

Sentiment Mining & Convolutional Networks

(With slides from Yoav Goldberg, Jan Wiebe, Kavita Ganesan,
Heng Ji, Dan Jurafsky, Chris Manning)

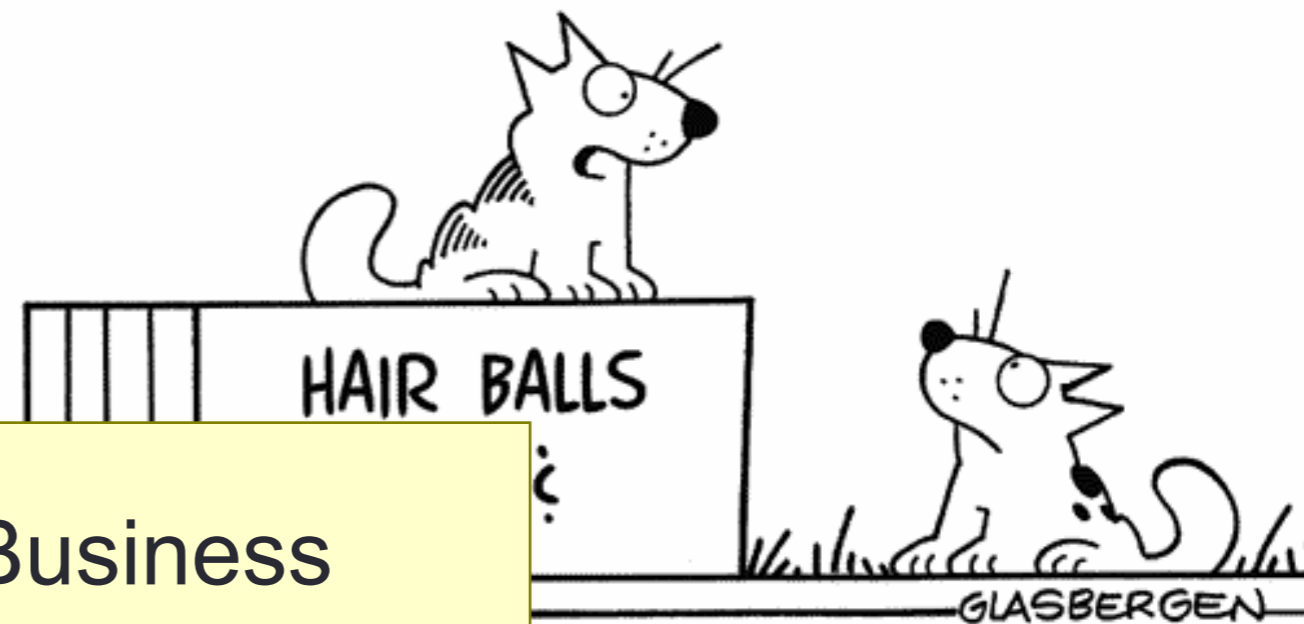
Mausam

Sentiment Analysis on iPhone Reviews

Business' Perspective

- **Apple:** What do consumers think about iPhone?
 - Do they like it?
 - What do they dislike?
 - What are the major complaints?
 - What features should we add?
- **Apple's competitor:**
 - What are iPhone's weaknesses?
 - How can we compete with them?
 - Do people like

Known as Business
Intelligence




sy. Maybe I should have
arket research first.”

Sentiment on Shopping Sites

HP Officejet 6500A E710N Multifunction Printer

[Product summary](#) [Find best price](#) **Customer reviews** [Specifications](#) [Related items](#)



\$121.53 - \$242.39 (14 stores)

Compare

Average rating **★★★★★** (144)



Most mentioned

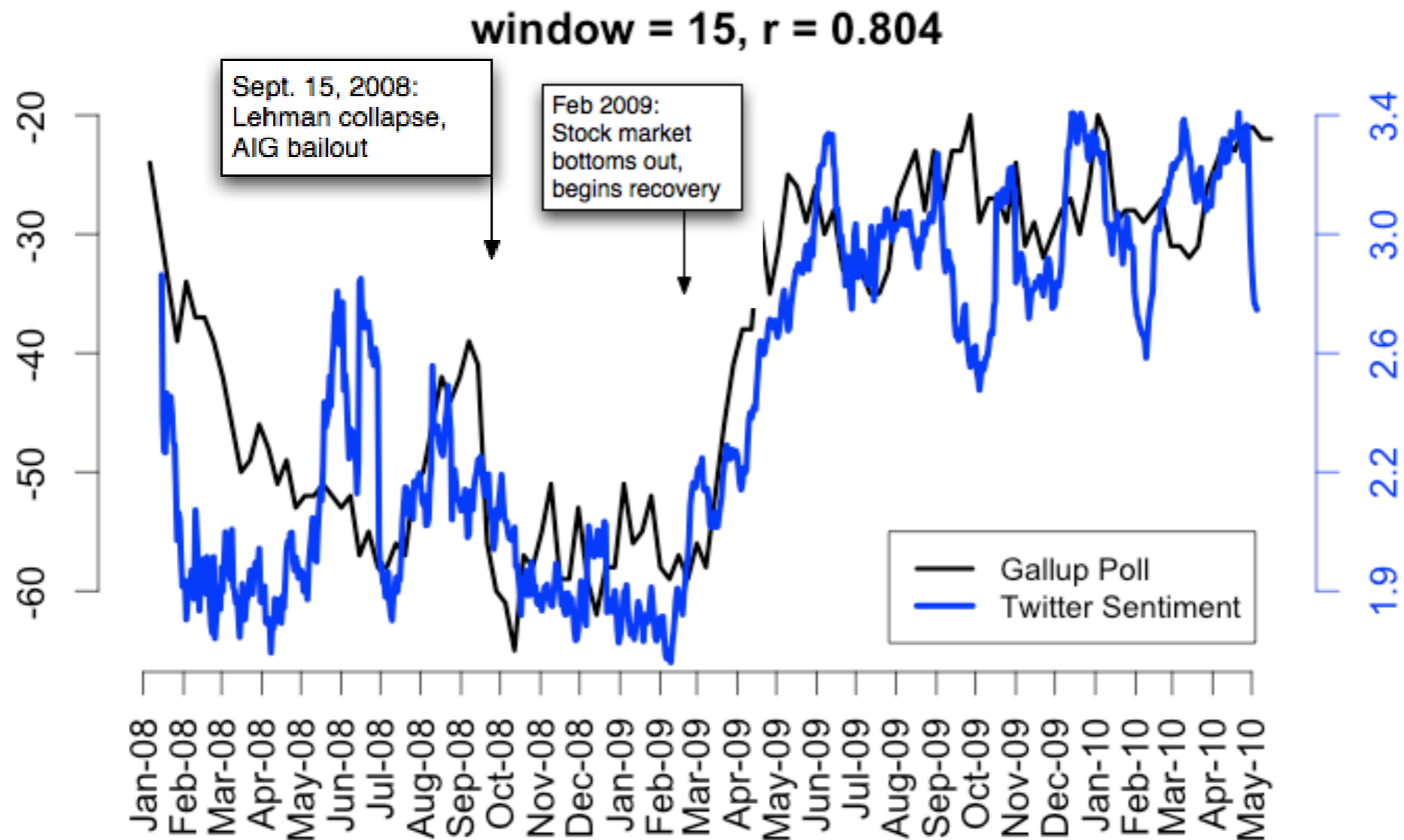


Show reviews by source



Twitter sentiment versus Gallup Poll of Consumer Confidence

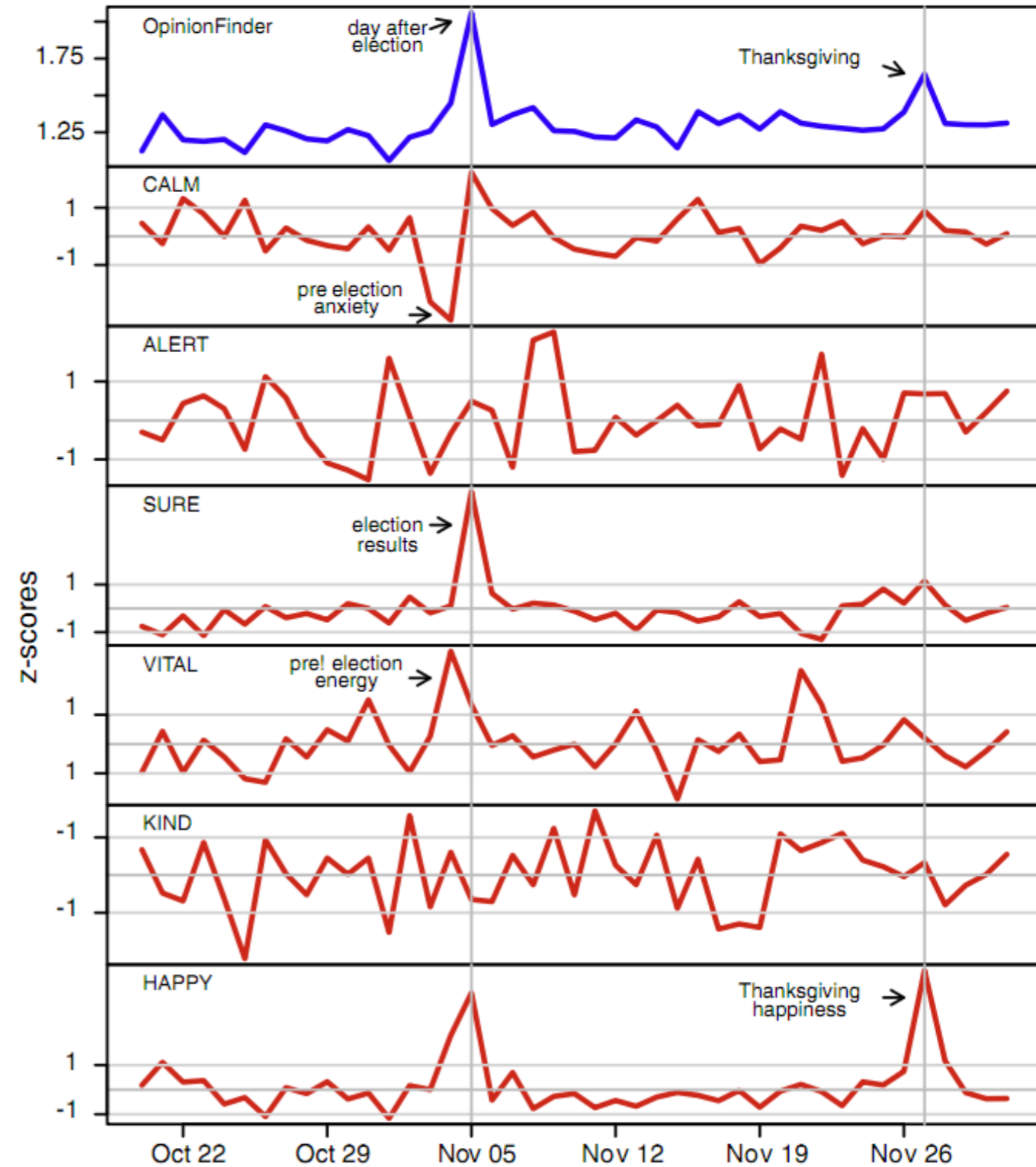
Brendan O'Connor, Ramnath Balasubramanian, Bryan R. Routledge, and Noah A. Smith. 2010. From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series. In ICWSM-2010



Twitter sentiment:

Johan Bollen, Huina Mao, Xiaojun Zeng.
2011. [Twitter mood predicts the stock market](#),

