Tendencies and Trends in NLP

Word Embedding, Representation Learning, Document Classification, …

Navid Rekabsaz
Idiap Research Institute

navid.rekabsaz@idiap.ch @navidrekabsaz

Natural Language Understanding

- A great challenge in AI
  - How can we think, understand, and communicate with each other through sequences of symbols?

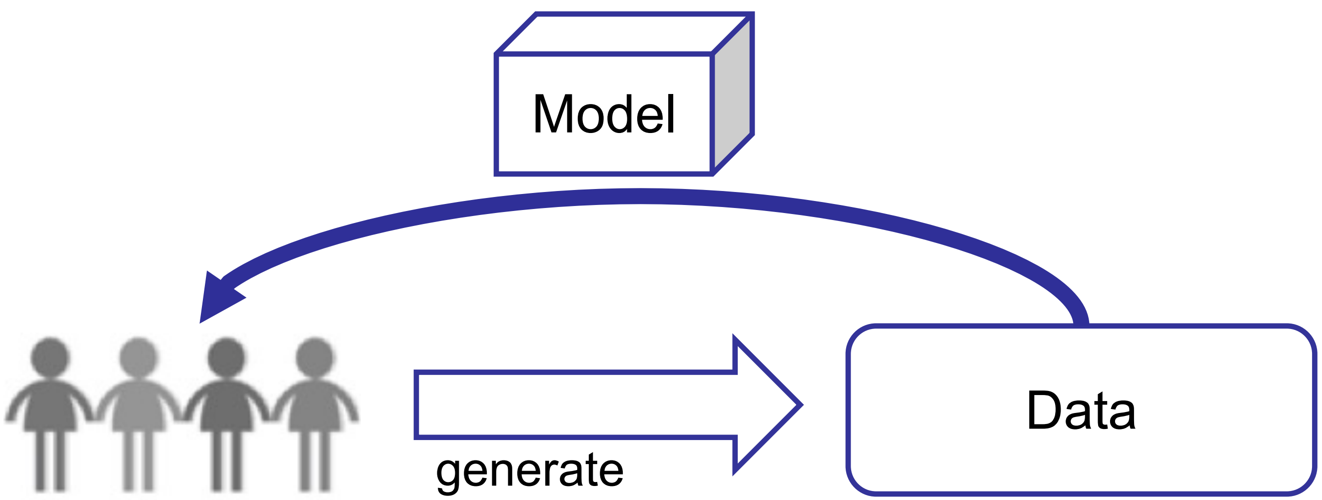
- What is the structure of human thoughts?

- How can we calculate a structured thought from an observed sequence?

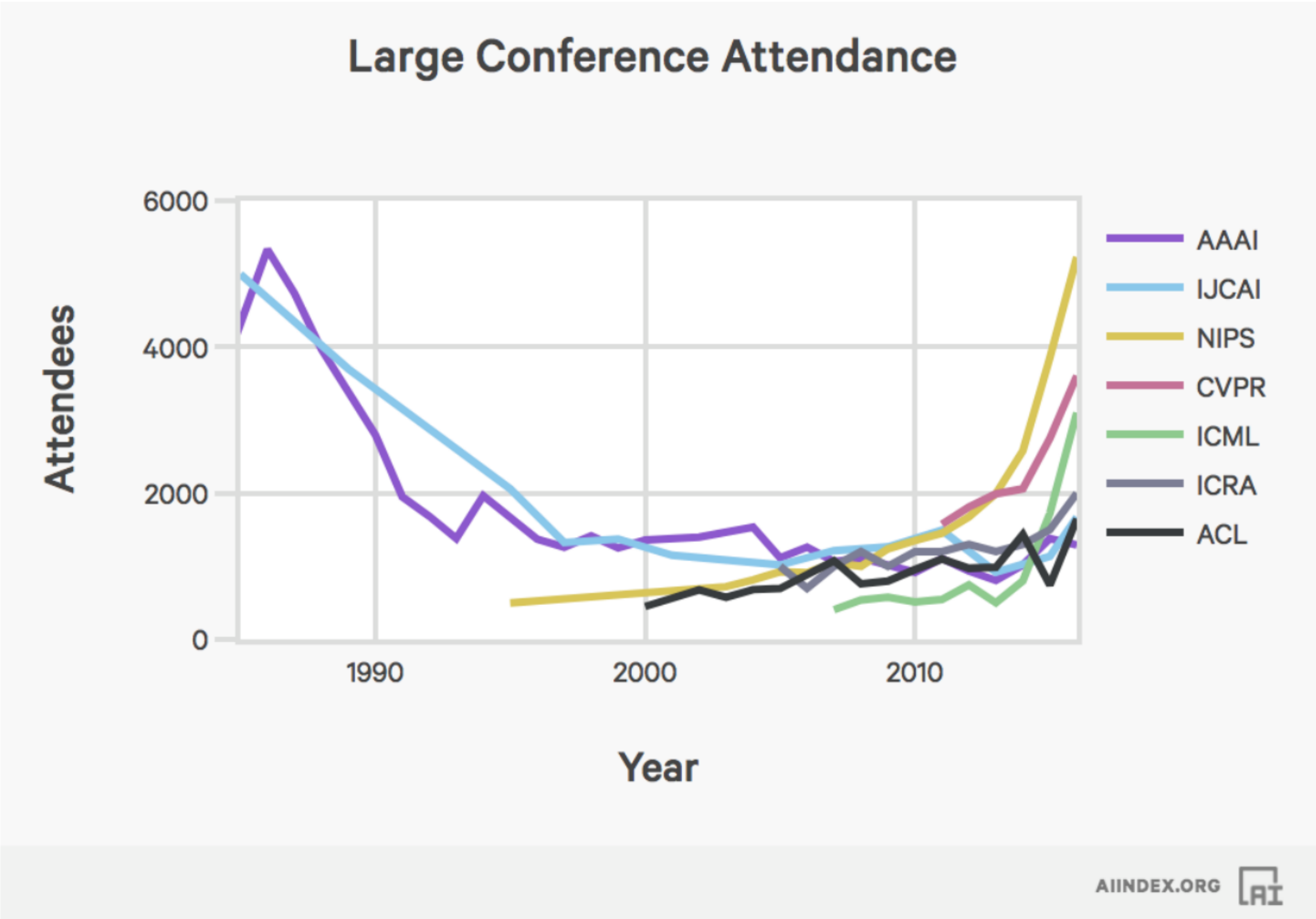
- How can calculate a sequence to convey a structured thought?

