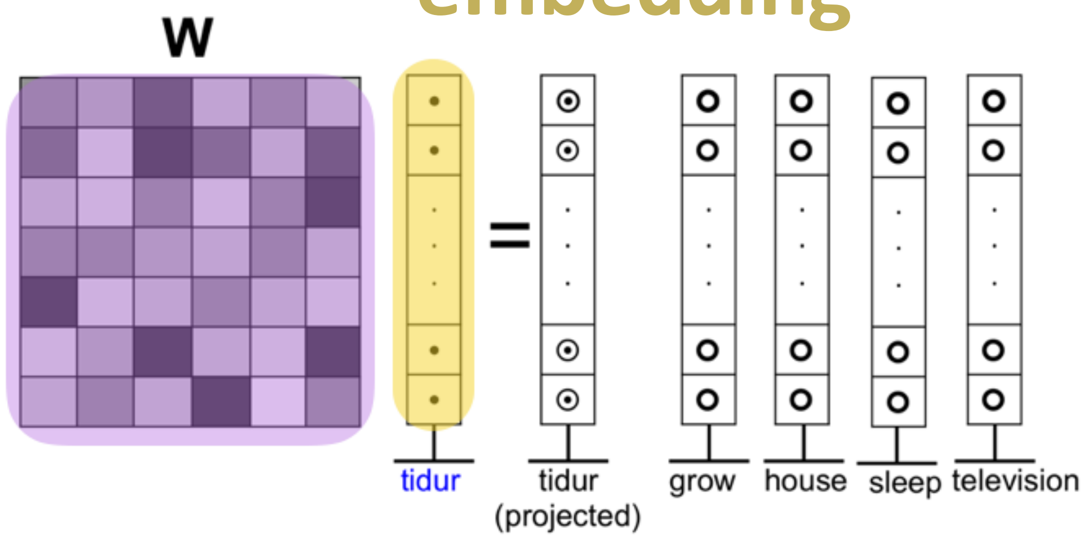
Learning Bilingual Embeddings

bilingual dictionaries or
Wikipedia inter-language links



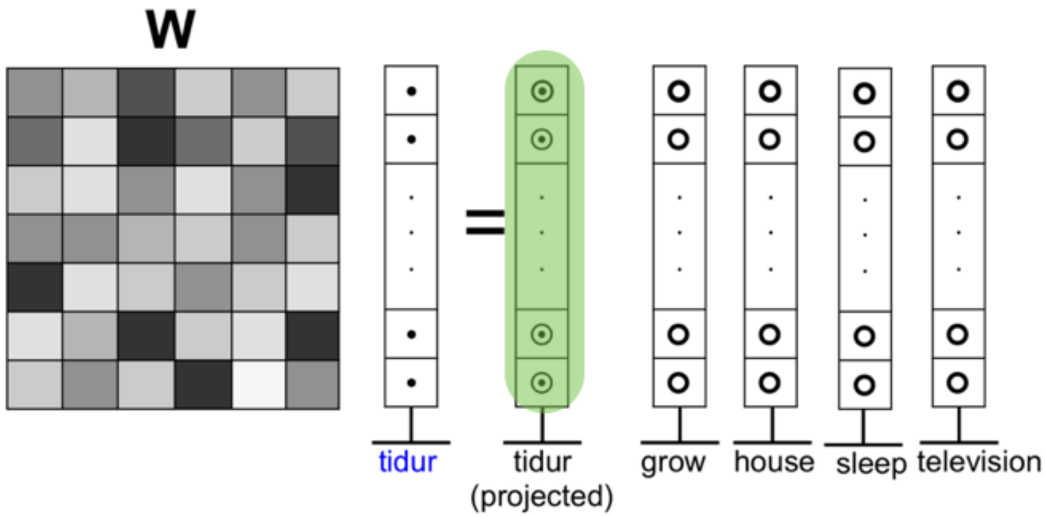
Learning Bilingual Embeddings

Apply **W** to a **source language embedding**



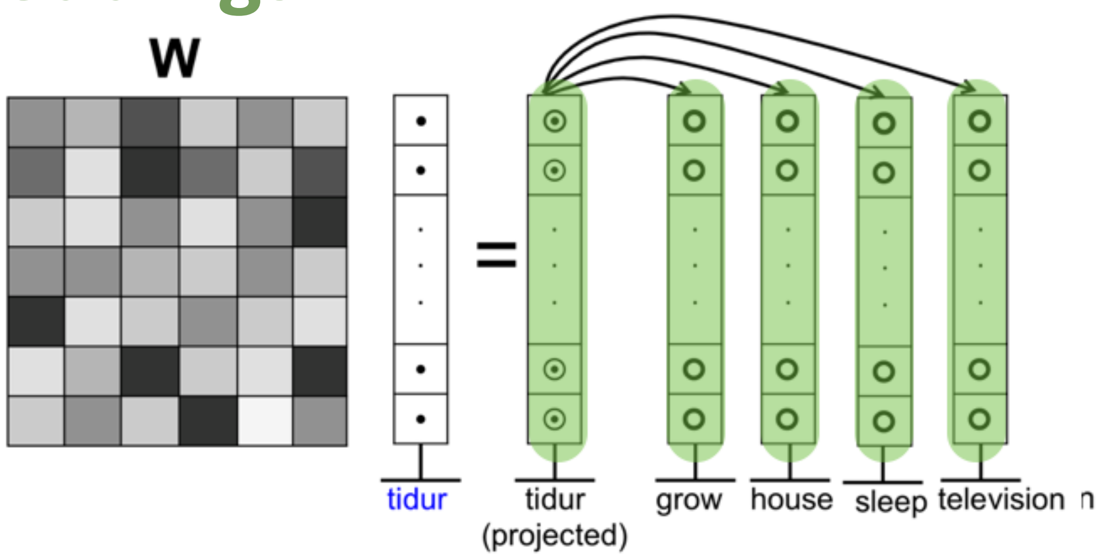
Learning Bilingual Embeddings

Project it into **the target language space**



Learning Bilingual Embeddings

Compare against **all target language embeddings**



use nearest neighbors as **translations**

Example Translations for Indonesian

mediterania	pertumbuhannya	solusi	pagar	armada
mediterranean	**growth**	**solutions**	**fence**	armada
aegean	exponentially	solver	tail	**fleet**
atlantic	germination	alternatives	**fences**	ships
baltic	rapidly	solving	info	warships
levantine	regrowth	bootstrapping	perimeter	freighter
europe	thrive	**solution**	biography	tanker
adriatic	/year	objective	around	oiler
pacific	growing	problem	moat	lst
marmara	steadily	enabler	embankment	frigate
caribbean	stunted	solvers	clothing	squadron

Ways to learn W

Linear Mapping $||\mathbf{W}\mathbf{X_E} - \mathbf{X_F}||_F^2$

