Language is funny

"Red tape holds up new bridges"

"Hospitals are sued by 7 foot doctors"

"Local high school dropouts cut in half"

"Tesla crashed today"

"Obama announced that he will run again"

"Kipchoge announced that he will run again"

"She made him duck"

"Will you visit the bank across from the river bank? You can bank on it"

"Yes" vs "Yes." vs "YES" vs "YES!" vs "YAS" vs "Yea"

Language is funny

“Maria likes May”

“Maria likes May and Joe”

“Maria likes May and June”

“May likes Maria”

“Maria hit May, then she [fell/ran]”

“Maria and Anqi bullied May, so they got in trouble”

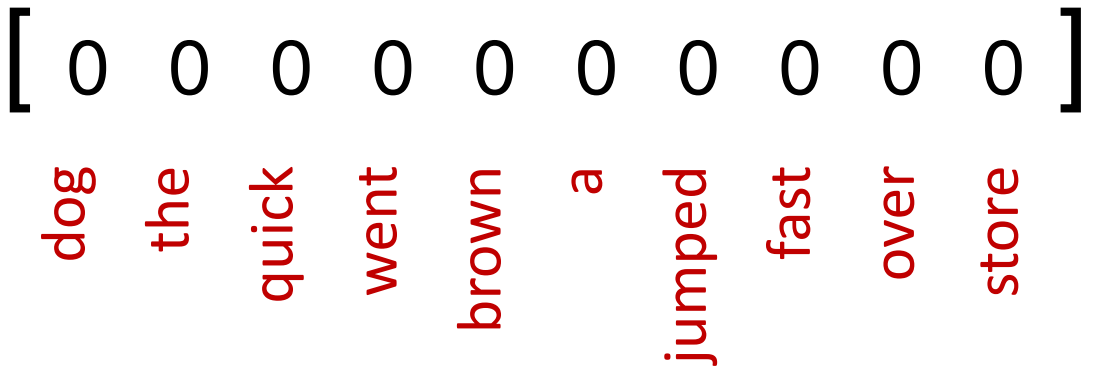
“Maria and Anqi convinced May to prank the teacher, so they got in trouble”

“May may like May, but she really likes June.”

Bag-of-words (BoW)

Let's say our dataset's entire *vocabulary* is just 10 words.

Each unique word can have its own dimension (feature index).

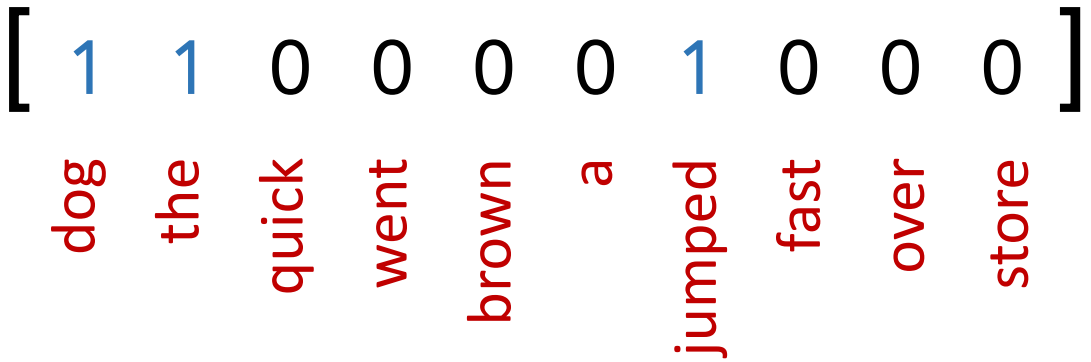


NOTE: This is the Boolean version, which isn't the most popular BoW representation

Bag-of-words (BoW)

Each document's vector has a 1 if the word is present. Otherwise, 0.