Natural Language Understanding

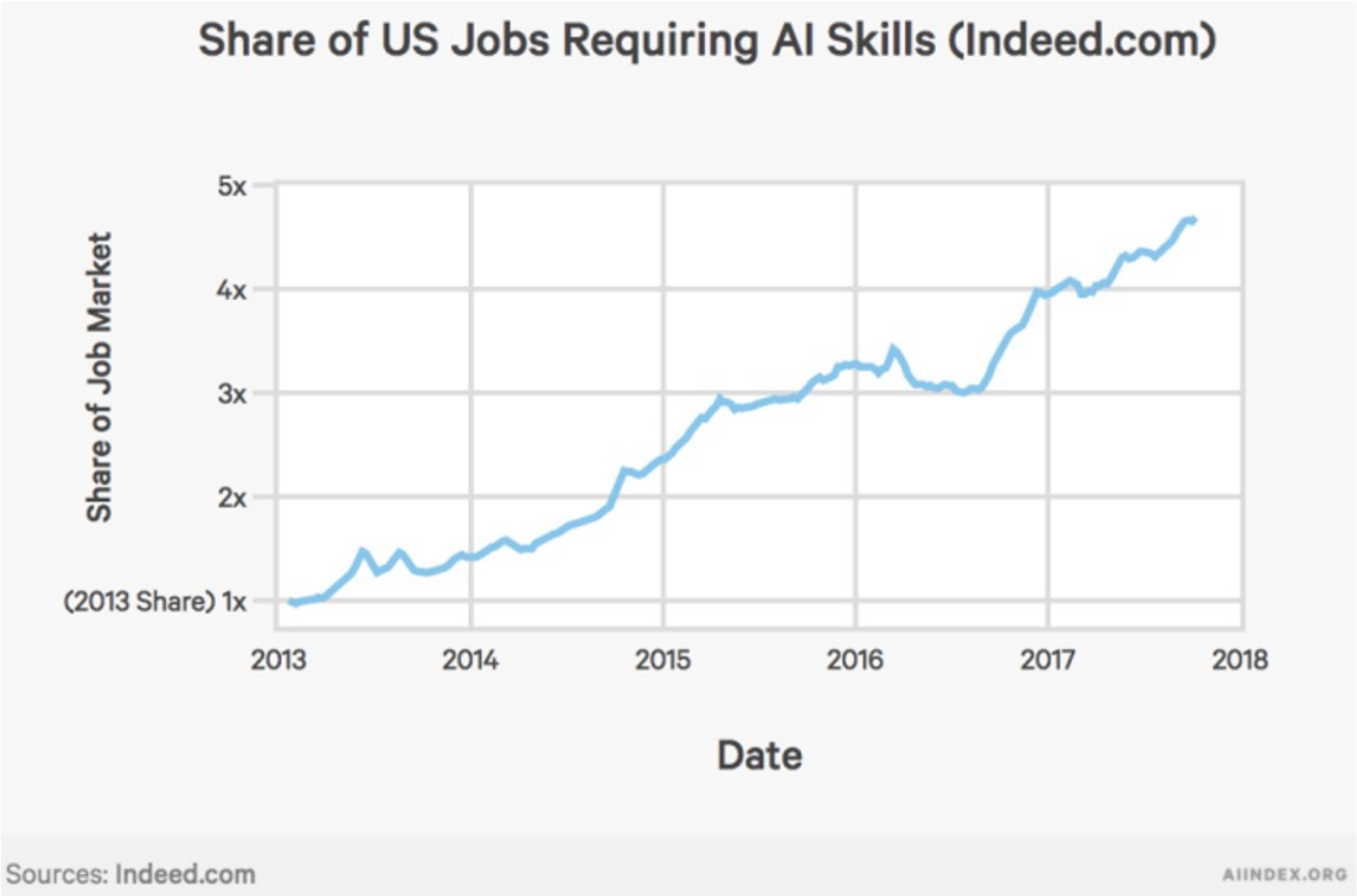
- We do not know how to directly conceptualize intelligence.
- Instead, we usually rely on **learning** from experience and **data**
- Modeling can lead to new insights on the origin



Research on AI and NLP



Job Market



Share of US Jobs Requiring AI Skills (Indeed.com)
5x
4x
3x
2x
(2013 Share) 1x
Share of Job Market
2013
2014
2015
2016
2017
2018
Date
Sources: Indeed.com
AIINDEX.ORG

Agenda

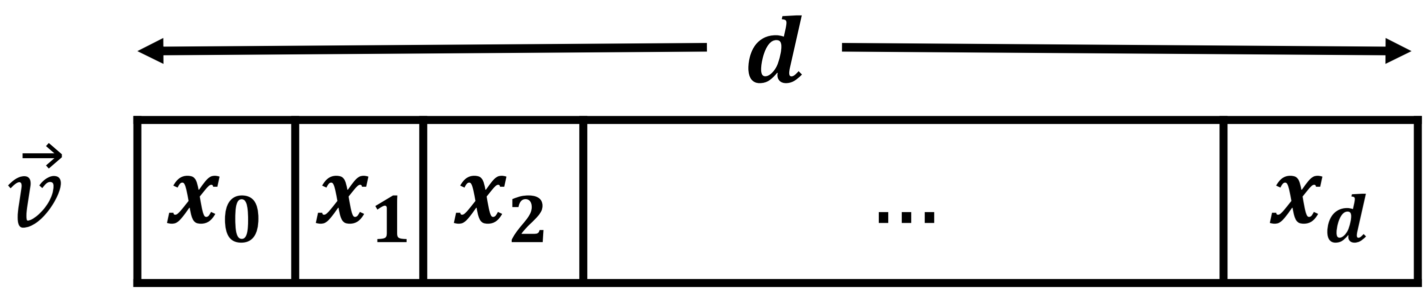
- Crash course (1)
  - Vector representation
  - Neural Networks
- Word Representation Learning
  - Neural word representation
  - word2vec