Neural Net $\sum_{\mathbf{x}_f \in \mathbf{X_F}} \sum_{\mathbf{x}_e \in \mathbf{X_E}} ||\mathbf{x}_f - \phi^{(4)} s(\phi^{(3)} s(\phi^{(2)} s(\phi^{(1)} \mathbf{x}_e)))||^2$

Matrix Factorization with Bayesian Personalized Ranking

[Full details in Wijaya et al (EMNLP 2017)]

Derry Wijaya's postdoc was funded by LORELEI. She is now an assistant professor at Boston University.

Bilingual Dictionaries

We need seed bilingual dictionaries to learn the mapping between source and target language embeddings.

We previously created bilingual dictionaries via crowdsourcing between English and 100 other languages.

Derry tested her models on **more than 2 dozen** high and low resource languages.

[Full details in Pavlick et al (TACL 2014)]

Can we use images instead of bilingual dictionaries?

Massively Multilingual Image Dataset (MMID)

100 languages, 10,000 words per language, plus 250K English word translations

100 images per word, 35M images, plus text of web pages they appeared on (20TB of data)

Hosted by Amazon Public Datasets

multilingual-images.org

[Full details in Hewitt, Ippolito et al (ACL 2018)]()

Image-based Translation

Previous papers have tried to learn translations based on visual similarity of images.