e.g., "the dog jumped" is represented as

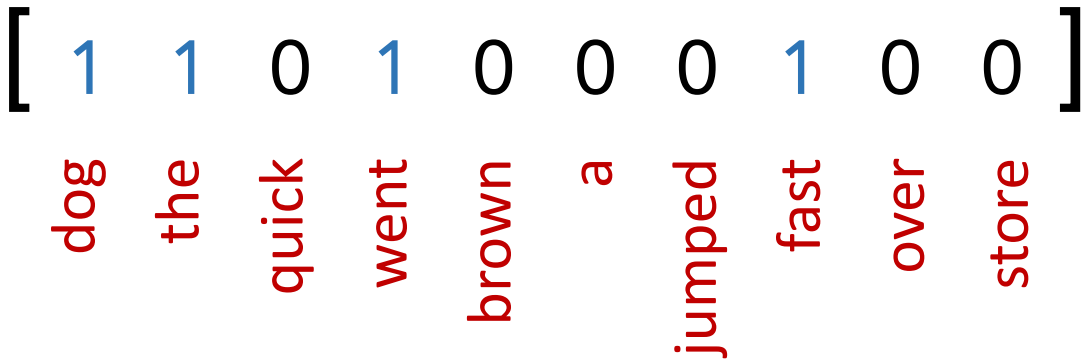


NOTE: This is the Boolean version, which isn't the most popular BoW representation

Bag-of-words (BoW)

Each document's vector has a 1 if the word is present. Otherwise, 0.

e.g., "the dog went fast" is represented as



NOTE: This is the Boolean version, which isn't the most popular BoW representation

Bag-of-words (BoW)

NOTE: The most common way of referring to this is as a "bag-of-words model". Technically, the "bag-of-words" is referring to the representation, not the model.

"bag-of-words model" actually means "Model that uses a bag-of-words representation"

Bag-of-words (BoW)

Are there any weaknesses with this type of representation?

Bag-of-words (BoW)

Weaknesses:

- Flattened view of the document
- Context-insensitive ("the horse ate" = "ate the horse")
- Curse of Dimensionality (vocab could be over 100k)
- Orthogonality: no concept of semantic similarity at the word-level
  - e.g., d(dog, cat) =d(dog, chair)

Bag-of-words (BoW)

Let's address the "flattened view of the document"

Bag-of-words (BoW)

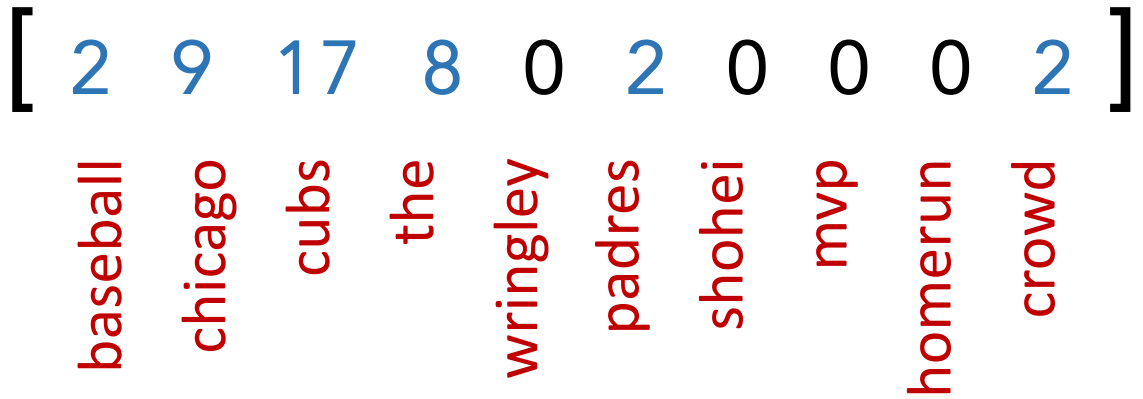
Imagine a document is a sports broadcast transcript, which concerns a few teams but mostly discusses the local home team, the Cubs

Bag-of-words (BoW)

We have no indication of *how much* the document is about the Cubs.

Bag-of-words (BoW)

Now we can see that it's much more about the Chicago Cubs than the Padres.