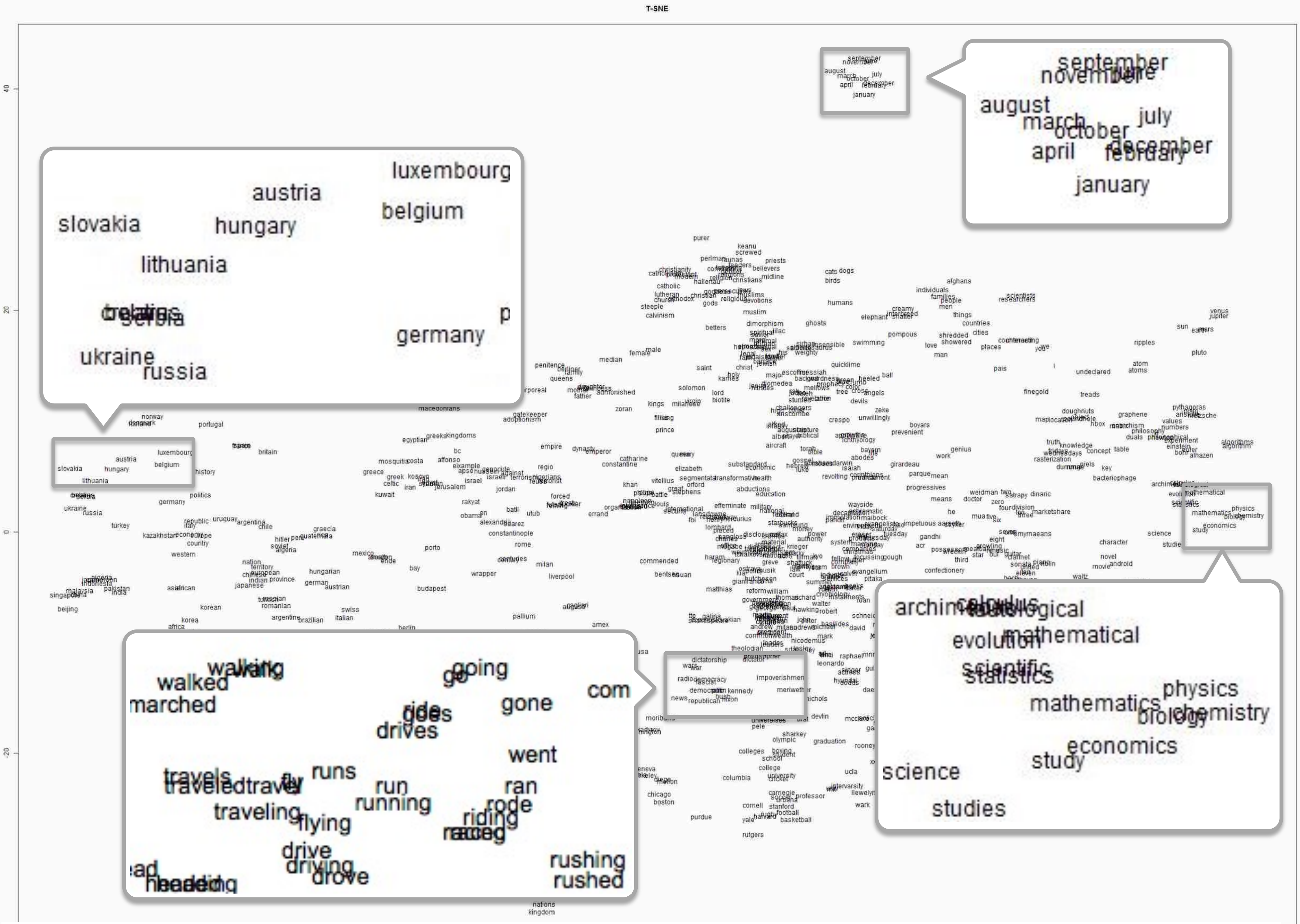
---Break---

- Crash course (2)
  - Recurrent Neural Networks
  - Attention Mechanism
- Document Classification
- Applications and challenges

Vector Representation

- Computation starts with representation of entities
- Common entities in NLP: word, sentence, document, etc.
- An entity is represented with a vector of *d* dimensions
- The dimensions usually reflects concepts, related to an entity





Vector representations of words projected in two-dimensional space

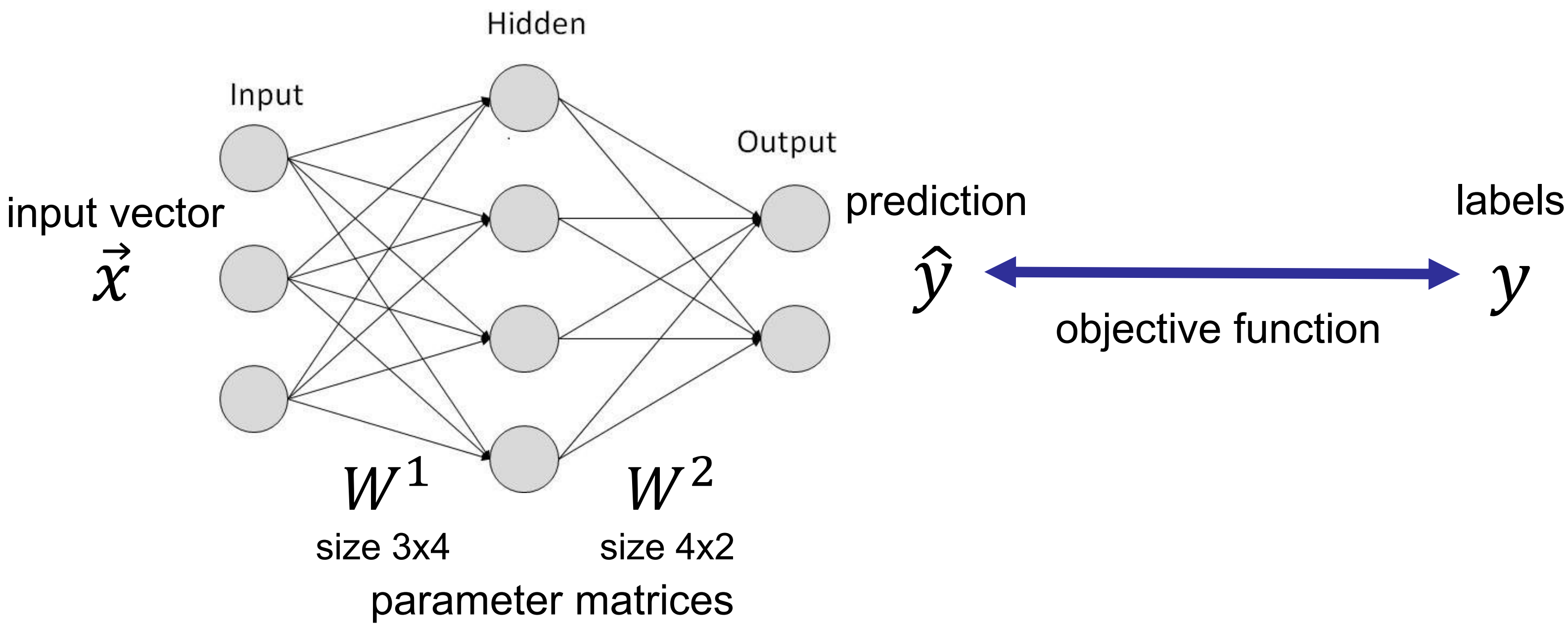
Vector Representation

- Vector representations (in this course) are usually:
  - in dimensions around $d \sim [10\text{-}1000]$
  - dense
  - referred to as embeddings: e.g. word embedding

- The similarity between the entities can be computed by any distance/similarity measure:
  - Usually by dot product or cosine (normalized dot product) between the vectors

$$similarity(\mathrm{cat}, \mathrm{sloth}) = cosine(\vec{v}_{\mathrm{cat}}, \vec{v}_{\mathrm{sloth}}) = \frac{\vec{v}_{\mathrm{cat}} \cdot \vec{v}_{\mathrm{sloth}}}{|\vec{v}_{\mathrm{cat}}||\vec{v}_{\mathrm{sloth}}|}$$

Neural Networks

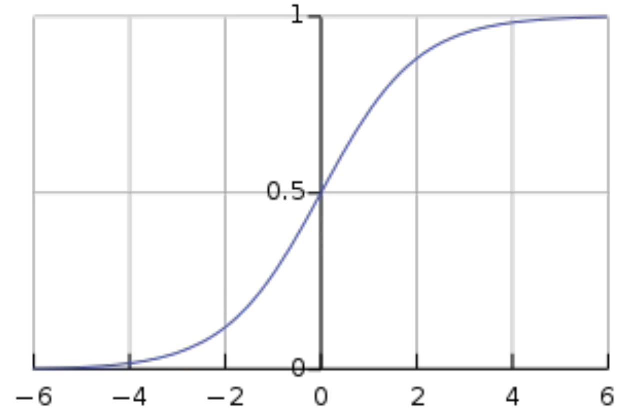


- Training Steps
  - Forward pass calculates output $\hat{y}$ from input $\vec{x}$
  - Objective function calculates error by comparing $\hat{y}$ with y
  - Backpropagation calculates the gradient of each parameter in regard to the error
  - Optimize updates network parameters using the gradients in the hope of reducing error (Stochastic Gradient Descent)