- Bergsma and Van Durme (2011) used SIFT+Histogram features
- Kiela et al (2015) used Convolutional Neural Network features

They focused on translating nouns in high resource languages.

New multilingual image corpus

We collect images for the 100 bilingual dictionaries created by Pavlick et al (2014)

100 languages, 10,000 words per language, ~263K English word translations (all POS)

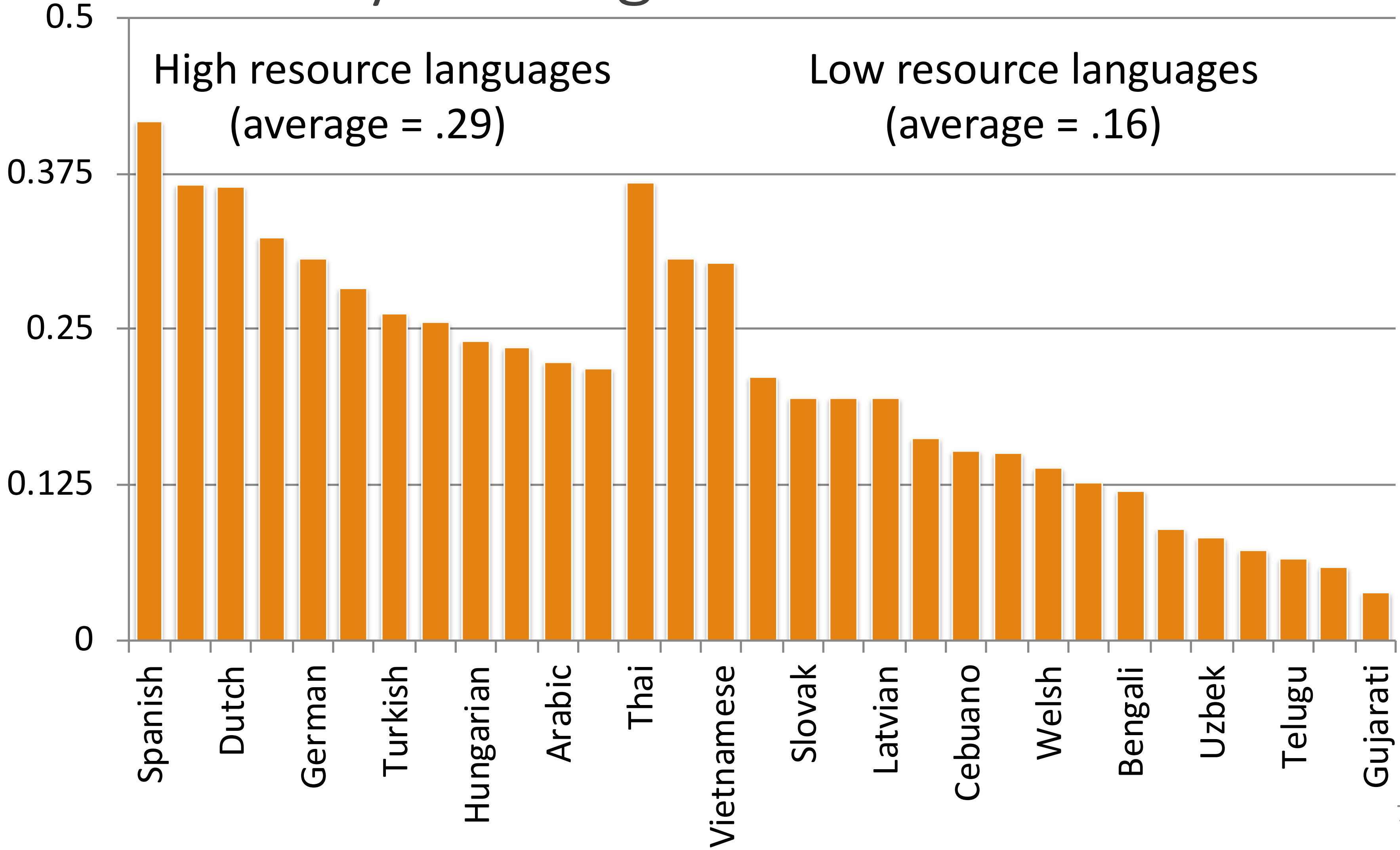
We collected images with Google Image Search

100 images per word, 35M images, 21 TB of data

Example translations

Top 4 English translations for Indonesian word *kucing* by finding k-NN English images using CNN vectors

Accuracy of image-based translation



When does it work?