TF-IDF

Notice that longer documents will naturally have higher counts than shorter documents.

baseball	chicago	cubs	the	wringley	padres	shohei	mvp	homerun	crowd
2	9	17	8	0	2	0	0	0	2

TF-IDF

Also notice that “the” has a fairly high count, too.

baseball	chicago	cubs	the	wringley	padres	shohei	mvp	homerun	crowd
2	9	17	8	0	2	0	0	0	2

TF-IDF

Simple ideas. Let's:

- disproportionately weight the common words that appear in many documents
- Use that info and combine it with the word frequency info

TF-IDF

TF (term frequency) = f_{w_i} = # times word w_i appeared in the document

IDF (inverse document frequency) = $log\left(\frac{\#\text{ docs in corpus}}{\#\text{ docs containing } w_i}\right)$

$$\text{TFIDF} = f_{w_i} * log\left(\frac{\#\text{ docs in corpus}}{\#\text{ docs containing } w_i}\right)$$

TF-IDF

Weaknesses:

- ~~Flattened view of the document~~
- Context-insensitive ("the horse ate" = "ate the horse")
- Curse of Dimensionality (vocab could be over 100k)
- Orthogonality: no concept of semantic similarity at the word-level
  - e.g., d(dog, cat) =d(dog, chair)

Masked language models

I bought a _____ from the bakery

Good morning, _____. Rise and shine!

I got my _____ license last week

word2vec

Two approaches:

1. Continuous Bag-of-Words (CBOW)
2. Skip-gram w/ negative sampling

word2vec: CBOW

Step 1: Iterate through your entire corpus, with sliding context windows of size N and step size 1

Step 2: Using all $2N$ context words, except the center word, try to predict the center word.

Step 3: Calculate your loss and update parameters (like always)

word2vec: CBOW

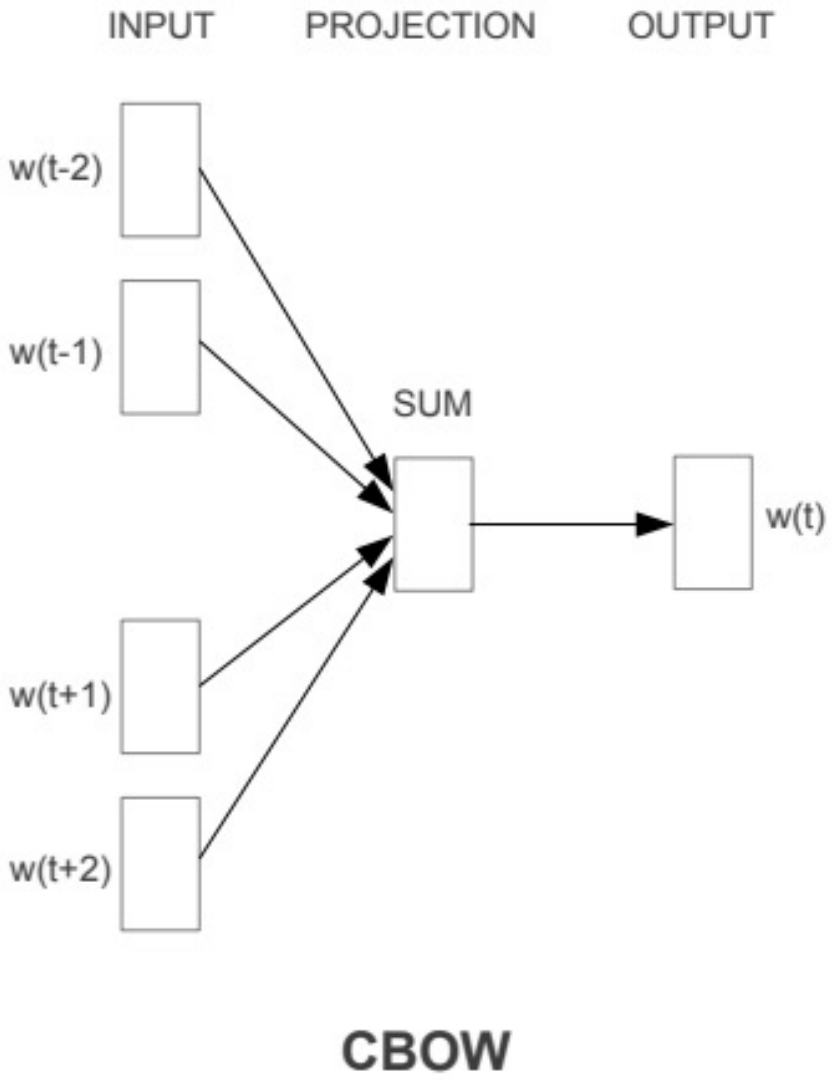


Figure 1: New model architectures. The CBOW architecture predicts the current word based on the context, and the Skip-gram predicts surrounding words given the current word.