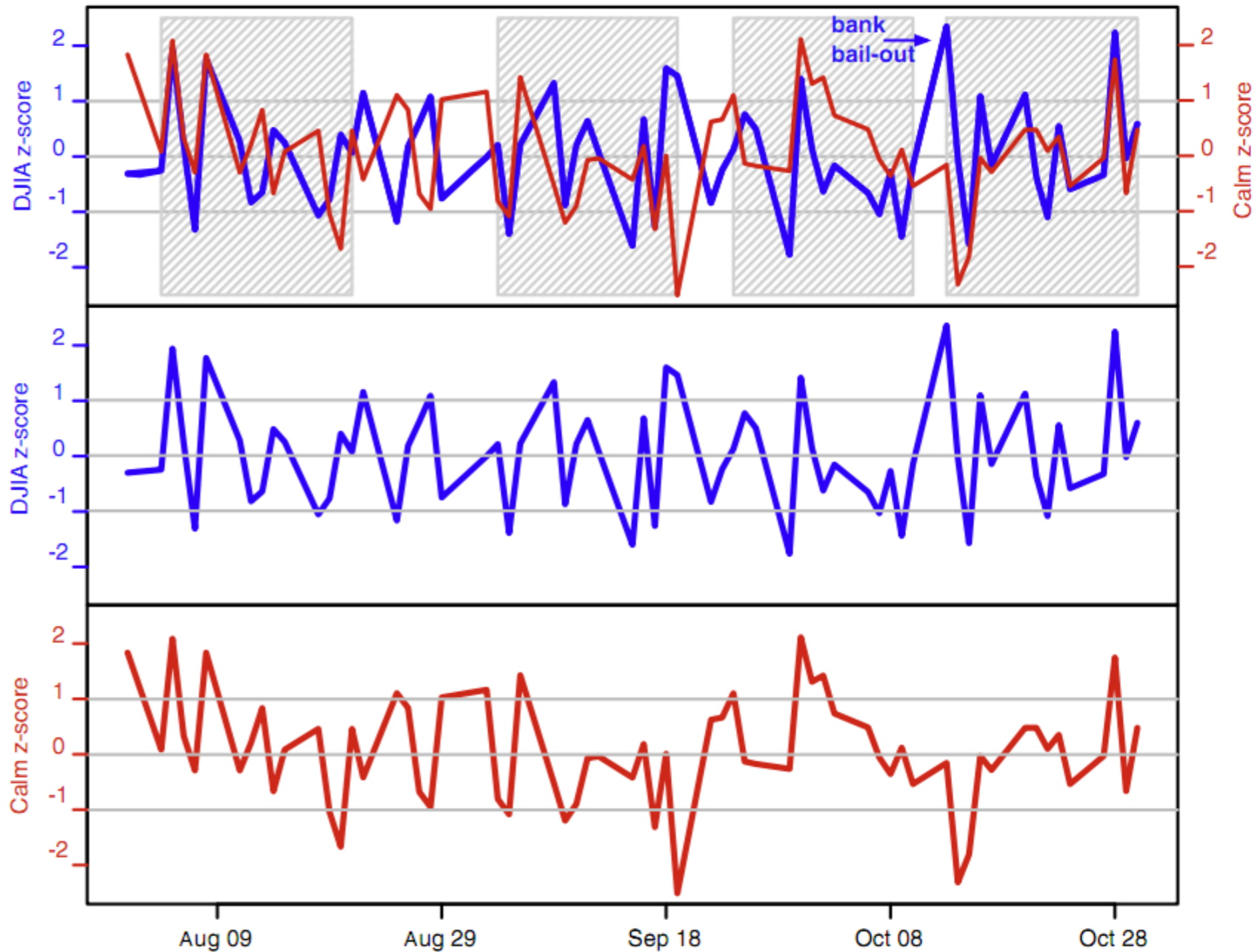
Journal of Computational Science 2:1, 1-8.
[10.1016/j.jocs.2010.12.007](https://doi.org/10.1016/j.jocs.2010.12.007).



Bollen et al. (2011)

- CALM today predicts DJIA 3 days later
- At least one current hedge fund uses this algorithm

CALM Dow Jones



Definition

Sentiment Analysis

- Sentiment analysis is the detection of **attitudes**
 - “enduring, affectively colored beliefs, dispositions towards objects or persons”
 - 1. **Holder (source)** of attitude
 - 2. **Target (aspect)** of attitude
 - 3. **Type** of attitude
 - From a set of types
 - *Like, love, hate, value, desire, etc.*
 - Or (more commonly) simple weighted **polarity**:
 - *positive, negative, neutral, together with strength*
 - 4. **Text** containing the attitude
 - Sentence or entire document

Sentiment Analysis

- Simplest task:
 - Is the attitude of this text positive or negative?
- More complex:
 - Rank the attitude of this text from 1 to 5
- Advanced:
 - Detect the target, source, or complex attitude types

Baseline Algorithms

Sentiment Classification in Movie Reviews

Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up? Sentiment Classification using Machine Learning Techniques. EMNLP-2002, 79—86.

Bo Pang and Lillian Lee. 2004. A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts. ACL, 271-278

- Polarity detection:
 - Is an IMDB movie review positive or negative?
- Data: *Polarity Data 2.0*:
 - <http://www.cs.cornell.edu/people/pabo/movie-review-data>

IMDB data in the Pang and Lee database



when `_star wars_` came out some twenty years ago , the image of traveling throughout the stars has become a commonplace image . [...]

when han solo goes light speed , the stars change to bright lines , going towards the viewer in lines that converge at an invisible point .

cool .

`_october sky_` offers a much simpler image—that of a single white dot , traveling horizontally across the night sky . [...]

“ snake eyes ” is the most aggravating kind of movie : the kind that shows so much potential then becomes unbelievably disappointing .

it's not just because this is a brian depalma film , and since he's a great director and one who's films are always greeted with at least some fanfare .

and it's not even because this was a film starring nicolas cage and since he gives a brauvara performance , this film is hardly worth his talents .

Starting Point

- BoW features
- Word2Vec → Paragraph Vector → Classifier
- Challenge?

Extracting Features for Sentiment Classification

- How to handle negation
 - I **didn't** like this movie
 - vs
 - I really like this movie
- Which words to use?
 - Only adjectives
 - All words
 - All words turns out to work better, at least on this data

Negation

Das, Sanjiv and Mike Chen. 2001. Yahoo! for Amazon: Extracting market sentiment from stock message boards. In Proceedings of the Asia Pacific Finance Association Annual Conference (APFA).

Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up? Sentiment Classification using Machine Learning Techniques. EMNLP-2002, 79—86.

Add NOT_ to every word between negation and following punctuation:

didn't like this movie , but I



didn't NOT_like NOT_this NOT_movie but I

Accounting for Negation

- Let us consider the following positive sentence:
 - Example: *Luckily, the **smelly poo** did not leave **awfully nasty stains** on my **favorite** shoes!*
- Rest of Sentence (RoS):
 - Following: *Luckily, the **smelly poo** did not leave awfully nasty stains on my favorite shoes!*
 - Around: *Luckily, the smelly poo did not leave awfully nasty stains on my favorite shoes!*
- First Sentiment-Carrying Word (FSW):
 - Following: *Luckily, the **smelly poo** did not leave awfully nasty stains on my **favorite** shoes!*
 - Around: *Luckily, the **smelly poo** did not leave awfully nasty stains on my **favorite** shoes!*

Accounting for Negation

- Let us consider the following positive sentence:
 - Example: *Luckily, the **smelly poo** did not leave **awfully nasty stains** on my **favorite** shoes!*
- Next Non-Adverb (NNA):
 - Following: *Luckily, the **smelly poo** did not leave **awfully nasty stains** on my **favorite** shoes!*
- Fixed Window Length (FWL):
 - Following (3): *Luckily, the **smelly poo** did not leave awfully nasty stains on my **favorite** shoes!*
 - Around (3): *Luckily, the smelly poo did not leave awfully nasty stains on my **favorite** shoes!*

KEYWORDS SELECTION FROM TEXT

- Pang et. al. (2002)
 - Binary Classification of unigrams
 - Positive
 - Negative
 - Unigram method reached 80% accuracy.

N-GRAM BASED CLASSIFICATION

- Learn N-Grams (frequencies) from pre-annotated training data.
- Use this model to classify new incoming sample.

Importance of Ngrams

- What is the equivalent of bigram features in embedding world?
- While we can ignore global order in many cases...
- ... local ordering is still often very important.
- Local sub-sequences encode useful structures.
E.g., negation

(so why not just assign a vector to each ngram?)

"feature embeddings"

- Each feature is assigned a vector.
- The input is a combination of feature vectors.
- The feature vectors are **parameters of the model** and are trained jointly with the rest of the network.
- **Representation Learning**: similar features will receive similar vectors.

ConvNets

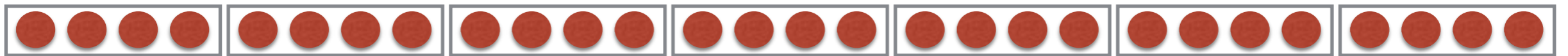
special architecture for local predictors

ConvNets

- CBOW allows encoding arbitrary length sequences, but loses all order information.
- Some local order (i.e. bigrams, trigrams) is informative. Yet, we do not care about exact position in the sequence. (think "good" vs. "not good")
- ConvNets (in language) allow to identify informative local predictors.
- Works by moving a shared function (feature extractor) over a sliding window, then pooling results.

ConvNets

- ConvNets have huge success in computer vision.
- It allows invariance to object position.
- It allows composing large predictors from small.