Nouns and adjectives translate better than verbs and adverbs

Abstract words translate poorly compared to concrete words

Most Concrete Words

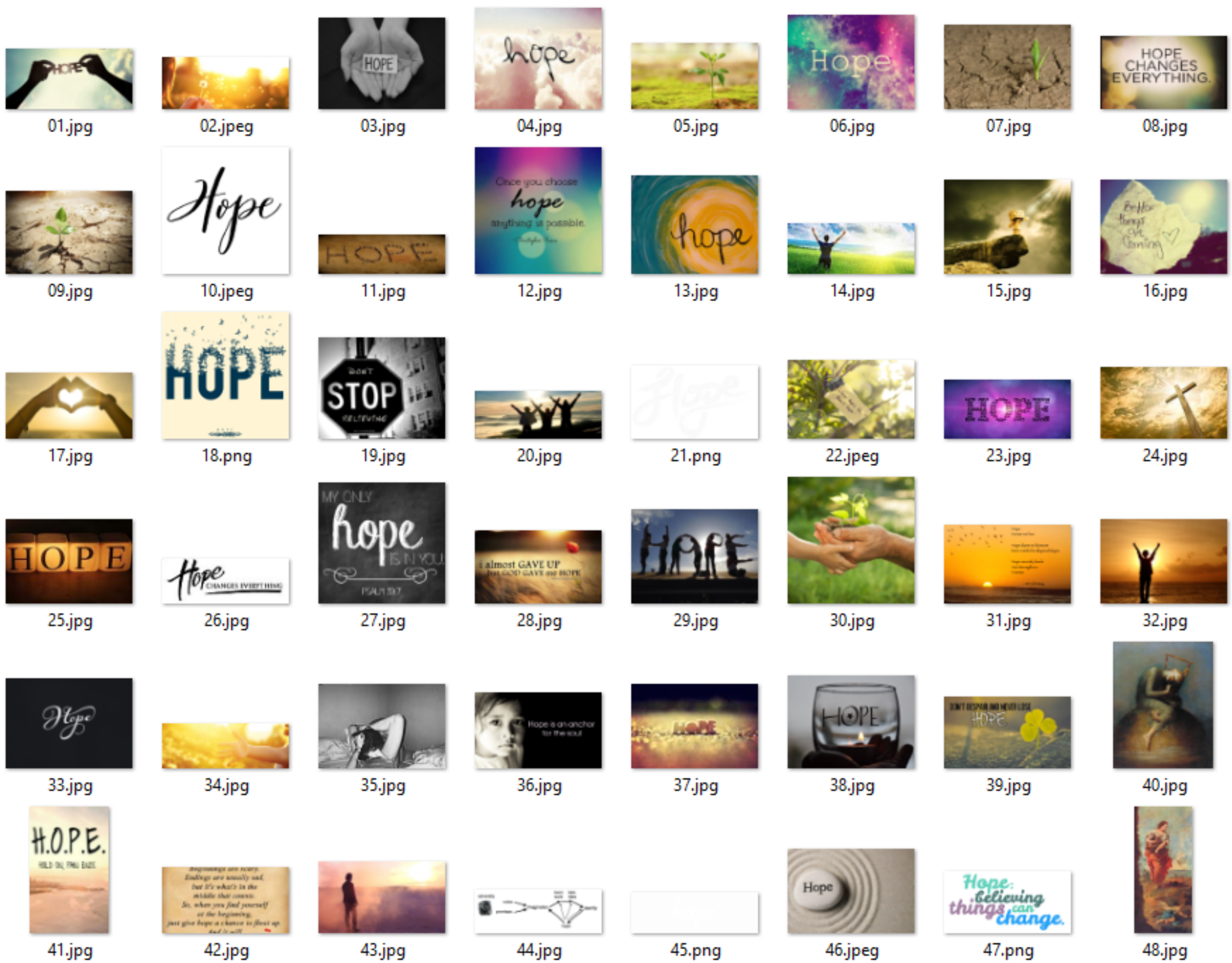
1. tulip - 5.0
2. telescope - 5.0
3. elephant - 5.0
4. bedsheet - 5.0
5. strawberry - 5.0