word2vec: CBOW

$y = U * \text{sum}(Hx)$

$x = [w_{t-2}, w_{t-1}, w_{t+1}, w_{t+2}]$

N = # total context words

D = embedding size

V = # word types

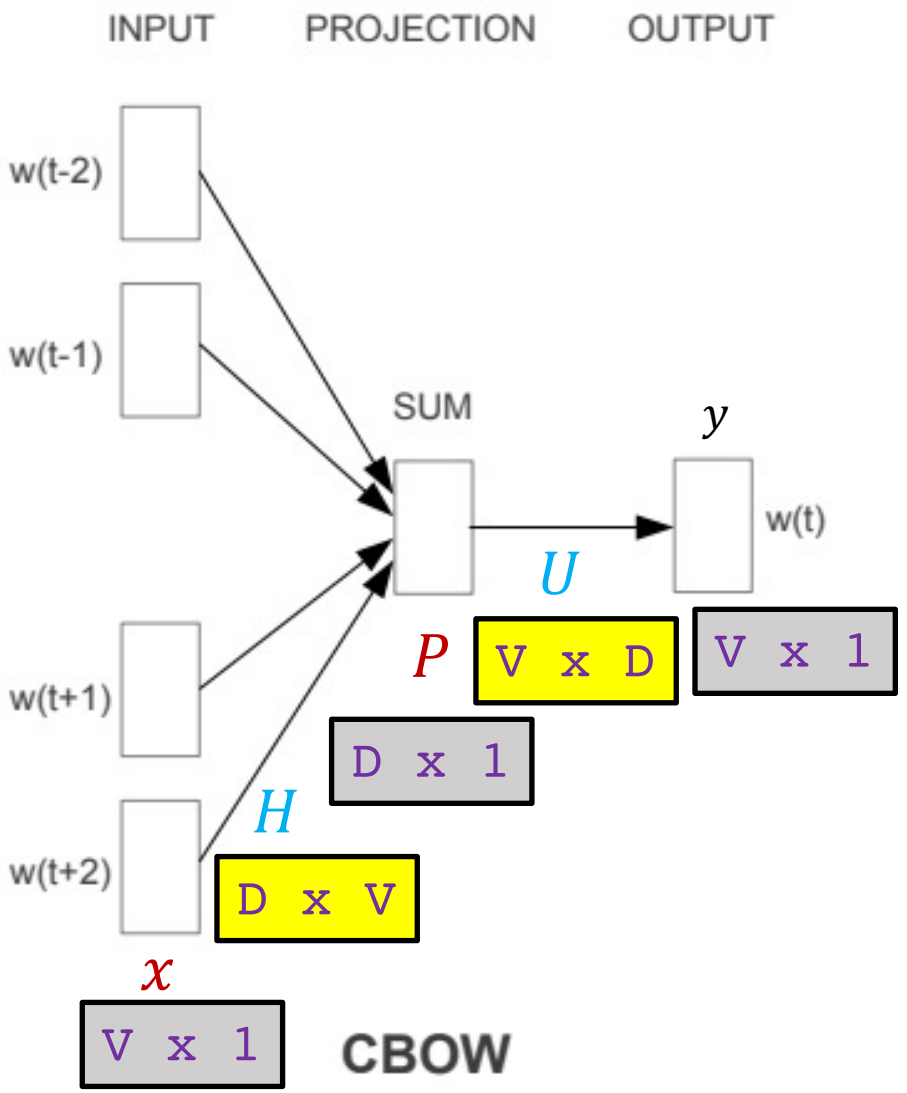


Figure 1: New model architectures. The CBOW architecture predicts the current word based on the context, and the Skip-gram predicts surrounding words given the current word.

word2vec: CBOW

- Linear projection layer
- Non-linear output layer (softmax)
- Training in batches helps a lot

word2vec: skip-gram

Step 1: Iterate through your entire corpus, with sliding context windows of size $\boldsymbol{N}$ and step size $\mathbf{1}$

Step 2: Using the masked center word, try to predict all 2N center words.