the actual service was not very good



dot



the

actual

service

was

not

very

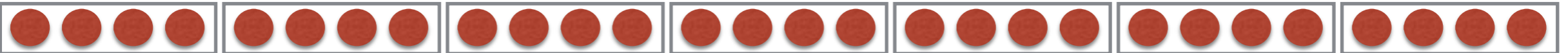
good



||



dot



the

actual

service

was

not

very

good

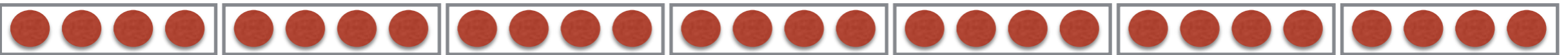
the actual



||



dot



the

actual

service

was

not

very

good

the actual

actual service



||



dot



the

actual

service

was

not

very

good

the actual

actual service

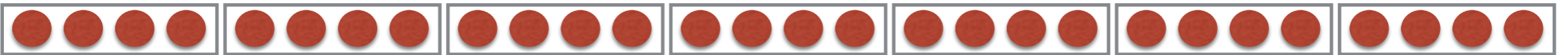
service was



||



dot



the

actual

service

was

not

very

good

the actual

actual service

service was

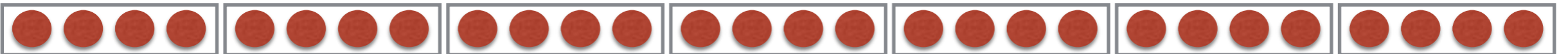
was not



||



dot



the

actual

service

was

not

very

good

the actual

actual service

service was

was not

not very



||



dot



the

actual

service

was

not

very

good

the actual

actual service

service was

was not

not very

very good



||



dot



the

actual

service

was

not

very

good

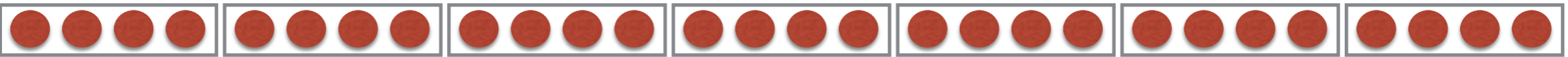
the actual



||



dot



the

actual

service

was

not

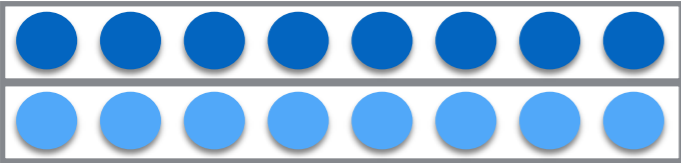
very

good

the actual



||



dot



the

actual

service

was

not

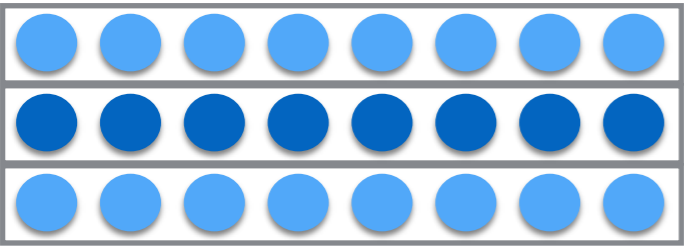
very

good

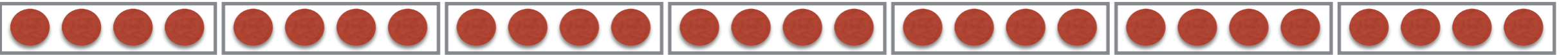
the actual



||



dot



the

actual

service

was

not

very

good

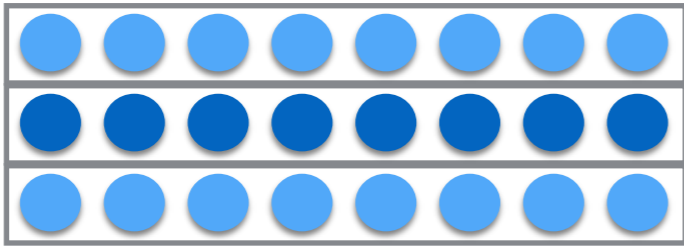
the actual



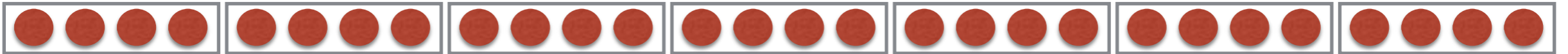
actual service



||



dot



the

actual

service

was

not

very

good

the actual



actual service



service was



was not



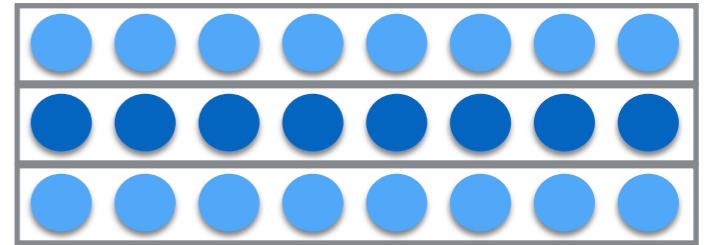
not very



very good



||



dot



the

actual

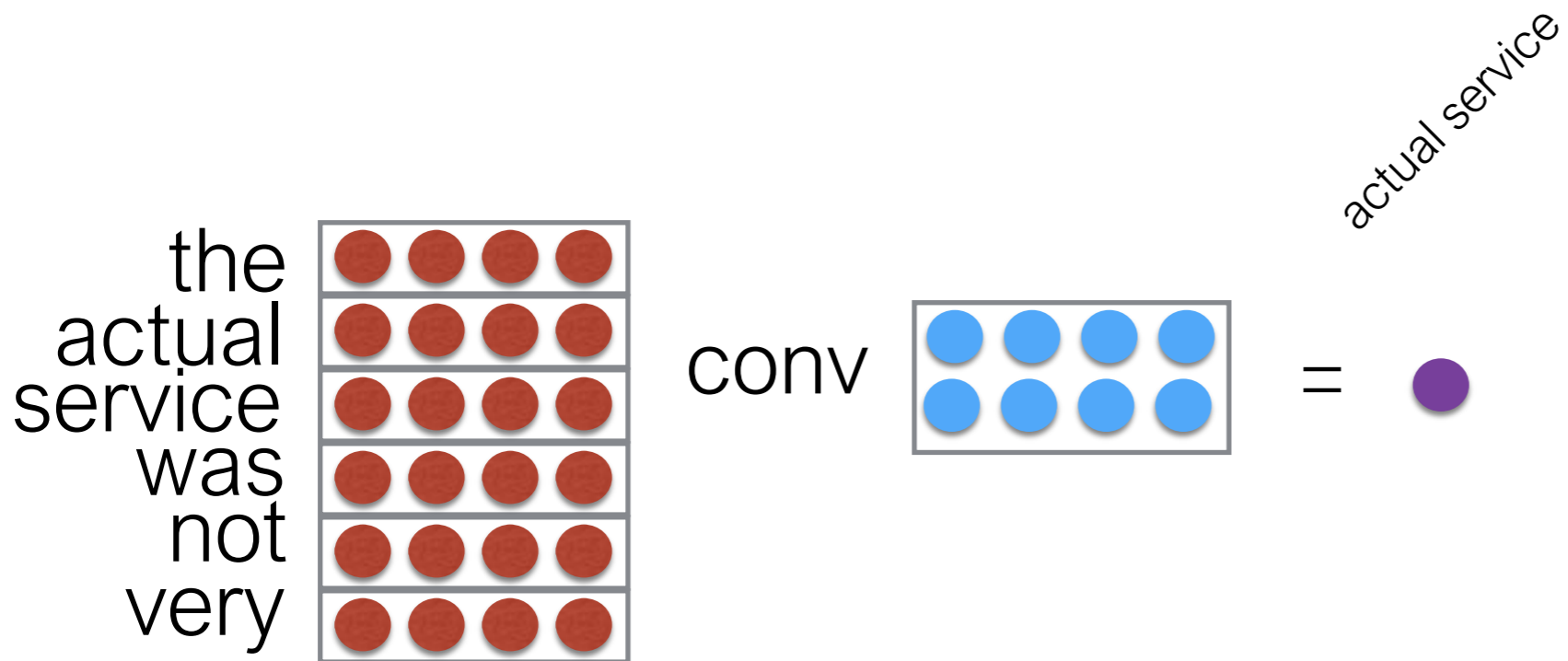
service

was

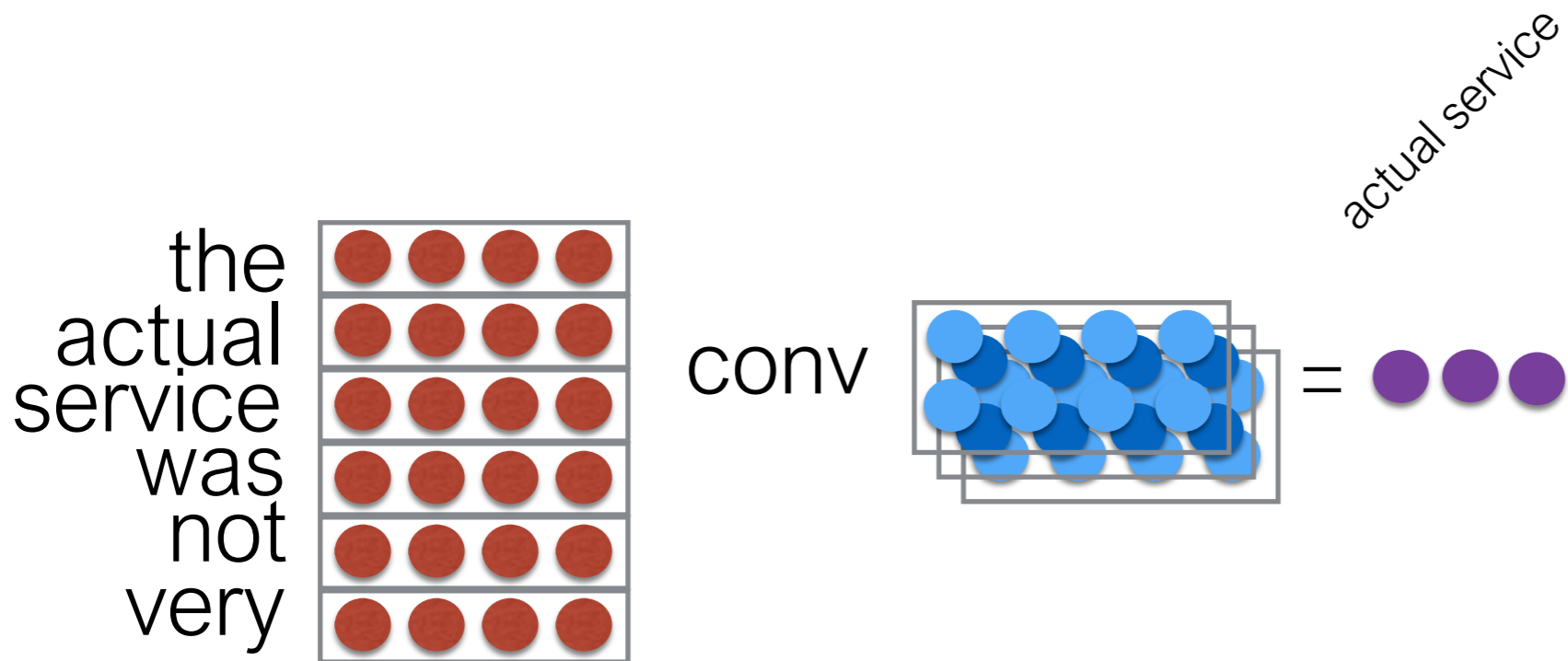
not

very

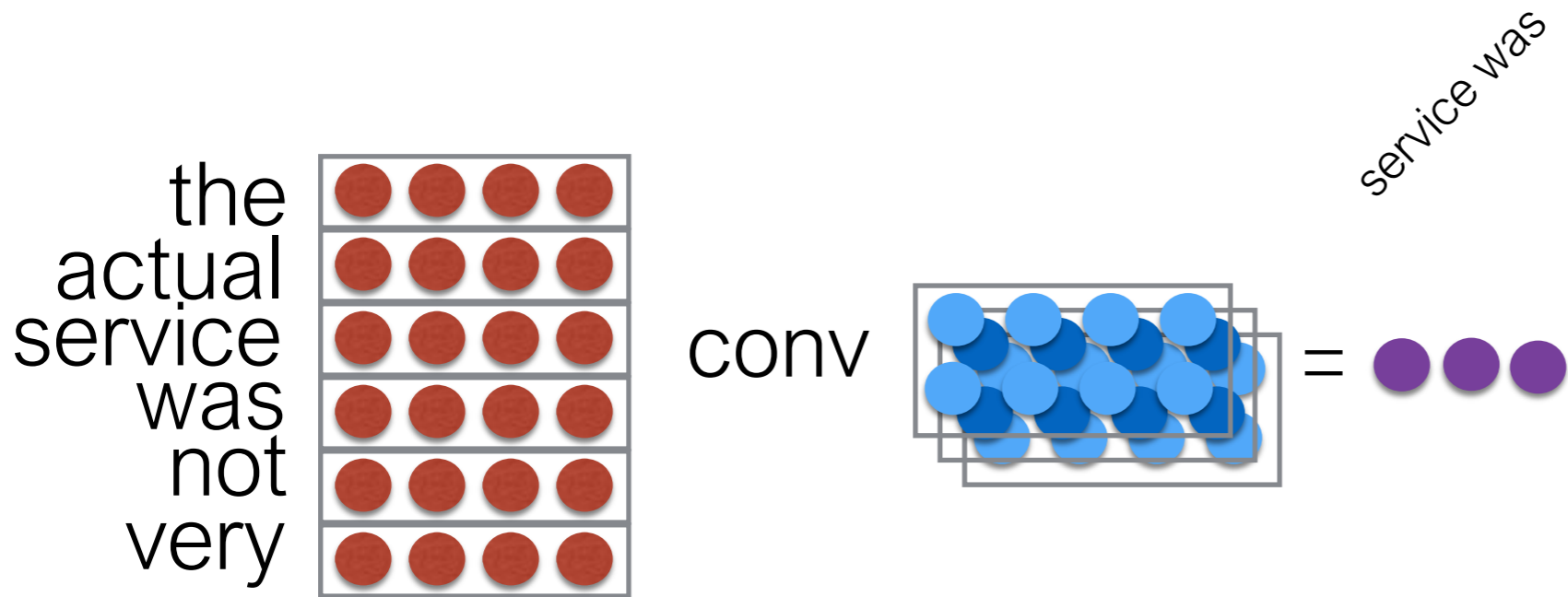
good



(another way to represent text convolutions)