Most Abstract Words

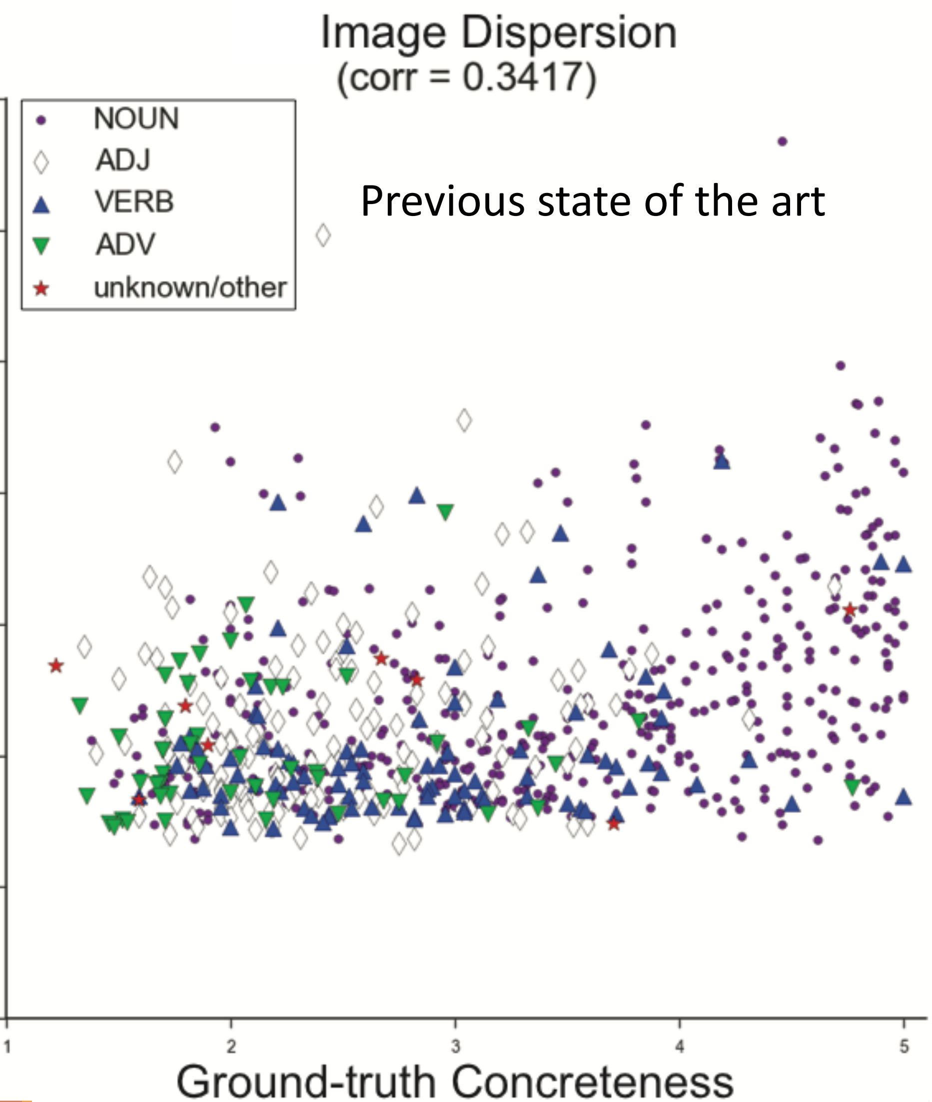
1. essentialness - 1.04
2. hope - 1.04
3. spirituality - 1.07
4. although - 1.07
5. possibility - 1.33