Step 3: Calculate your loss and update parameters (like always)

word2vec: skip-gram

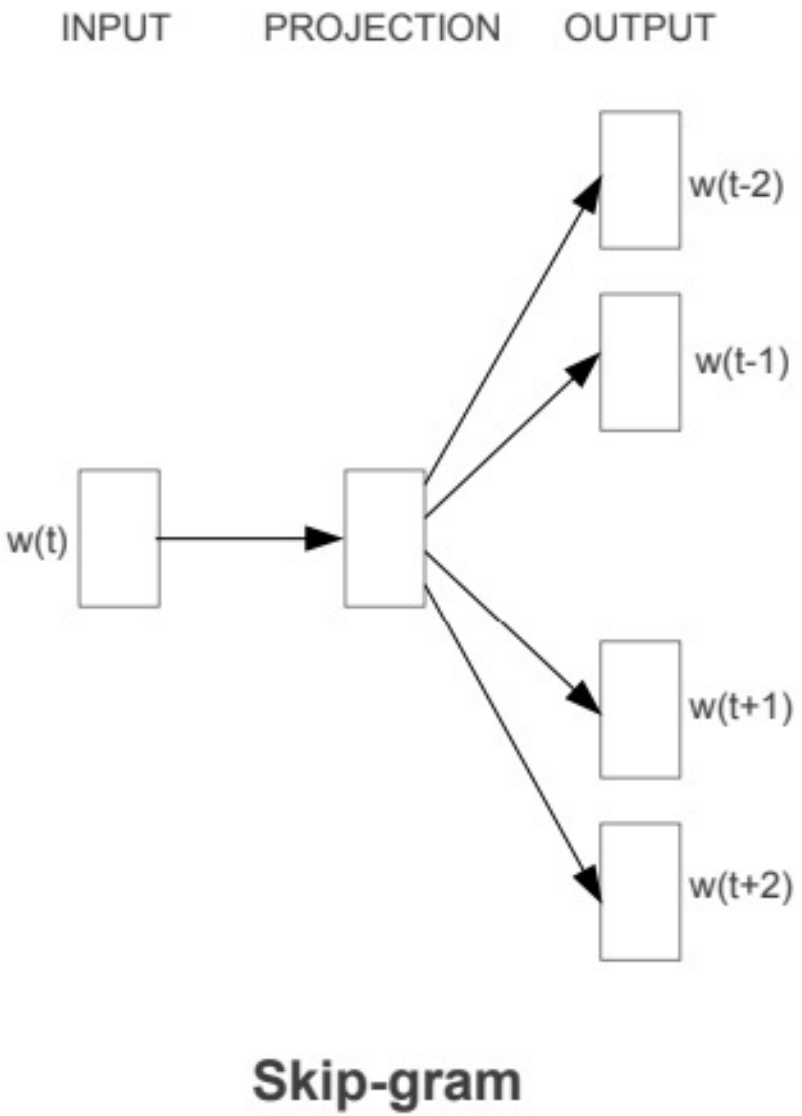


Figure 1: New model architectures. The CBOW architecture predicts the current word based on the context, and the Skip-gram predicts surrounding words given the current word.

word2vec: skip-gram

In practice, this softmax is painfully slow.

Instead, flip the modelling to be pairs of words:

e.g., (center word, context word$_i$)

It would learn to always predict 1. So, probabilistically sample negative examples based on their frequencies

word2vec: results

- Smaller window sizes yield embeddings such that high similarity scores indicates that the words are *interchangeable*
- Larger window sizes (e.g., 15+) yield embeddings such that high similarity is more indicative of *relatedness* of the words.

word2vec: results

- Words that appear in the same contexts are forced to gravitate toward having the same embeddings as one another

- Imagine two words, w_1 and w_2, that never appear together, but they each, individually have the exact same contexts with *other* words. w_1 and w_2 will have ~identical embeddings!