(another way to represent text convolutions)



(another way to represent text convolutions)

the actual



actual service



service was



was not



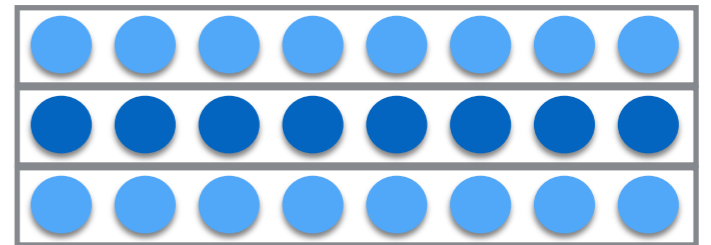
not very



very good



||



dot



the

actual

service

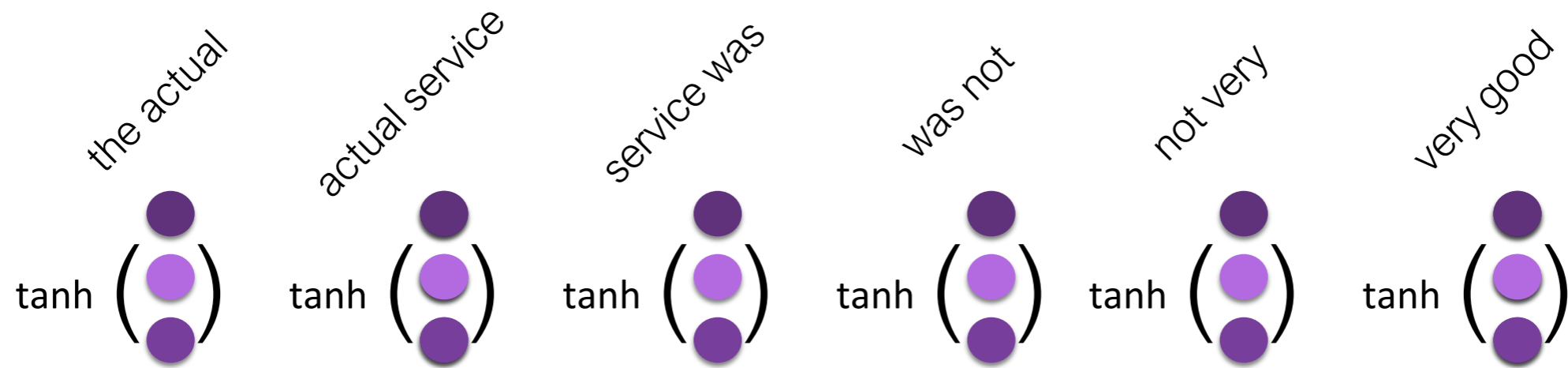
was

not

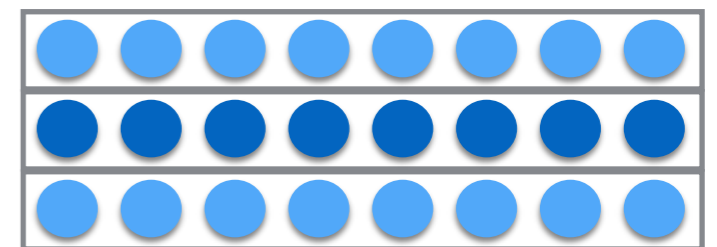
very

good

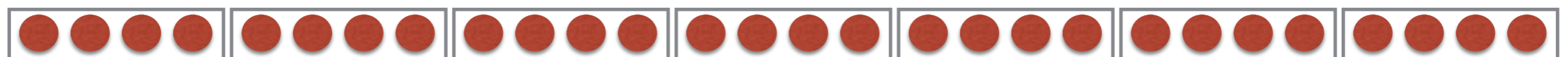
**(we'll focus on the 1-d view here,
but remember they are equivalent)**



||



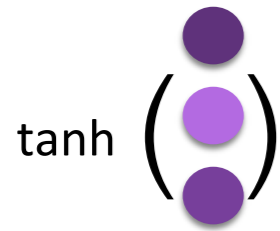
dot



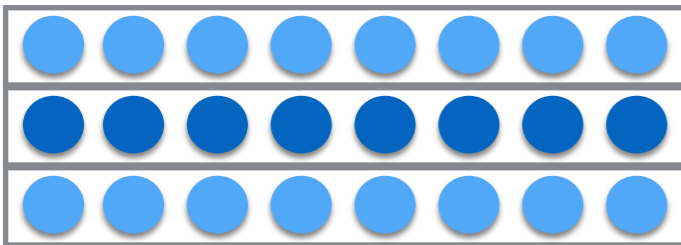
the actual service was not very good

(usually also add non linearity)

the actual



||



dot



the

actual

service

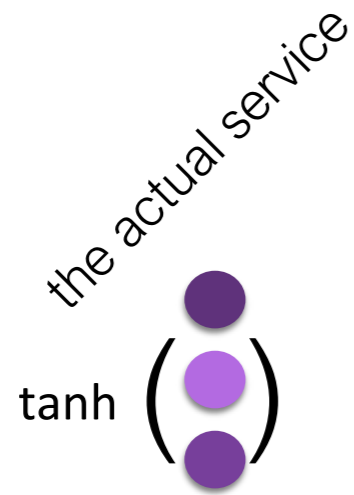
was

not

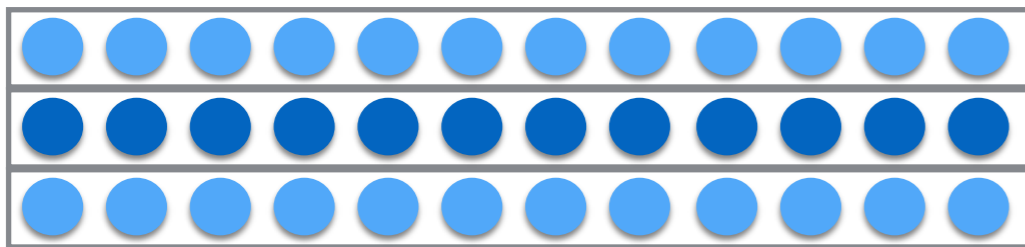
very

good

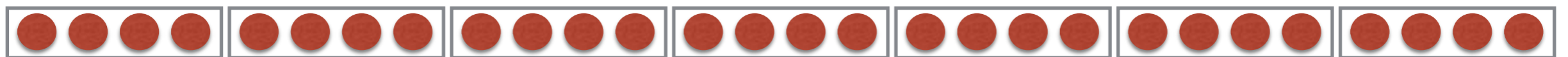
(can have larger filters)



||



dot



the

actual

service

was

not

very

good

(can have larger filters)

the actual



actual service



service was



was not



not very



very good



the

actual

service

was

not

very

good

we have the ngram vectors. now what?

the actual



+

actual service



+

service was



+

was not



+

not very



+

very good



=



the

actual

service

was

not

very

good

can do "pooling"

"Pooling"

Combine K vectors into a single vector

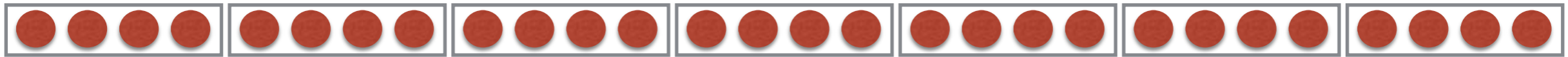
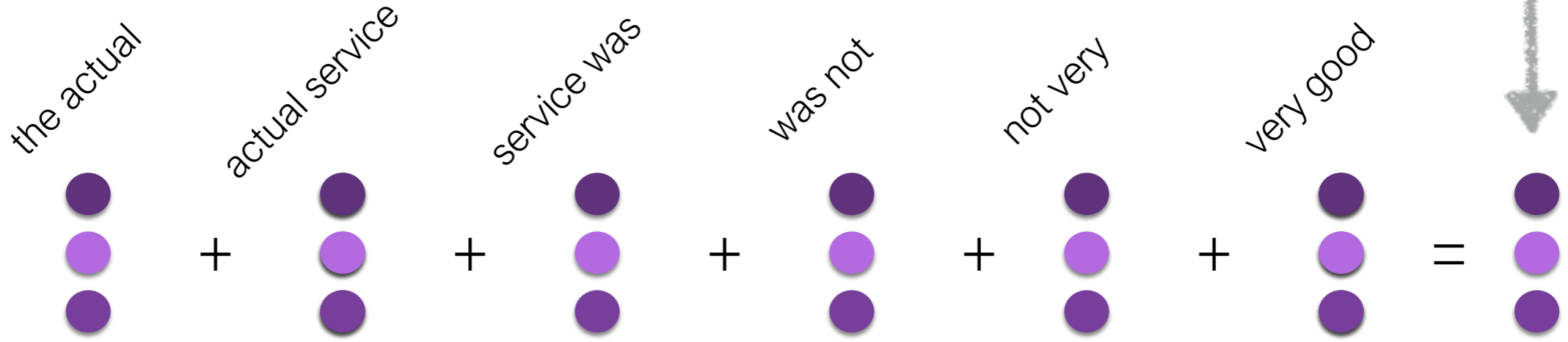
"Pooling"

Combine K vectors into a single vector

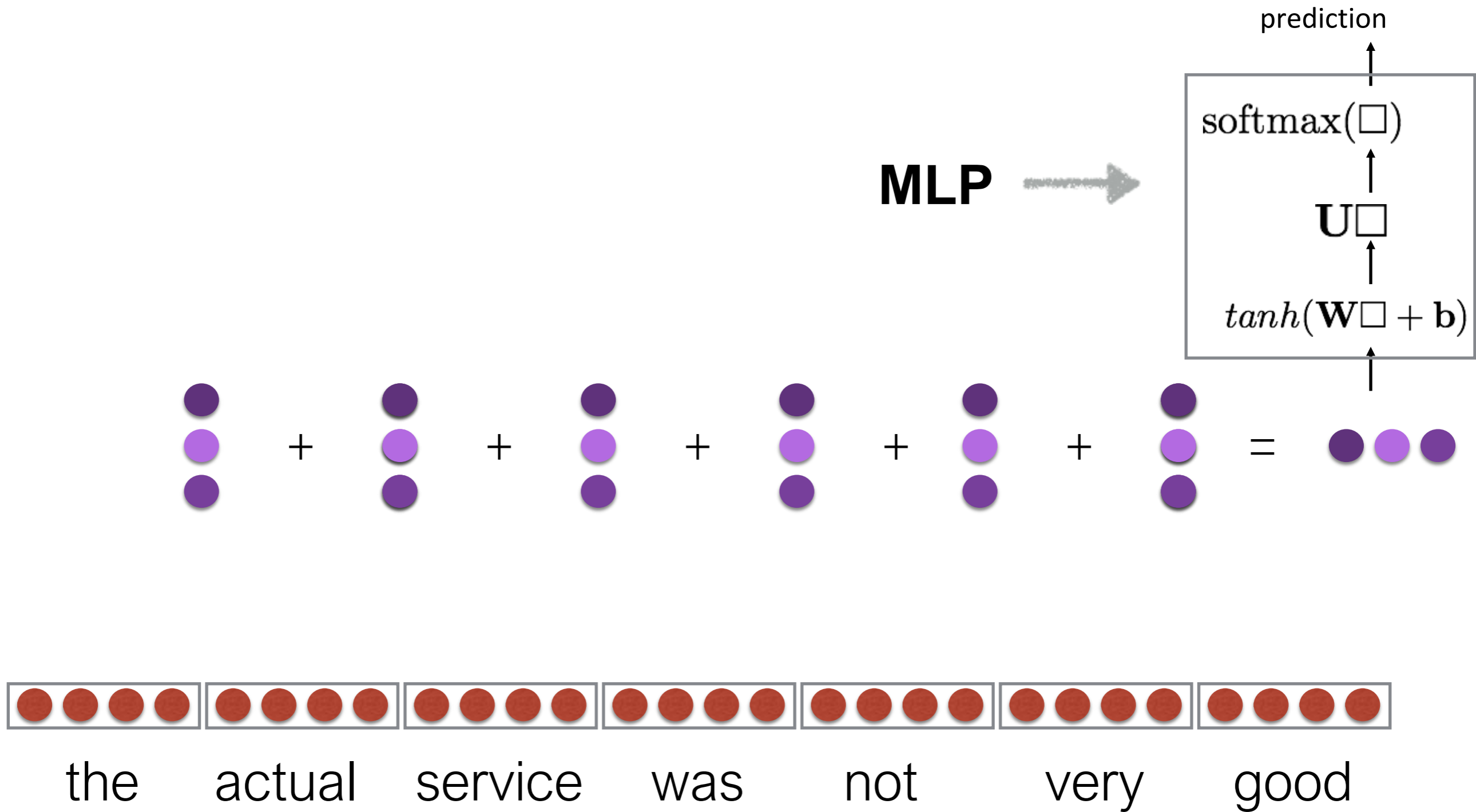
**This vector is a summary of the K vectors,
and can be used for prediction.**