Neural Networks

- Non-linearities on hidden layer
  - Sigmoid σ
  - Tanh
  - Relu

Sigmoid function

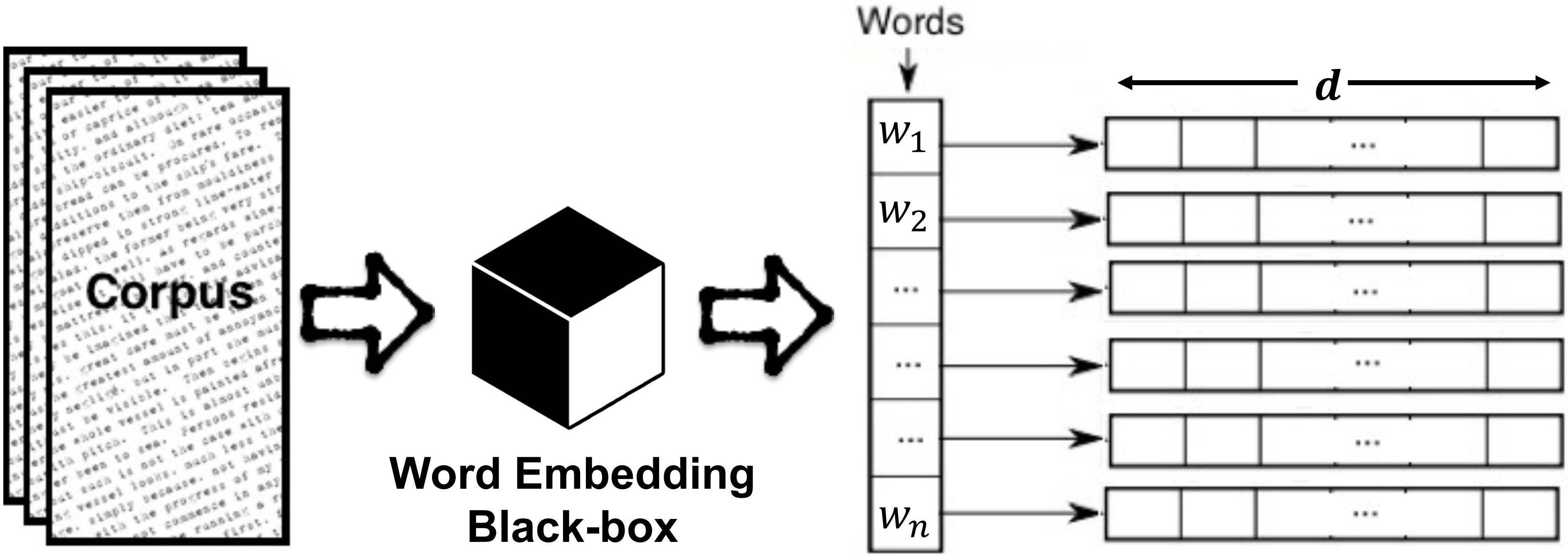


- Softmax on output layer
  - normalizes values of a vector to the range of (0,1) such that all the values sums up to 1 $\Longrightarrow$ probability distribution

$$softmax(\vec{v})_i = \frac{e^{v_i}}{\sum_{k=1}^{d} e^{v_k}}$$

$$\vec{v} = [2, 3, 5, 6] \qquad softmax(\vec{v}) = [0.01, 0.03, 0.26, 0.70]$$

Word Representation Learning



Intuition for Computational Semantics

"You shall know a word by the company it keeps!"

J. R. Firth, A synopsis of linguistic theory 1930–1955 (1957)

Tesgüino

drink

drunk

alcohol

on the table

make

out of corn

fermented

Mexico

bottle of

Nida[1975]

Dutch

drunk

pale

Heineken

brew

red star

bar

drink

green bottle

alcohol

Tesgüino $\longleftrightarrow$ Heineken

Algorithmic intuition:

Two words are related when
they have similar context words

Word-Context Matrix

- Number of times a word *c* appears in the context of the word *w* in a corpus

sugar, a sliced lemon, a tablespoonful of	**apricot**	preserve or jam, a pinch each of,
their enjoyment. Cautiously she sampled her first	**pineapple**	and another fruit whose taste she likened
well suited to programming on the digital	**computer**.	In finding the optimal R-stage policy from
for the purpose of gathering data and	**information**	necessary for the study authorized in the

	c_1 Aardvark	c_2 computer	c_3 data	c_4 pinch	c_5 result	c_6 sugar
w_1 apricot	0	0	0	1	0	1
w_2 pineapple	0	0	0	1	0	1
w_3 digital	0	2	1	0	1	0
w_4 information	0	1	6	0	4	0

- Our first word vector representation!!

[1]

Words Semantic Relations

	c_1 Aardvark	c_2 computer	c_3 data	c_4 pinch	c_5 result	c_6 sugar
w_1 apricot	0	0	0	1	0	1
w_2 pineapple	0	0	0	1	0	1
w_3 digital	0	2	1	0	1	0
w_4 information	0	1	6	0	4	0

- **Co-occurrence relation**
  - Words that appear near each other in the language
  - Like (*drink* and *beer*) or (*drink* and *wine*)
  - Measured by counting the co-occurrences
- **Similarity relation**
  - Words that appear in similar contexts
  - Like (*beer* and *wine*) or (*knowledge* and *wisdom*)
  - Measured by similarity metrics between the vectors

$$similarty(\text{digital}, \text{information}) = \text{cosine}\left(\vec{v}_{\text{digital}}, \vec{v}_{\text{information}}\right)$$

Sparse vs. Dense Vectors

- Such word representations are highly sparse
  - Number of dimensions is the same as the number of words in the corpus $n \sim$ [10000–500000]
  - Many zeros in the matrix as many words don't co-occur
    - Normally ~98% sparsity

- Dense representations → Embeddings
  - Number of dimensions usually between $d \sim$ [10–1000]

- Why dense vectors?
  - More efficient for storing and load
  - More suitable for machine learning algorithms as features
  - Generalize better by removing noise for unseen data

Word Embedding with Neural Networks

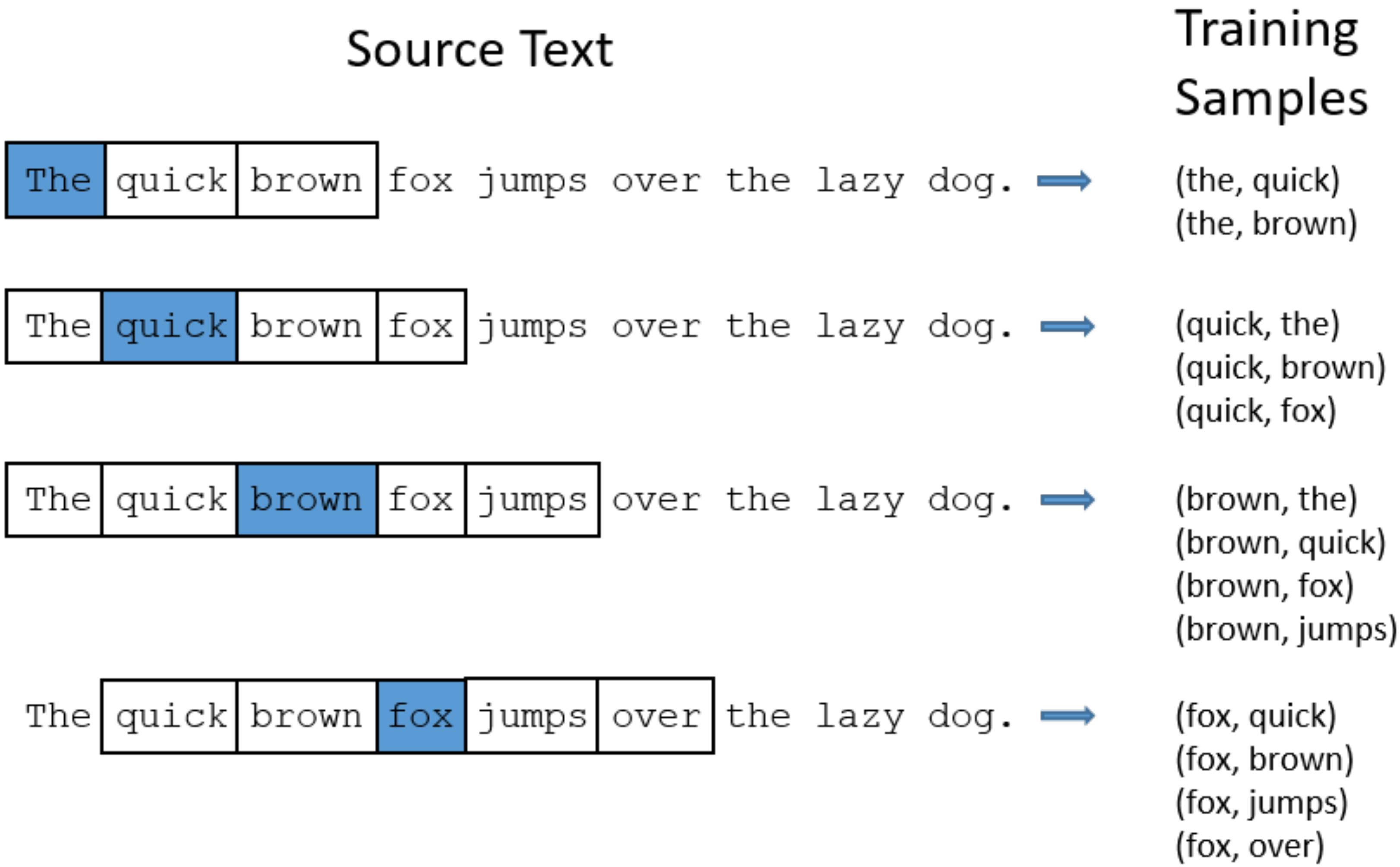
<u>Recipe for creating (dense) word embedding with neural networks</u>