- "The" appears the most. What do you imagine its embedding is like?

word2vec results

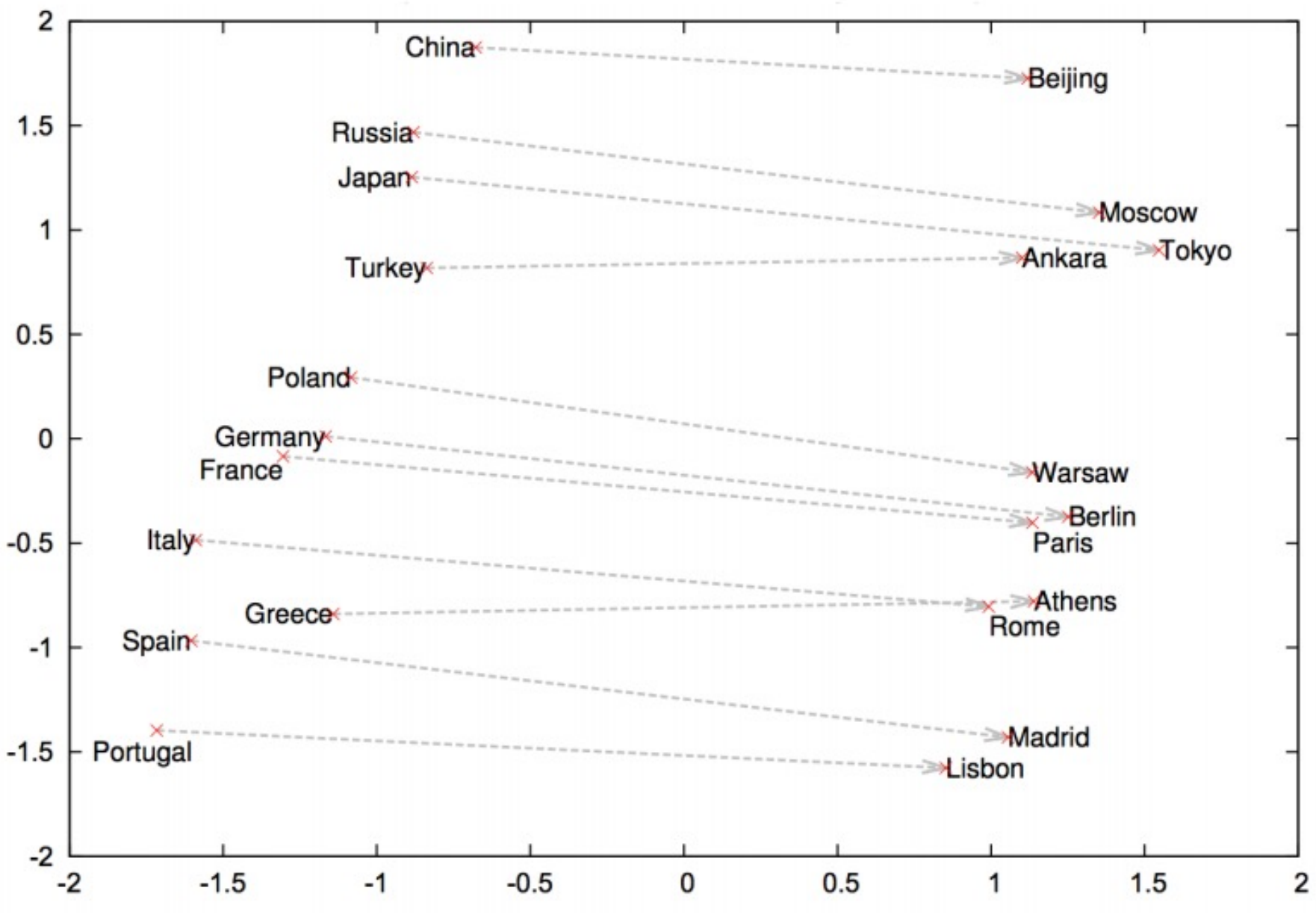


Table 1: *Examples of five types of semantic and nine types of syntactic questions in the Semantic-Syntactic Word Relationship test set.*

Type of relationship	Word Pair 1		Word Pair 2	
Common capital city	Athens	Greece	Oslo	Norway
All capital cities	Astana	Kazakhstan	Harare	Zimbabwe
Currency	Angola	kwanza	Iran	rial
City-in-state	Chicago	Illinois	Stockton	California
Man-Woman	brother	sister	grandson	granddaughter
Adjective to adverb	apparent	apparently	rapid	rapidly
Opposite	possibly	impossibly	ethical	unethical
Comparative	great	greater	tough	tougher
Superlative	easy	easiest	lucky	luckiest
Present Participle	think	thinking	read	reading
Nationality adjective	Switzerland	Swiss	Cambodia	Cambodian
Past tense	walking	walked	swimming	swam
Plural nouns	mouse	mice	dollar	dollars
Plural verbs	work	works	speak	speaks