average pooling

average vector

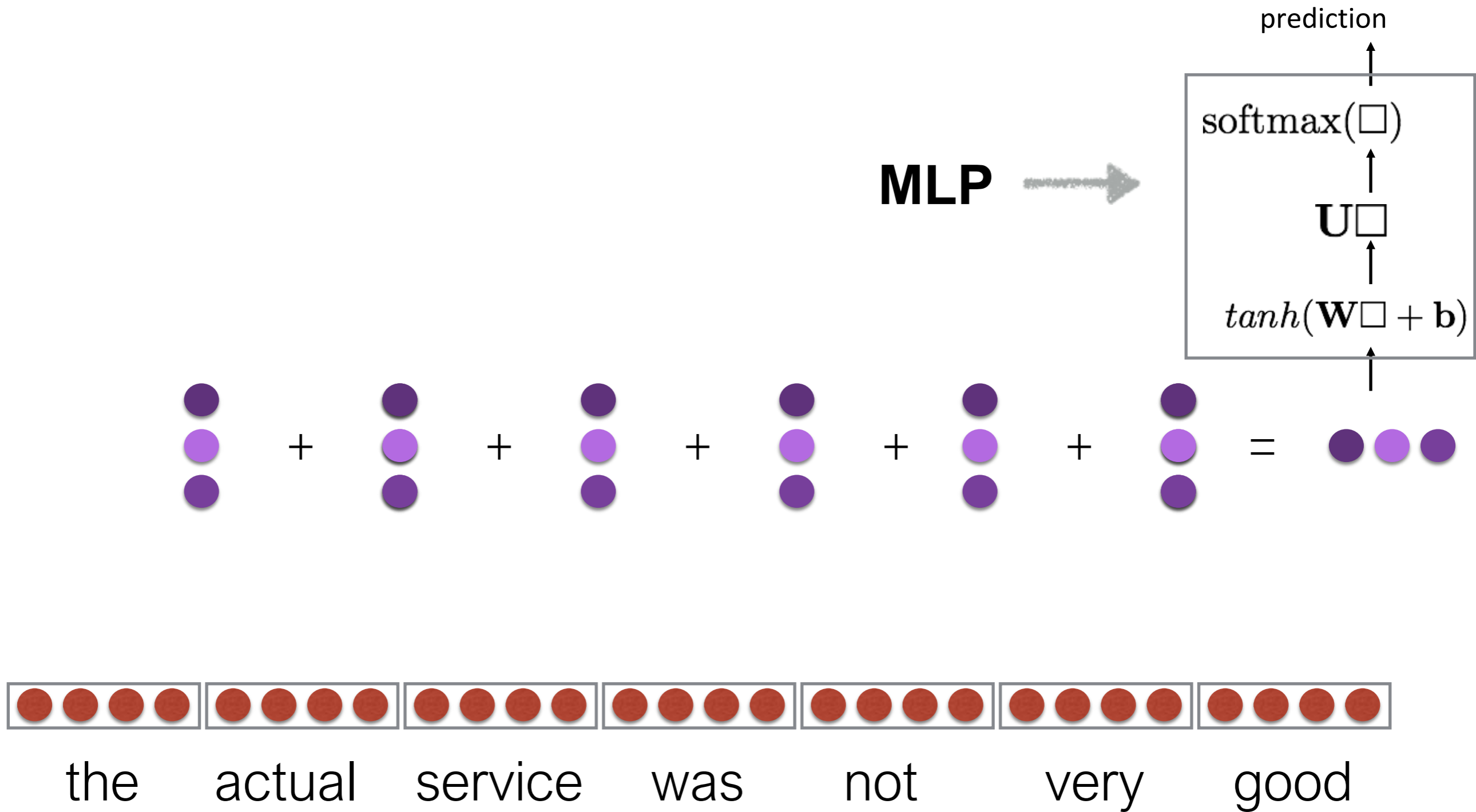


the actual service was not very good



train end-to-end for some task

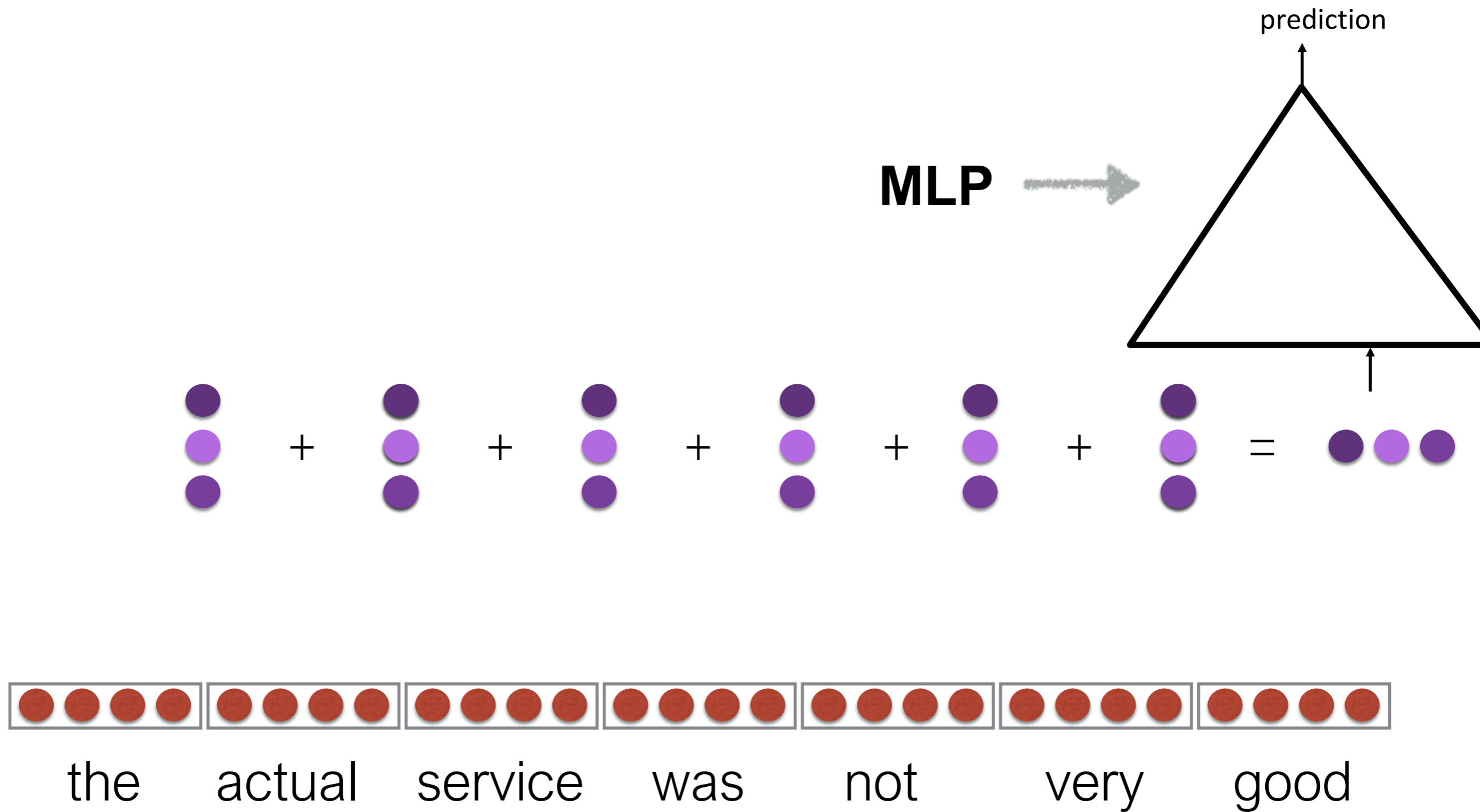
(train the MLP, the filter matrix, and the embeddings together)



train end-to-end for some task

(train the MLP, the filter matrix, and the embeddings together)

the vectors learn to capture what's important



train end-to-end for some task

(train the MLP, the filter matrix, and the embeddings together)
the vectors learn to capture what's important

we have the ngram vectors. now what?

Can look at the differences between terms.

microsoft <i>office</i> software		car <i>body</i> shop	
Free <i>office</i> 2000	0.550	car <i>body</i> kits	0.698
download <i>office</i> excel	0.541	auto <i>body</i> repair	0.578
word <i>office</i> online	0.502	auto <i>body</i> parts	0.555
apartment <i>office</i> hours	0.331	wave <i>body</i> language	0.301
massachusetts <i>office</i> location	0.293	calculate <i>body</i> fat	0.220
international <i>office</i> berkeley	0.274	forcefield <i>body</i> armour	0.165

Table 2: Sample word n-grams and the cosine similarities between the learned word-n-gram feature vectors of “*office*” and “*body*” in different contexts after the CLSM is trained.