1. Design a neural network architecture!
2. Loop over training data (w, c)
   a. Set word w as input and context word c as output
   b. Calculate the output of network, namely
      The probability of observing context word c given word w
      $$P(c|w)$$
   c. Optimize the network to increase this probability
3. Repeat

<u>Details come next!</u>

Prepare Training Samples

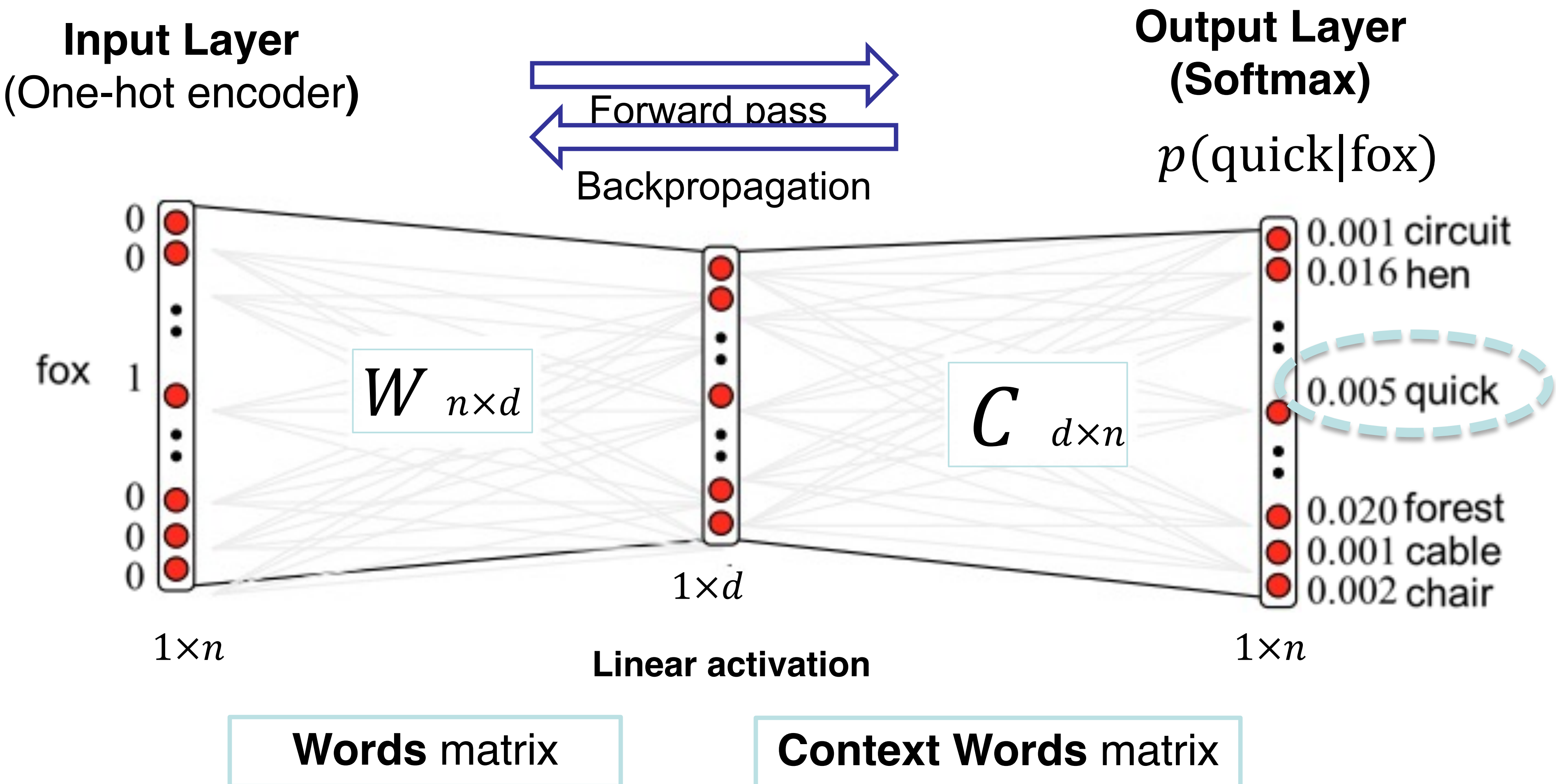
Window size of 2



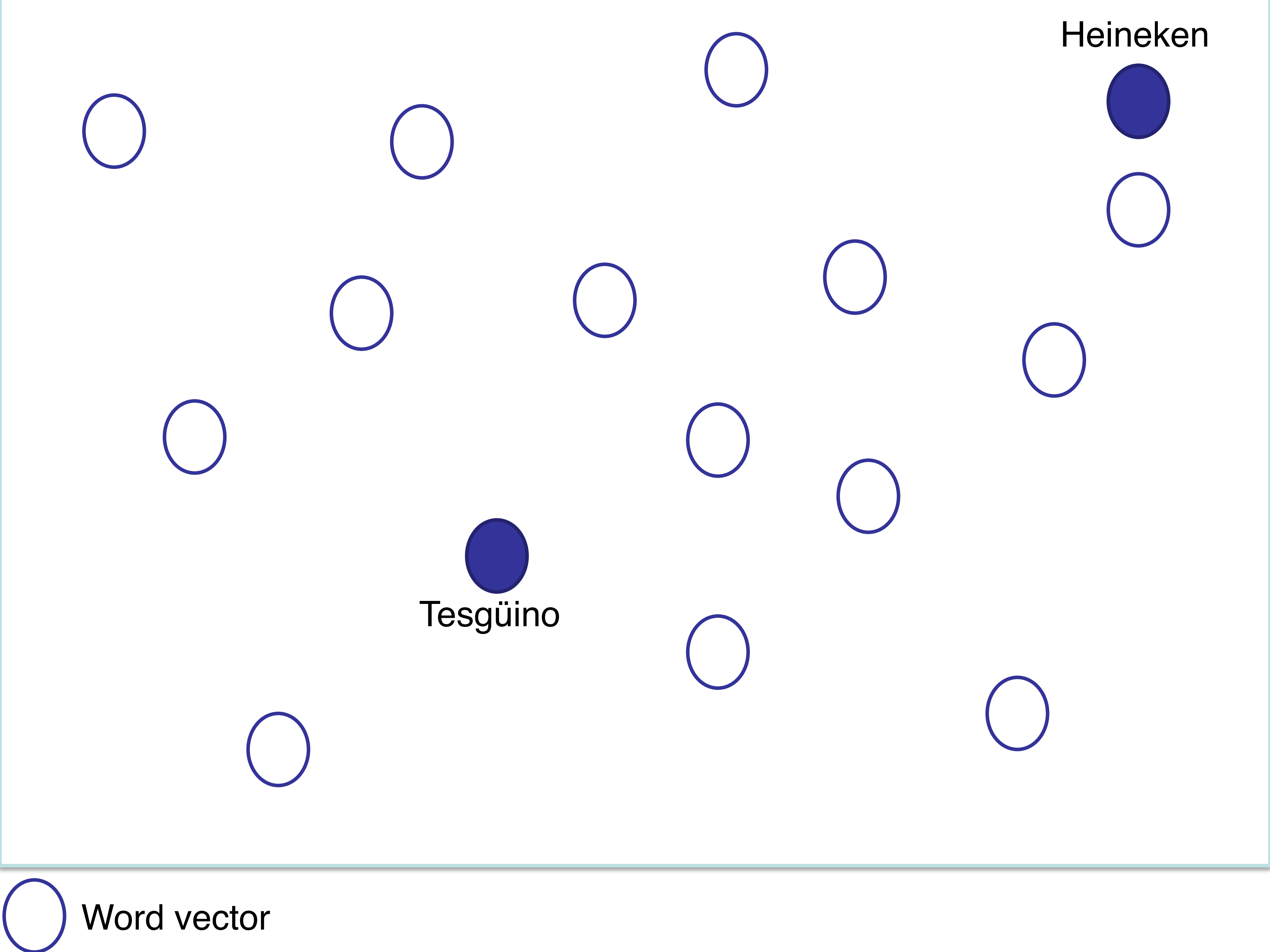
http://mccormickml.com/2016/04/19/word2vec-tutorial-the-skip-gram-model/