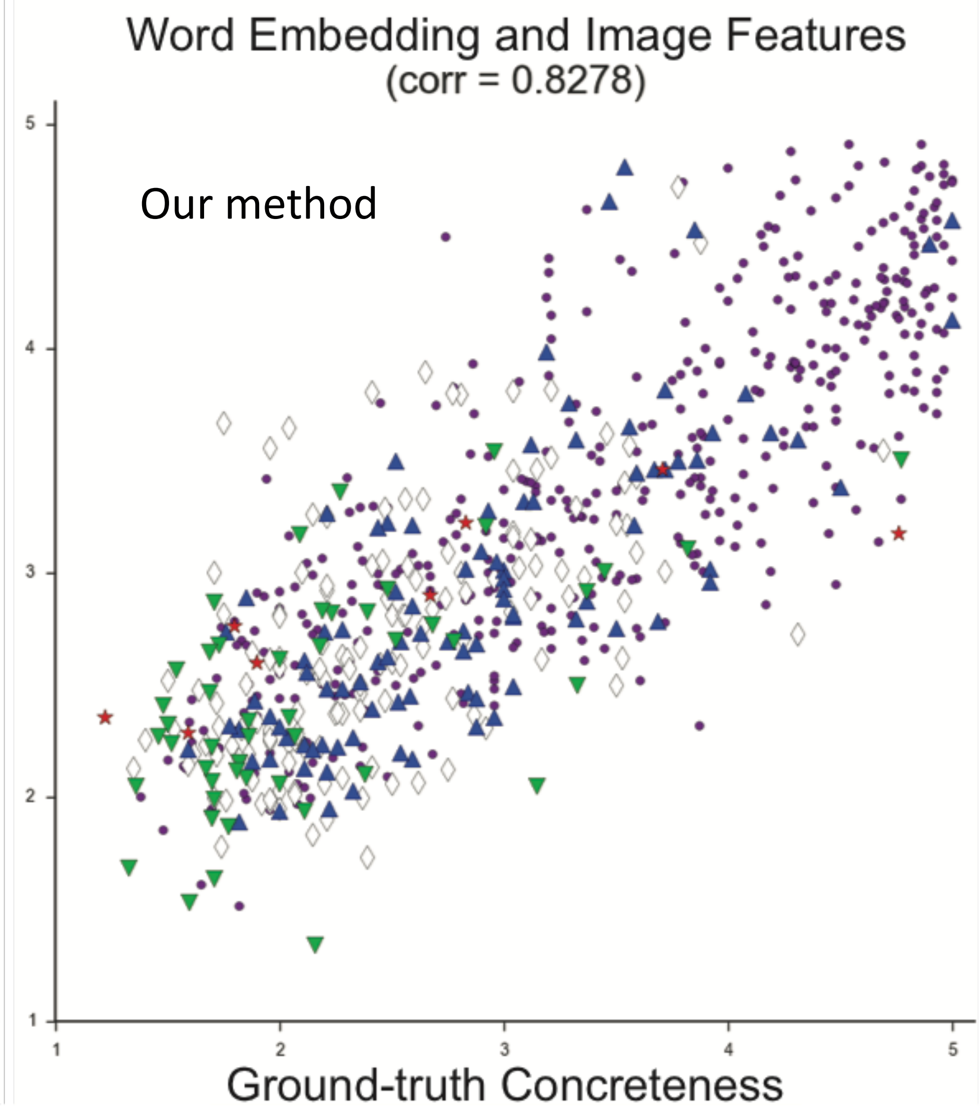
Concrete example

Abstract example