A Latent Semantic Model with Convolutional-Pooling Structure for Information Retrieval

Yelong Shen
Microsoft Research
Redmond, WA, USA
yeshen@microsoft.com

Xiaodong He
Microsoft Research
Redmond, WA, USA
xiahe@microsoft.com

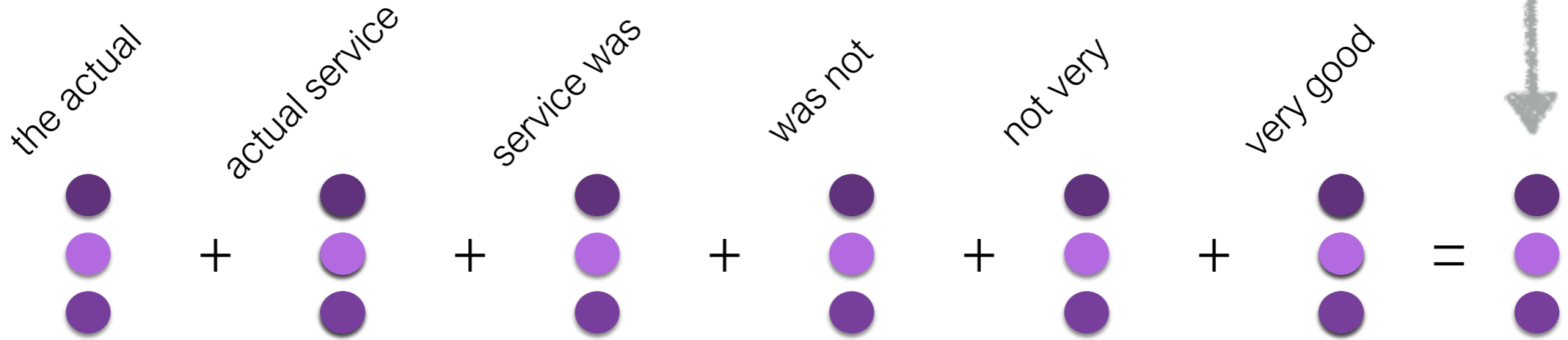
Jianfeng Gao
Microsoft Research
Redmond, WA, USA
jfgao@microsoft.com

Li Deng
Microsoft Research
Redmond, WA, USA
deng@microsoft.com

Grégoire Mesnil
University of Montréal
Montréal, Canada
gregoire.mesnil@umontreal.ca

average pooling

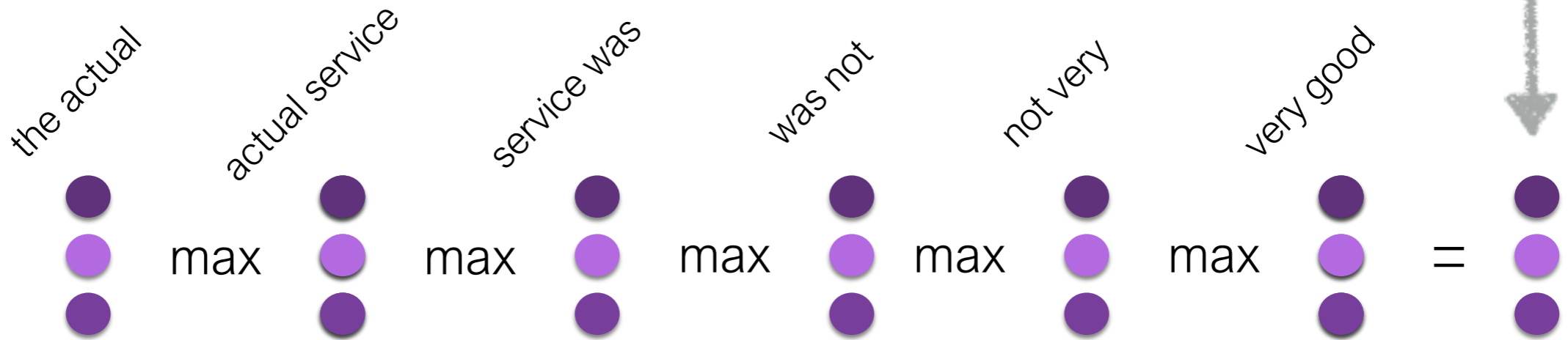
average vector



the actual service was not very good

max pooling

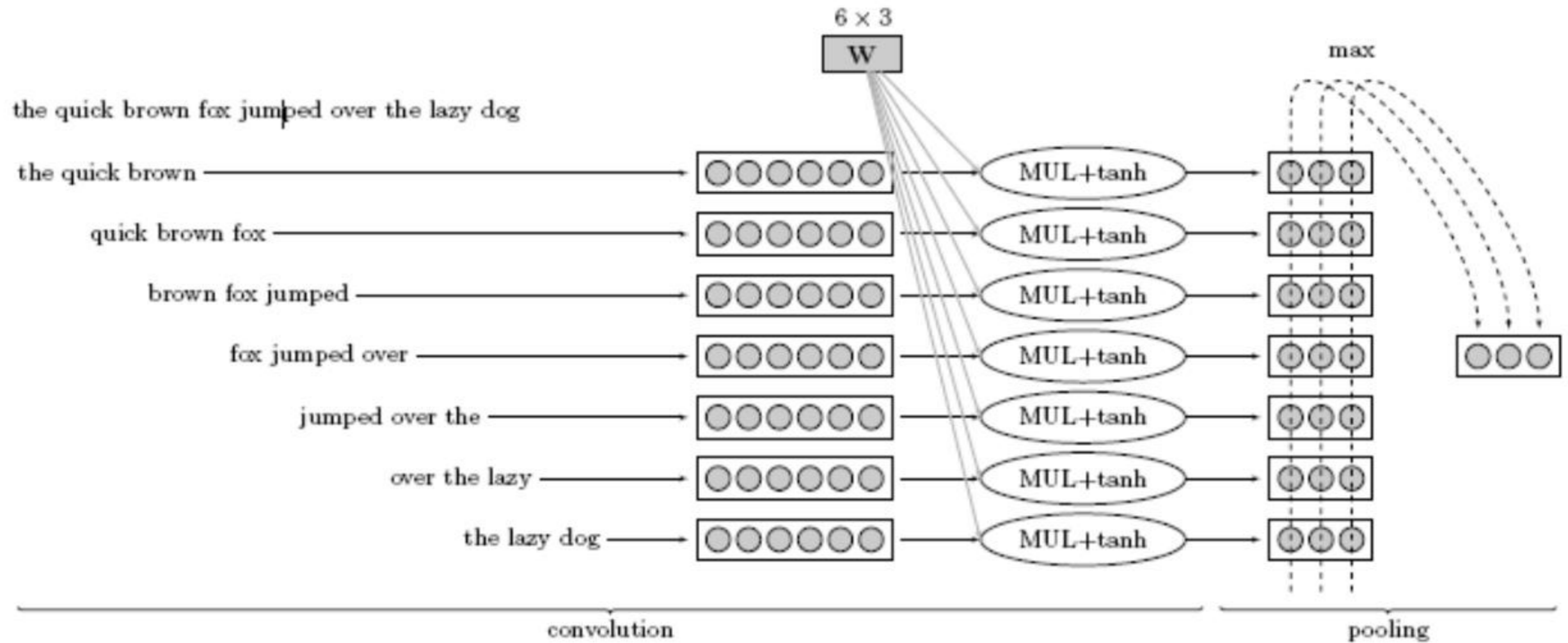
max vector



the actual service was not very good

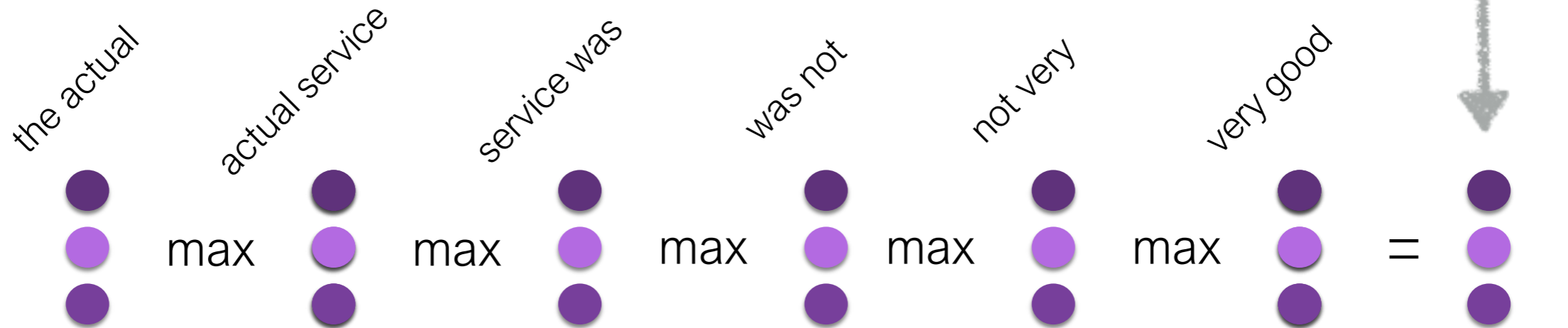
(max in each coordinate)

Another way to draw this:



max pooling

max vector



the actual service was not very good

max vs average – discuss

Zhang, Y., & Wallace, B. (2015). A Sensitivity Analysis of (and Practitioners' Guide to) Convolutional Neural Networks for Sentence Classification

Aspect	Max Pooling	Average Pooling	Sum Pooling
Captures	Strongest cue	Overall presence	Total evidence
Sensitivity to length	None	Mild	High
Robust to noise	✗	✓	✗
Trigger detection	✓ Excellent	✗ Weak	⚠ Mixed
Frequency awareness	✗	⚠ Partial	✓
Typical NLP usage	Very common	Occasional	Rare

Max pooling: *Did it appear at all?*

Sum pooling: *How much did it appear (total evidence)?*

Average pooling: *How prevalent was it?*

one benefit of max-pooling: it's "**interpretable**"

we can know where each element
in the summary vector came from

Examples of resulting "summaries"

microsoft **office excel** could allow remote **code execution**

welcome to the **apartment office**

online **body fat** percentage **calculator**

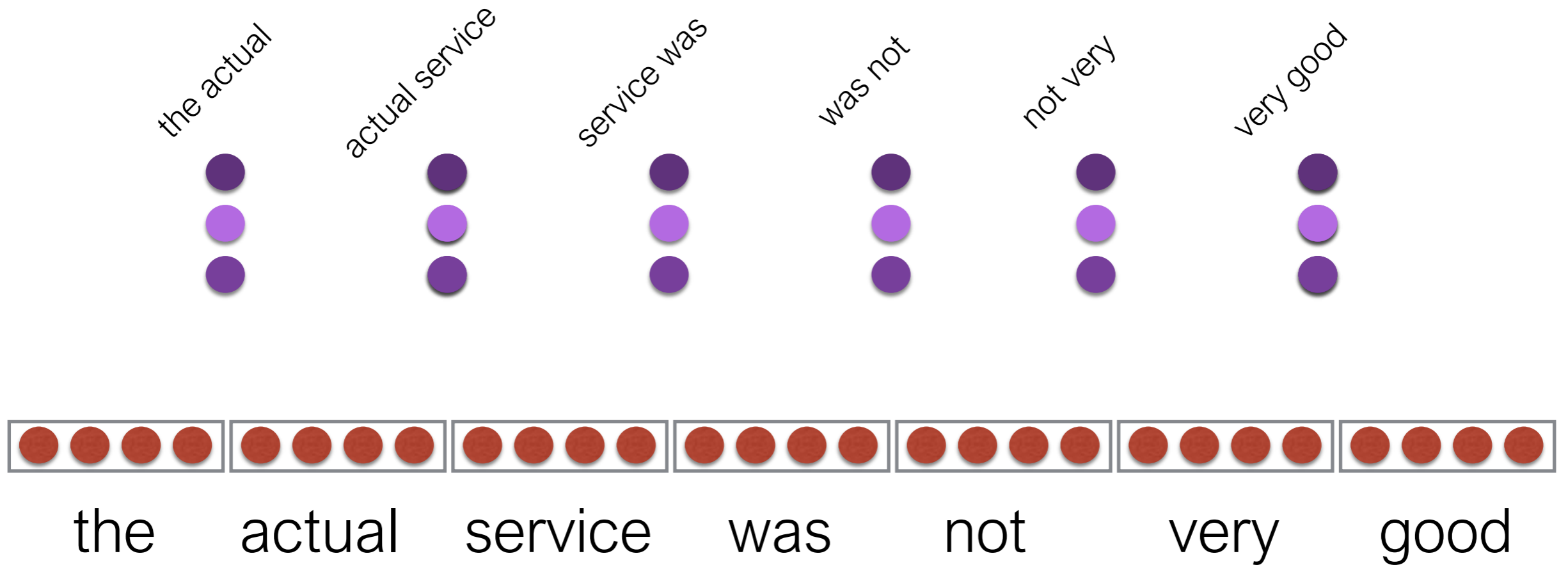
online **auto body** repair **estimates**

vitamin a the **health** benefits given by **carrots**

calcium supplements and **vitamin d** discussion stop **sarcoidosis**

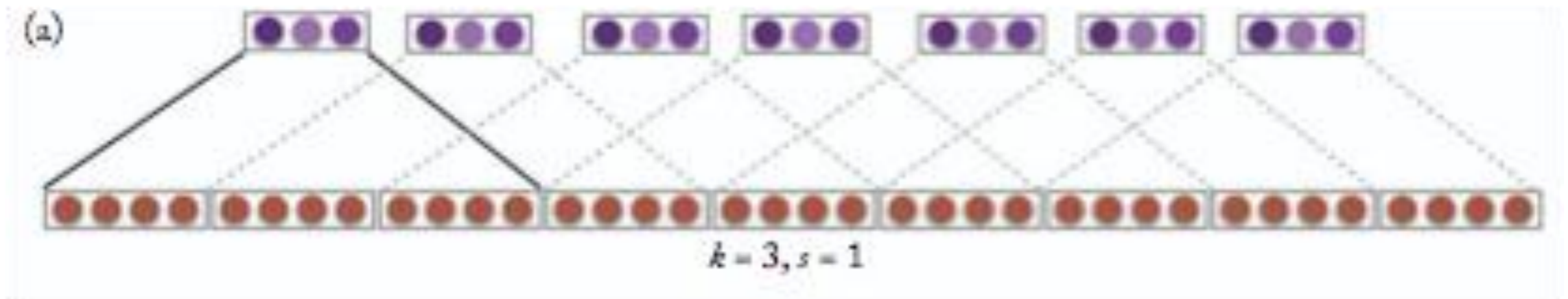
Table 3: Sample document titles. We examine the five most active neurons at the max-pooling layer and highlight the words in **bold** who win at these five neurons in the *max* operation. Note that, the feature of a word is extracted from that word together with the context words around it, but only the center word is highlighted in bold.