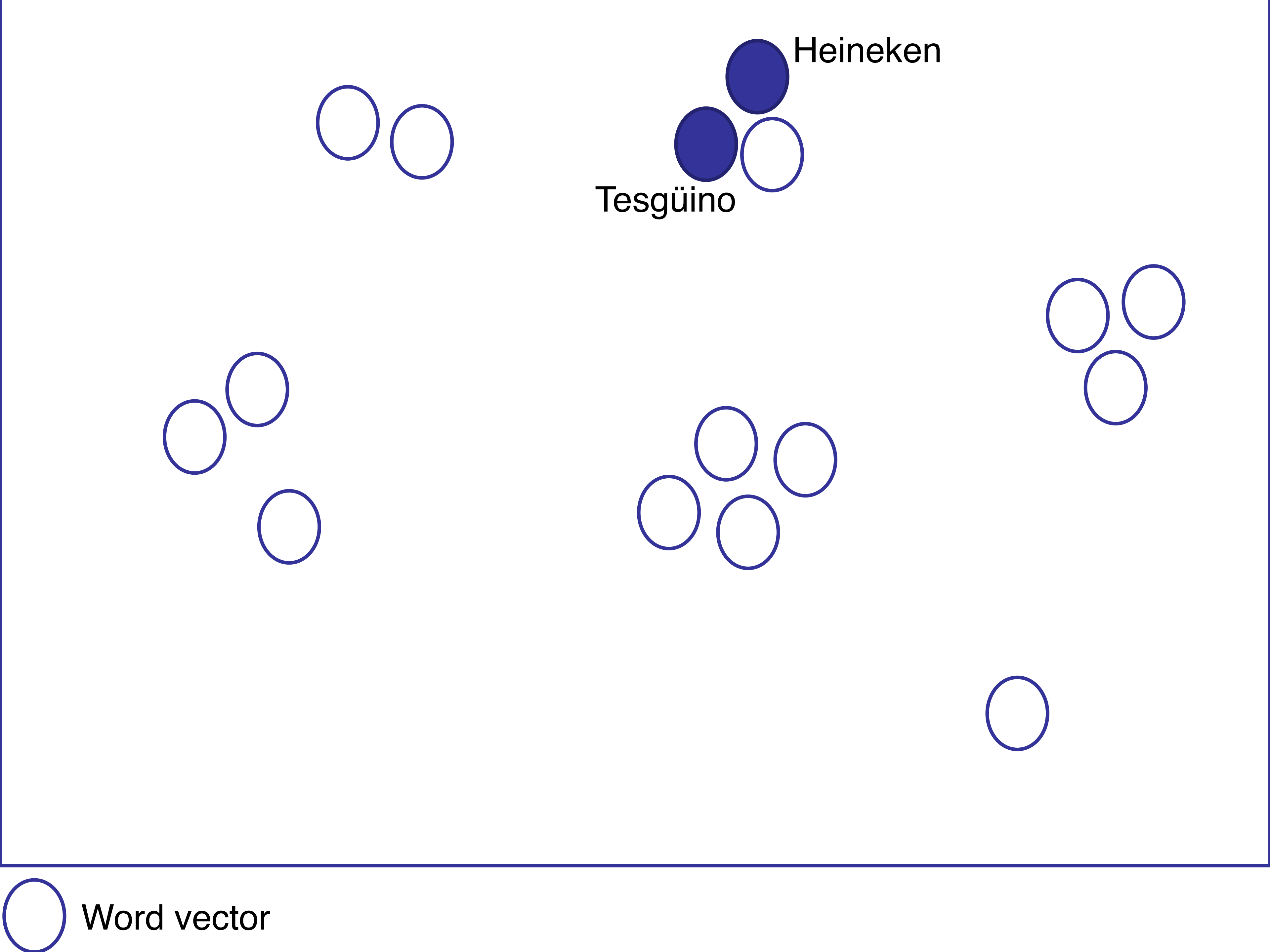
Neural Word Embedding Architecture

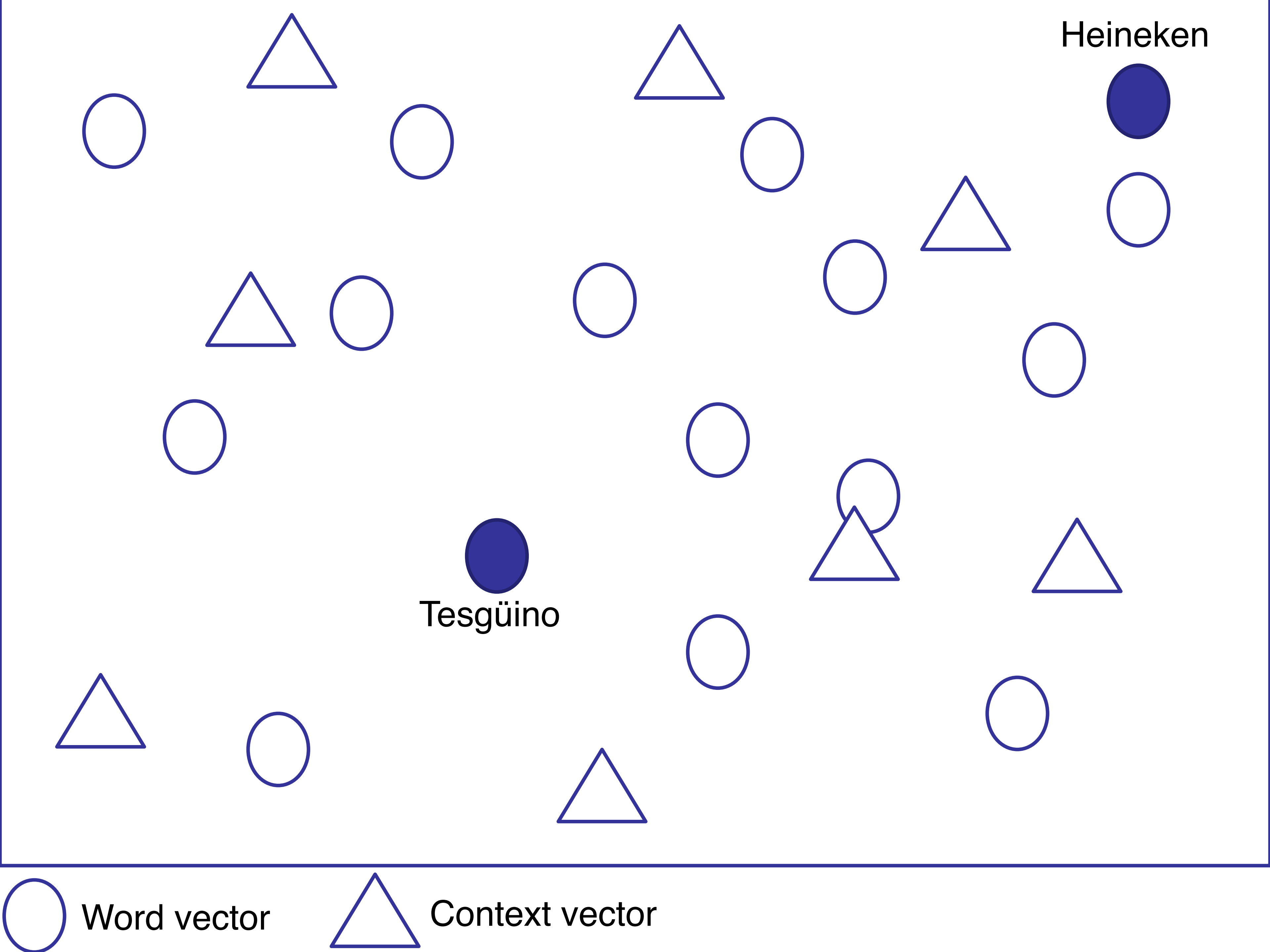
Train sample: (*fox*, *quick*)