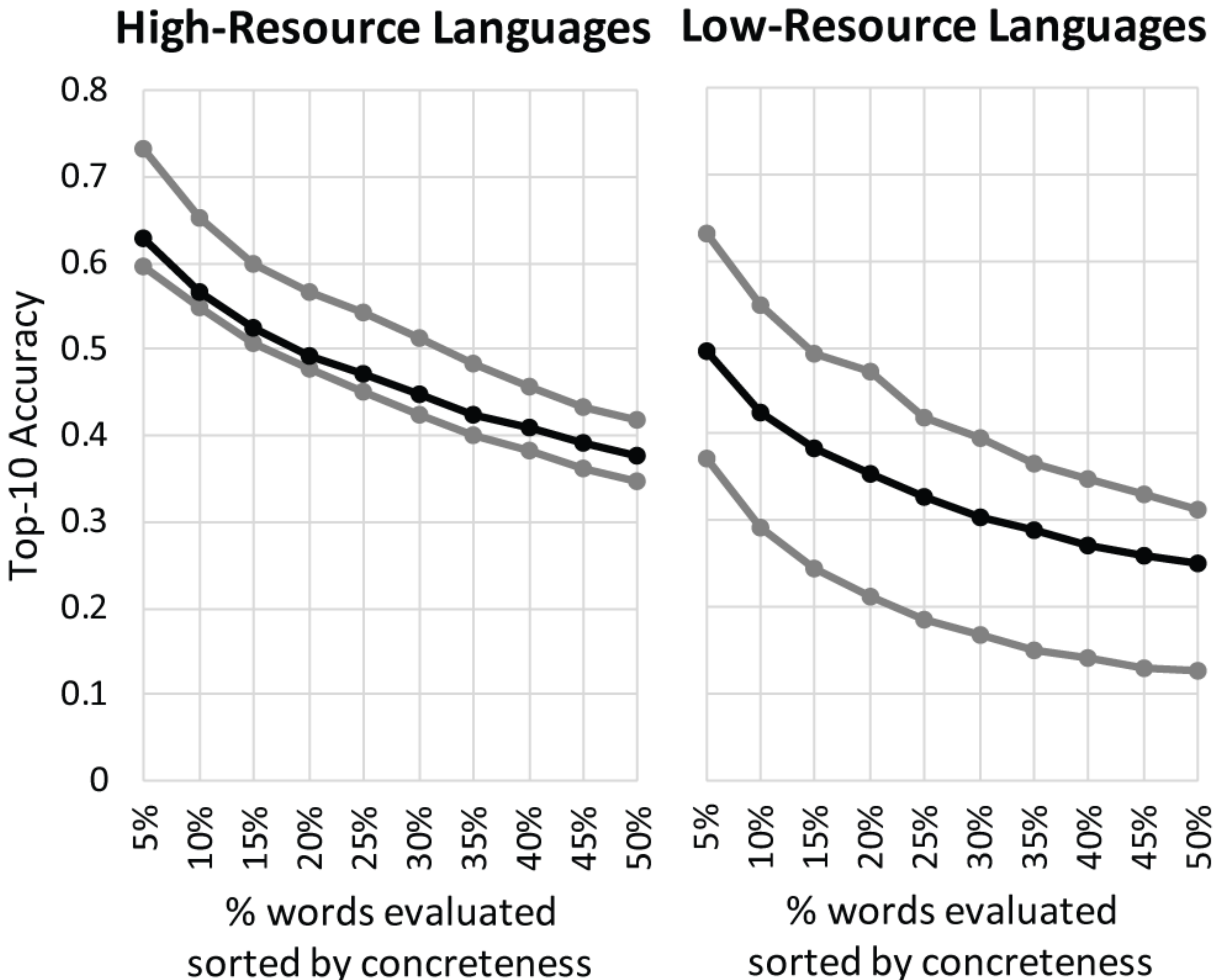


Can we predict concreteness?