Strides



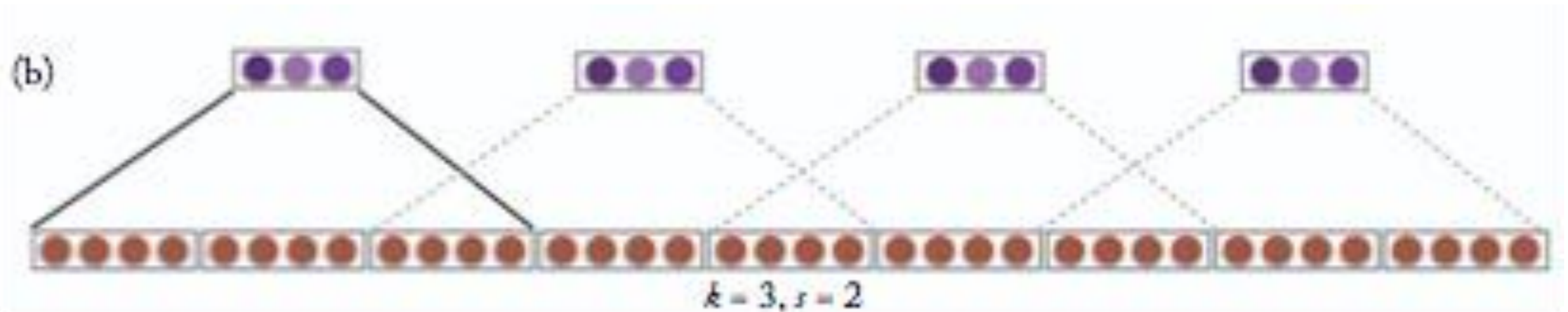
strides = how much you move

Strides



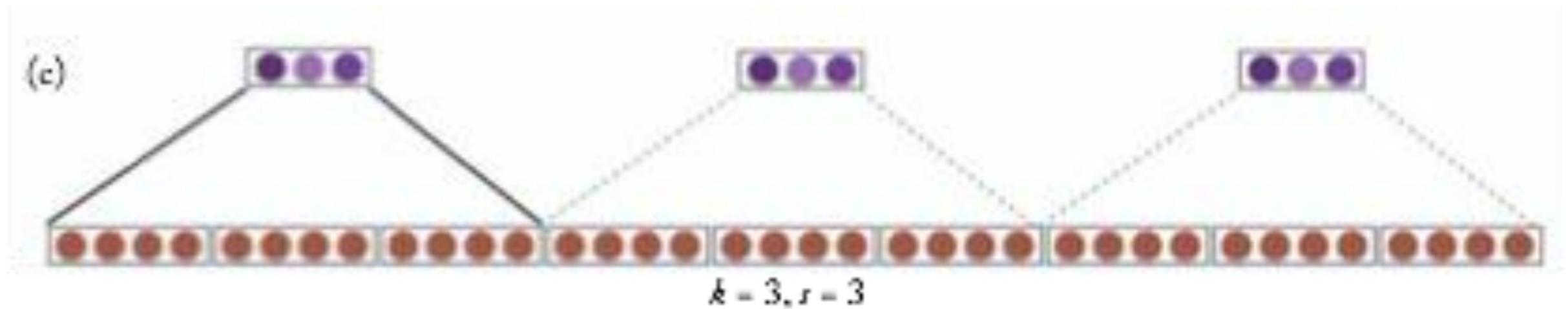
$k = 3, \text{ stride} = 1$

Strides



$k = 3, \text{ stride} = 2$

Strides



$k = 3, \text{ stride} = 3$

Hierarchy

Hierarchy



the actual



actual service



service was



was not



not very



very good



the

actual

service

was

not

very

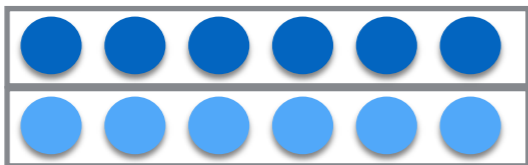
good

can have hierarchy

the actual service



||



dot



the actual



actual service



service was



was not



not very



very good



the

actual

service

was

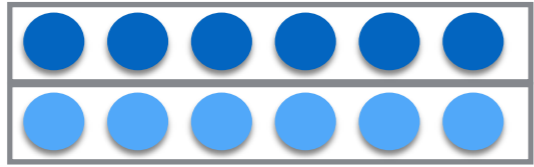
not

very

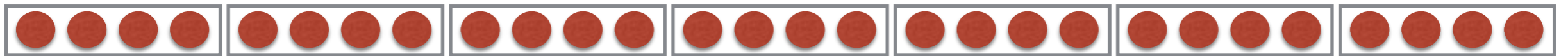
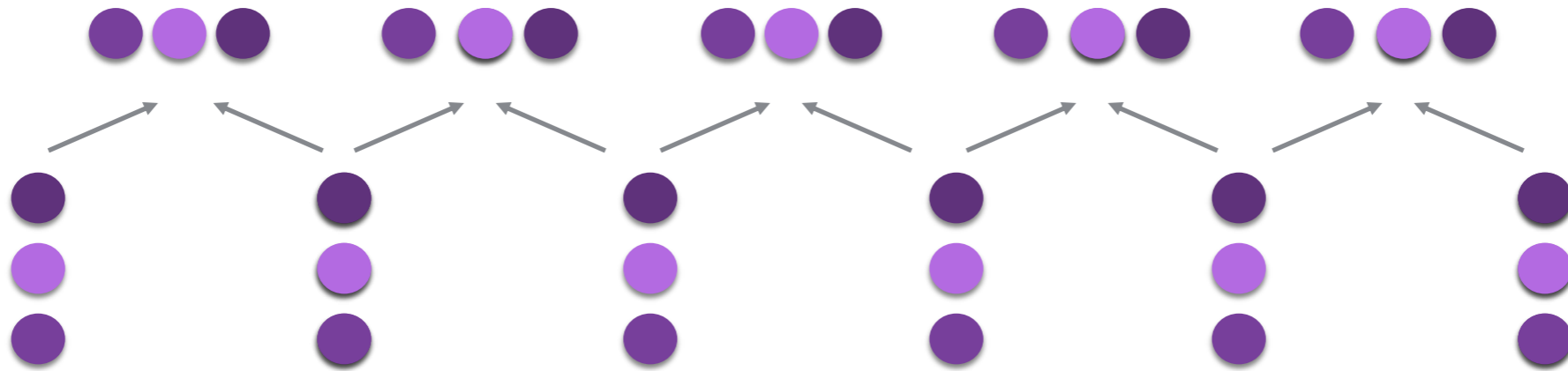
good

can have hierarchy

II



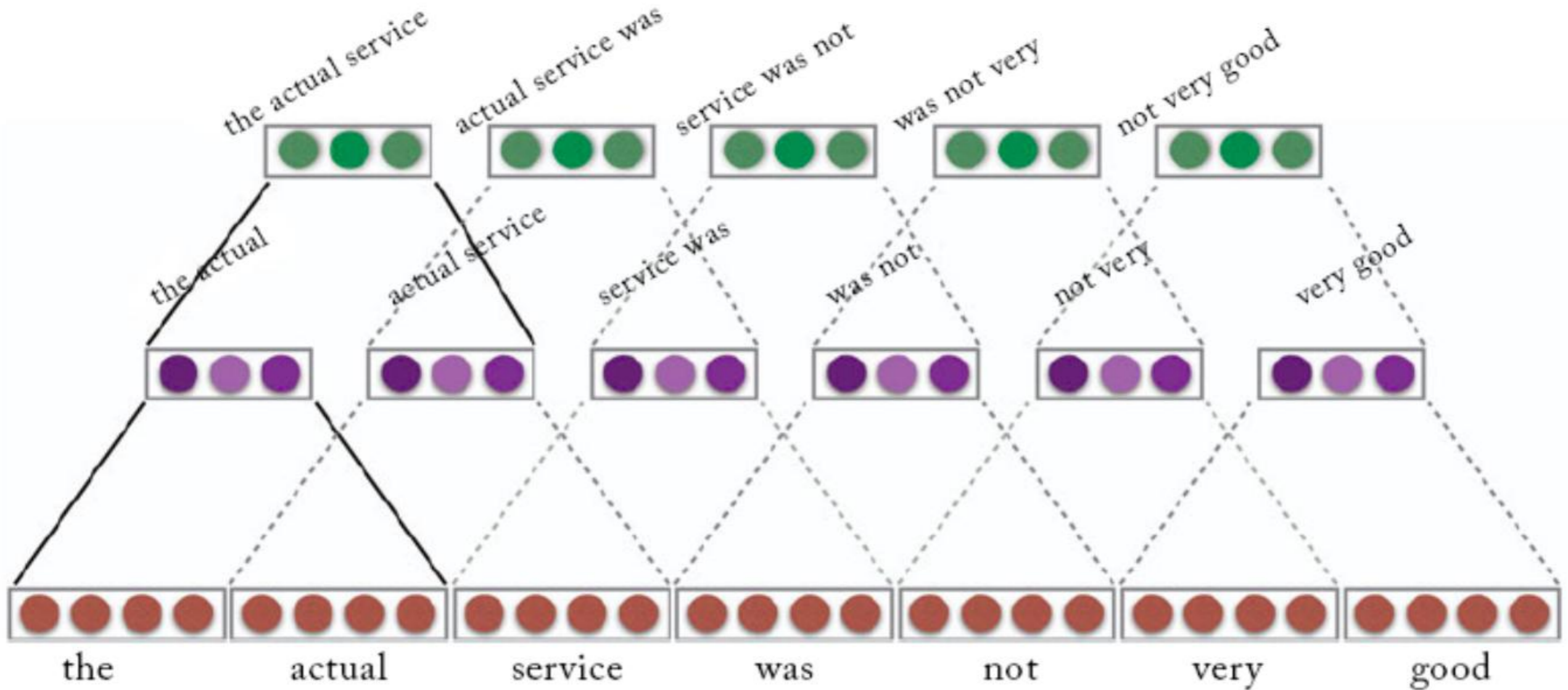
dot



the actual service was not very good

(can combine: **pooling + hierarchy**)