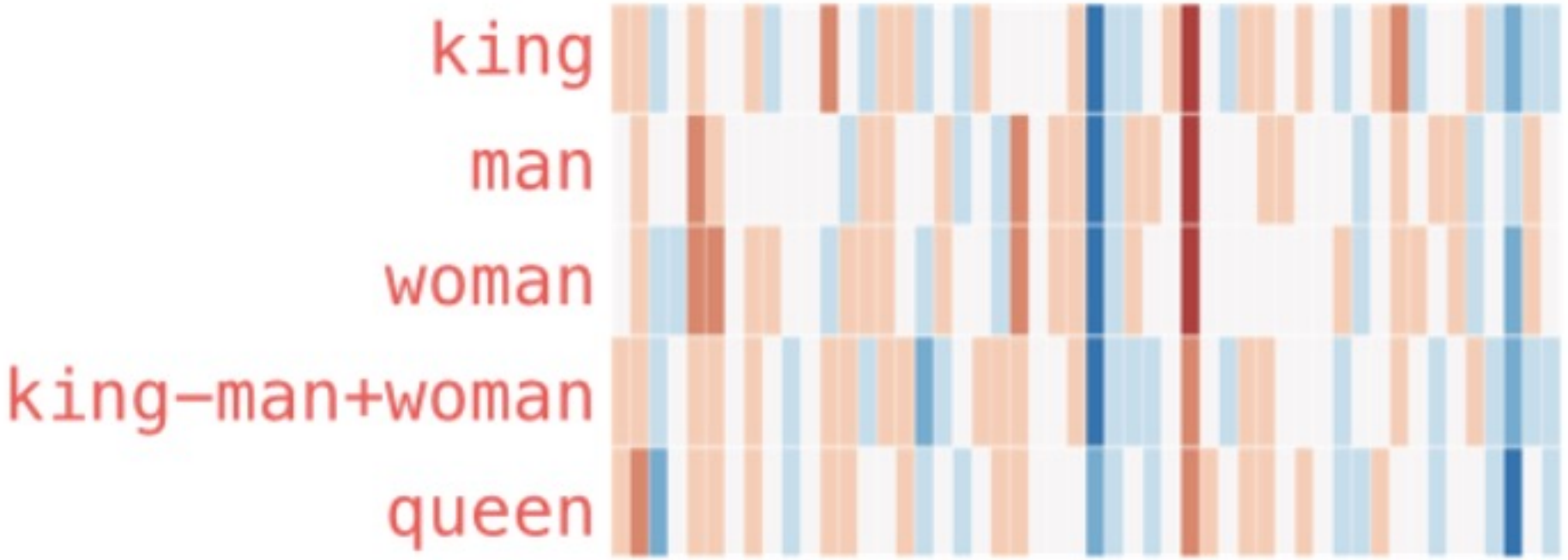
Table 8: *Examples of the word pair relationships, using the best word vectors from Table 4 (Skip-gram model trained on 783M words with 300 dimensionality).*

Relationship	Example 1	Example 2	Example 3
France - Paris	Italy: Rome	Japan: Tokyo	Florida: Tallahassee
big - bigger	small: larger	cold: colder	quick: quicker
Miami - Florida	Baltimore: Maryland	Dallas: Texas	Kona: Hawaii
Einstein - scientist	Messi: midfielder	Mozart: violinist	Picasso: painter
Sarkozy - France	Berlusconi: Italy	Merkel: Germany	Koizumi: Japan
copper - Cu	zinc: Zn	gold: Au	uranium: plutonium
Berlusconi - Silvio	Sarkozy: Nicolas	Putin: Medvedev	Obama: Barack
Microsoft - Windows	Google: Android	IBM: Linux	Apple: iPhone
Microsoft - Ballmer	Google: Yahoo	IBM: McNealy	Apple: Jobs
Japan - sushi	Germany: bratwurst	France: tapas	USA: pizza

word2vec results

Incredible finding!!

king − man + woman ~= queen



word2vec results

Disclaimer: As a heads-up, no models create embeddings such that the dimensions actually correspond to linguistic or real-world phenomenon.

The embeddings are often really great and useful, but no single embedding (in the absence of others) is interpretable.