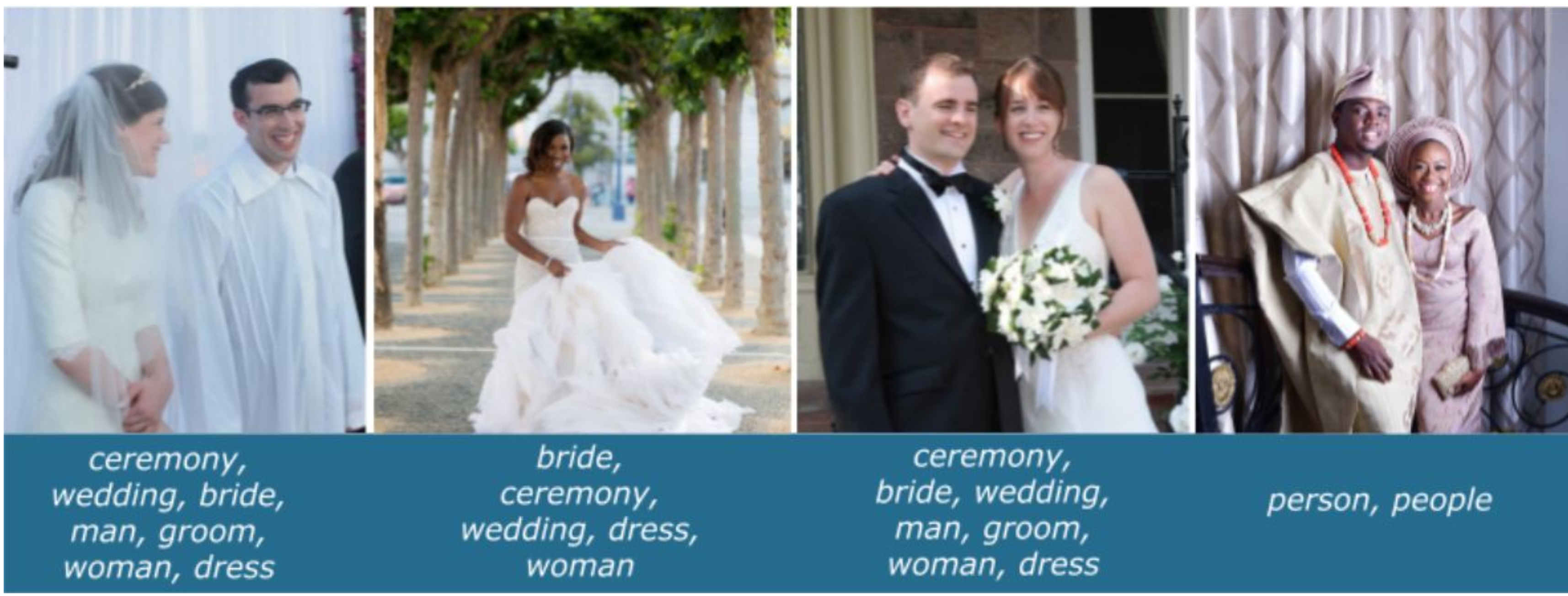


We can produce better translations



Mitigating Geographic Bias of Image Classifiers with MMID

Question: What's wrong with these predictions?



Problem: ~75% of images in ImageNet are from Western countries.

Indian Weddings

Find images by translating wedding into Bengali, Bishnupriya-Manipuri, Gujarati, Hindi, Kannada, Malayalam, Marathi, Punjabi, Tamil, and Telugu

Culturally Divergent Images

Athlete

Children

Farmer

Work by Penn students Yoni Nachmany, Nikhil Krishnan, Aditya Kashyap

Culturally Divergent Images

Police

Military

Wedding

Work by Penn students Yoni Nachmany, Nikhil Krishnan, Aditya Kashyap

References

Comparison of Diverse Decoding Methods from Conditional Language Models. Daphne Ippolito, Reno Kriz, Joao Sedoc, Maria Kustikova and Chris Callison-Burch. ACL 2019.

Magnitude: A Fast, Efficient Universal Vector Embedding Utility Package. Ajay Patel, Alex Sands, Marianna Apidianaki and Chris Callison-Burch. EMNLP 2018. Demo papers.

Learning Translations via Images with a Massively Multilingual Image Dataset. John Hewitt, Daphne Ippolito, Brendan Callahan, Reno Kriz, Derry Wijaya and Chris Callison-Burch. ACL 2018.

Learning Translations via Matrix Completion. Derry Wijaya, Brendan Callahan, John Hewitt, Jie Gao, Xiao Ling, Marianna Apidianaki and Chris Callison-Burch. EMNLP 2017.

A Comprehensive Analysis of Bilingual Lexicon Induction. Ann Irvine and Chris Callison-Burch. Computational Linguistics 2016.

The Language Demographics of Amazon Mechanical Turk. Ellie Pavlick, Matt Post, Ann Irvine, Dmitry Kachaev, and Chris Callison-Burch. TACL 2014.