Hierarchy



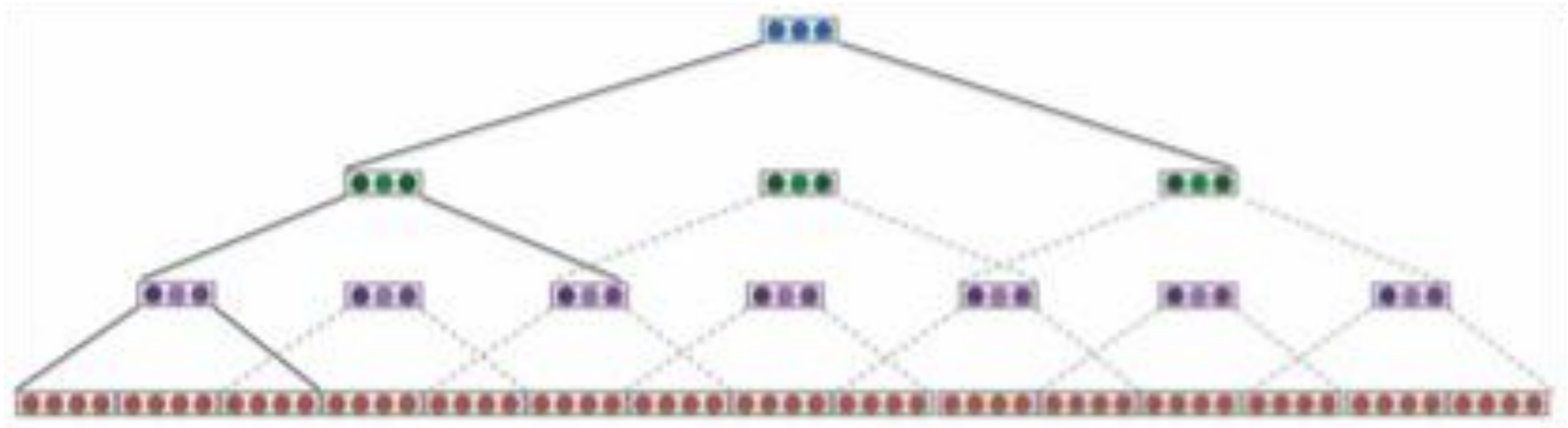
2-layer hierarchical conv with $k=2$

Dilated Convolutions

we want to cover more of the sequence

idea: strides + hierarchy

Dilated Convolutions



dilated convolution, $k=3$

idea: strides + hierarchy

ConvNets Summary

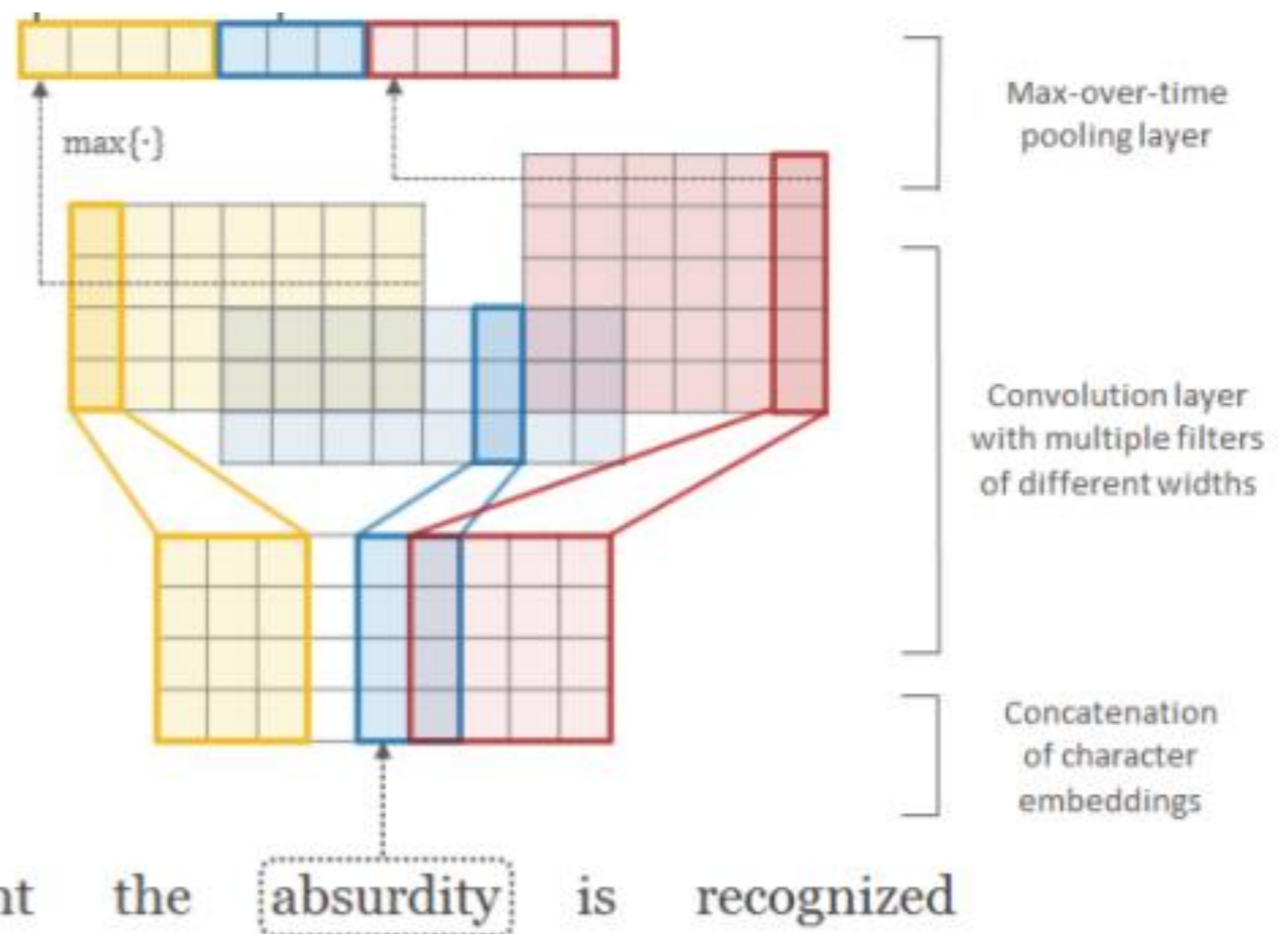
- Shared matrix used as feature detector.
- Extracts interesting ngrams.
- Pool ngrams to get fixed length representation.
- Max-pooling works well.
 - Max vs. Average pooling.
- Use hierarchy / dilation to expand coverage.
- Train end-to-end.

Character CNNs

- Fix the input OOV problem
 - Input: some insight in word shapes (xxxxing, xxxxly)
 - Output: can't ever output a word not in vocabulary
- Idea
 - Instead (or in addition of) word embedding
 - Use word = CNN over character sequences

Char CNN for Words

- Varied filter sizes
- Word embedding



Character-Aware Neural Language Models

Yoon Kim
School of Engineering
and Applied Sciences
Harvard University
yoonkim@seas.harvard.edu

Yacine Jernite
Courant Institute
of Mathematical Sciences
New York University
jernite@cs.nyu.edu

David Sontag
Courant Institute
of Mathematical Sciences
New York University
dsontag@cs.nyu.edu

Alexander M. Rush
School of Engineering
and Applied Sciences
Harvard University
srush@seas.harvard.edu

- Can't differentiate between words w similar spellings
- Solution: add small correction $[e_w = \text{CNN}(\text{chars}_w) + M.\text{corr}_w]$

Alternative: Hashing Trick

- ConvNet is an architecture for finding good ngrams.
- But if we know ngrams are important, why not just have ngram embeddings (ngram vectors)?
- --> for large vocabulary, not scalable.

Can't represent all ngrams, don't know which are important.

Alternative: Hashing Trick

- **Problem:** our ngram vocabulary size is 10^9
- **Solution:** use smaller space via hashing, allow feature clashes.

Hashing Trick

- We have $> 10^9$ different ngrams.
- We can afford $\sim 10^6$ different embeddings.
- Map each ngram to a number in $[0, 10^6]$
- Use the corresponding embedding vector.
- Clashes will happen, but it will probably be ok.
- Even safer: map each ngram to two numbers using two different hash functions, sum the vectors.

Hashing Trick vs ConvNets

- What are the benefits of using bag of ngrams?
- What are the benefits of using ConvNet (ngram detector)?
- Does it matter if the vocabulary size is small or large?

(discuss)

Problems:

What makes reviews hard to classify?

- Subtlety:
 - Perfume review in *Perfumes: the Guide*:
 - “If you are reading this because it is your darling fragrance, please wear it at home exclusively, and tape the windows shut.”
 - Dorothy Parker on Katherine Hepburn
 - “She runs the gamut of emotions from A to B”

CHALLENGES

- Ambiguous words
 - This music cd is literal waste of time. (negative)
 - Please throw your waste material here. (neutral)
- Sarcasm detection and handling
 - “All the features you want - too bad they don’t work. :-P”
- (Almost) No resources and tools for low/scarce resource languages like Indian languages.

User written: grammar, spellings...

Hi,




I have Haier phone.. It was good when i was buing this phone.. But I invented A lot of bad features by this phone those are It's cost is low but Software is not good and Battery is very bad... Ther are no signals at out side of the city...,, People can't understand this type of software...,, There aren't features in this phone, Design is better not good...,, Sound also bad.. So I'm not intrest this side They are diving heare phones it is good. They are these are also good.They are giv also good because other phones low wait.




**Lack of punctuation marks,
Grammatical errors**

Wait.. err.. Come again

From: www.mouthshut.com

Alternating Sentiment

I suggest that instead of fillings songs in tunes you should
 tunes (not made of songs) only. The phone  has good
popularity in old age people. Third  i had tried much for its
data cable but i find it nowhere. It should be supplied with
set with some extra cost.

Good  features of this phone are its cheapest price and
durability . It should  ve some features more than nokia
1200. it is easily available in market and  pair is also
available

Subject Centrality

- I have this personal experience of using this cell phone. I bought it one and half years back. It had modern features that a normal cell phone has, and the look is excellent. I was very impressed by the design. I bought it for Rs. 8000. It was a gift for someone. It worked fine for first one month, and then started the series of multiple faults it has. First the speaker didnt work, I took it to the service centre (which is like a govt. office with no work). It took 15 days to repair the handset, moreover they ~~charged me Rs. 500. Then after 15~~ days again the mike didnt work, then again same set of time was consumed for the repairs and it continued. Later the camera didnt work, the speakes were rubbish, it used to hang. It started restarting automatically. And the govt. office had staff which I doubt have any knoledge of cell phones??

These multiple faults continued for as long as one year, when the warranty period ended. In this period of time ~~I spent a considerable amount on the petrol, a lot of time (as~~ the service centre is a govt. office). And at last the phone is still working, but now it works as a paper weight. The company who produces such items must be sacked. I understand that it might be fault with one prticular handset, but the company itself never bothered for replacement and I have never seen such miserable cust service. For a comman man like me, Rs. 8000 is a big amount. And I spent almost the same amount to get it work, if any ~~has a good suggestion and can gude me how to sue such~~ companies, please guide.

For this the quality team is faulty, the cust service is really miserable and the worst condition of any organisation I have ever seen is ~~with the service centre for Fly and Sony Erricson, (it's near~~ Sancheti hospital, Pune). I dont have any thing else to say.

Thwarted Expectations and Ordering Effects

- “This film should be brilliant. It sounds like a great plot, the actors are first grade, and the supporting cast is good as well, and Stallone is attempting to deliver a good performance. However, it can’t hold up.”
- Well as usual Keanu Reeves is nothing special, but surprisingly, the very talented Laurence Fishbourne is not so good either, I was surprised.

Thwarted Expectations and Ordering Effects

- “This film should be **brilliant**. It sounds like a **great** plot, the actors are **first grade**, and the supporting cast is **good** as well, and Stallone is attempting to deliver a good performance. However, it **can't hold up**.”
- Well as usual Keanu Reeves is nothing special, but surprisingly, the **very talented** Laurence Fishbourne is **not so good** either, I was surprised.