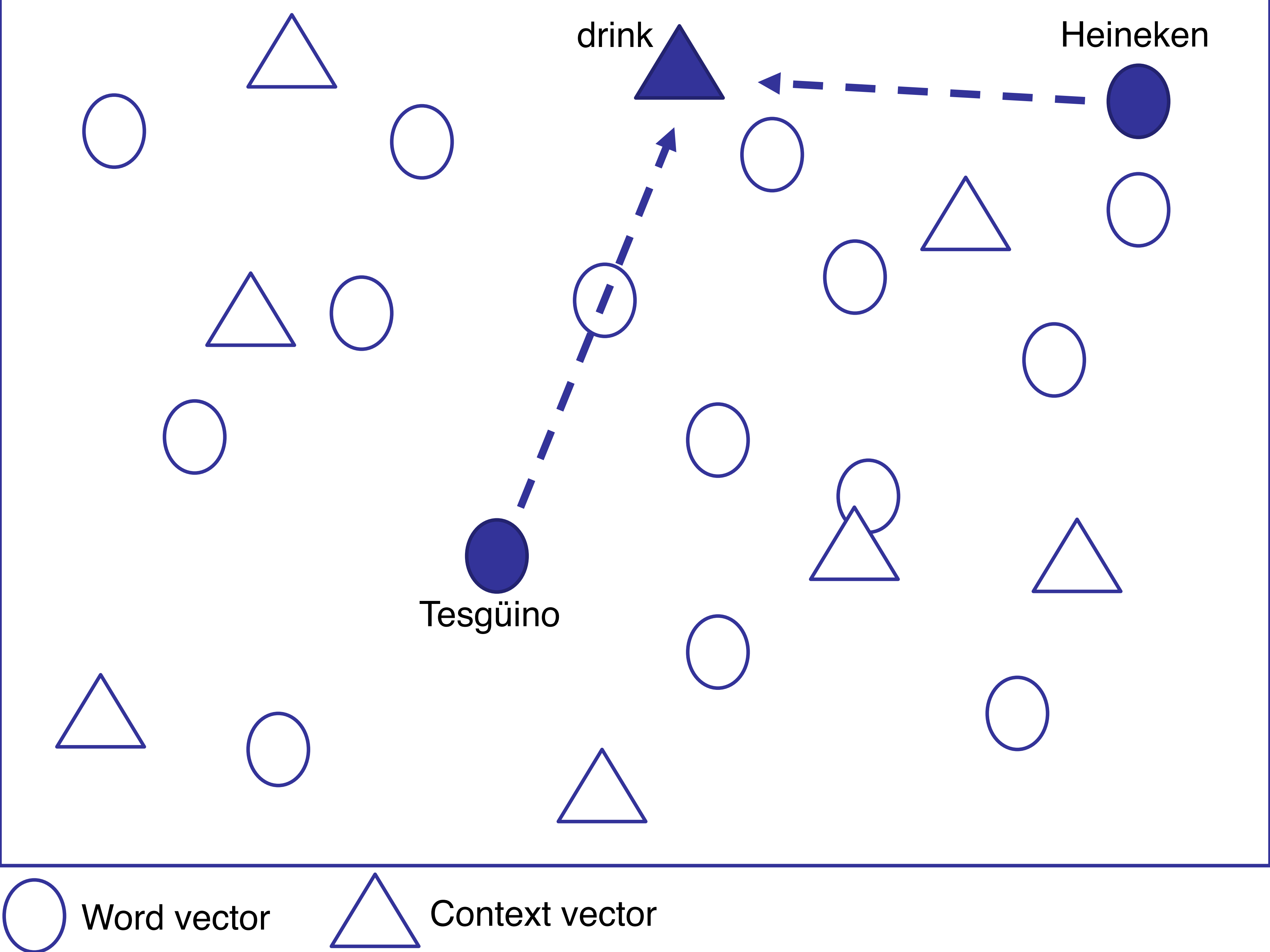


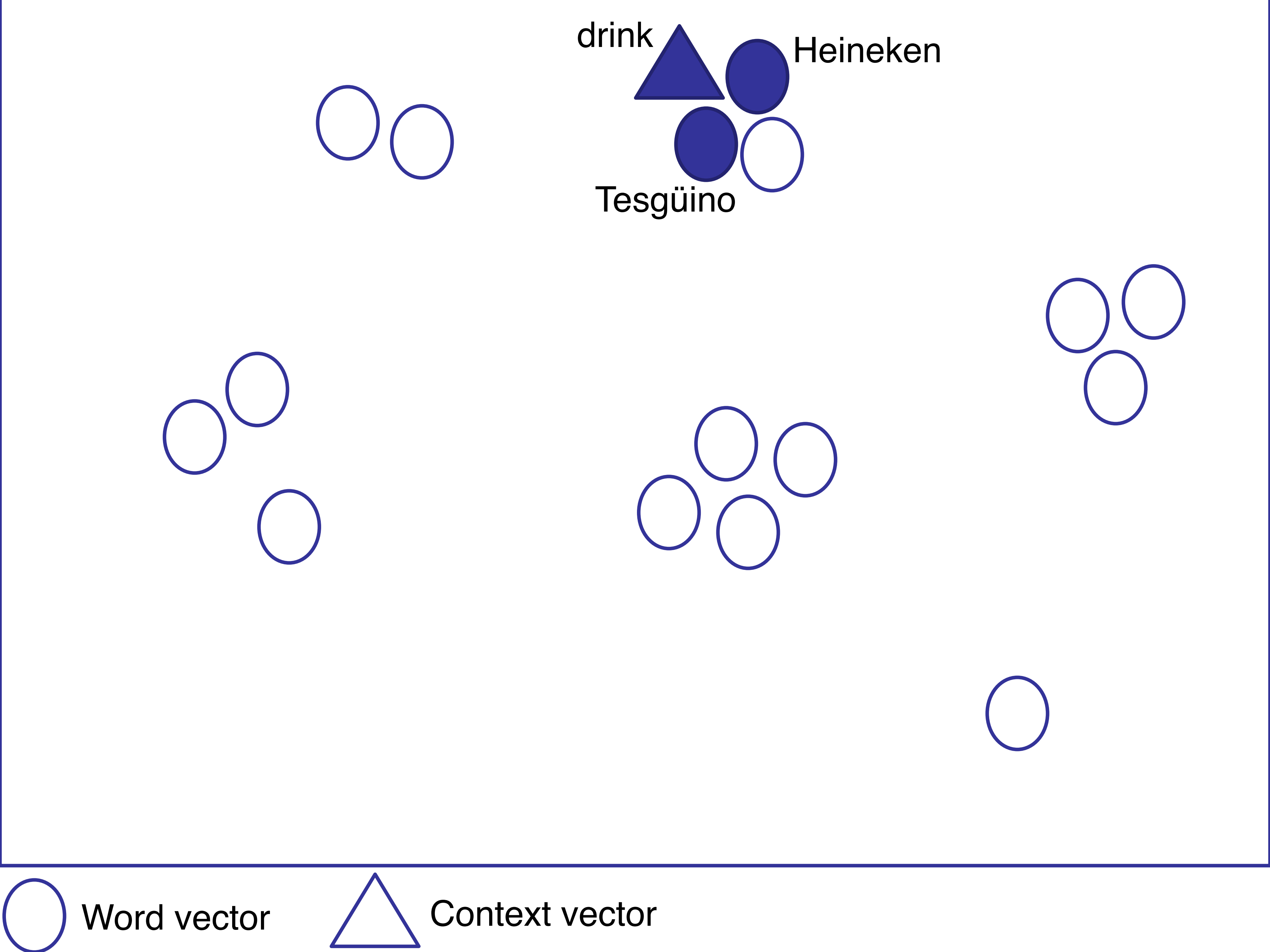
https://web.stanford.edu/~jurafsky/slp3/

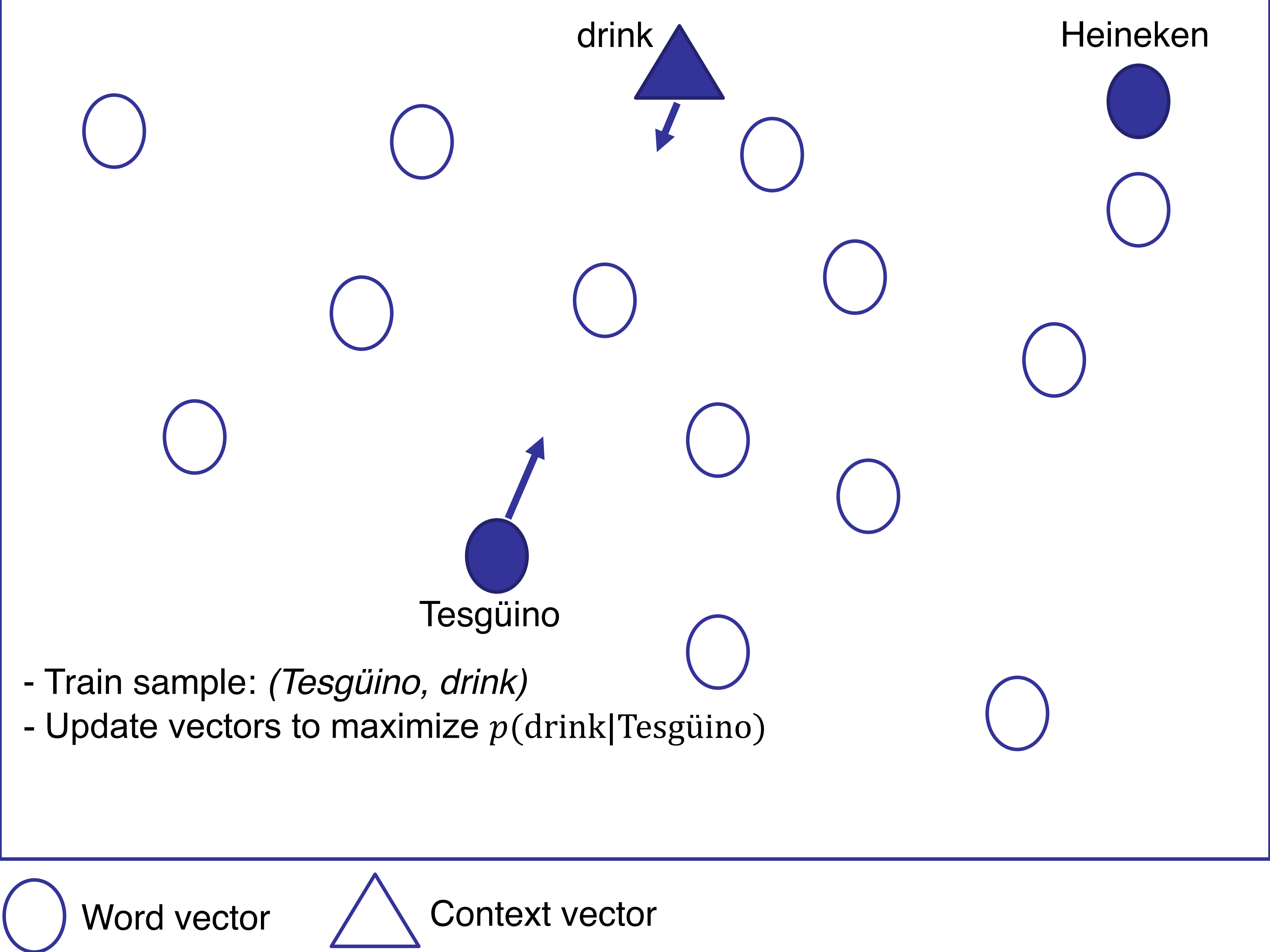

Heineken
Tesgüino
Word vector


Heineken
Tesgüino
Word vector


Heineken
Tesgüino
Word vector
Context vector


drink
Heineken
Tesgüino
Word vector
Context vector


drink
Heineken
Tesgüino
Word vector
Context vector


drink
Heineken
Tesgüino
- Train sample: *(Tesgüino, drink)*
- Update vectors to maximize $p(\text{drink}|\text{Tesgüino})$
Word vector
Context vector

Neural Word Embedding - Summary

- Output value is equal to: $\overrightarrow{W}_{\text{Tesgüino}} \cdot \vec{C}_{\text{drink}}$

- Output layer is normalized with Softmax

$$p(\text{drink}|\text{Tesgüino}) = \frac{e^{\overrightarrow{W}_{\text{Tesgüino}} \cdot \vec{C}_{\text{drink}}}}{\sum_{q \in V} e^{\overrightarrow{W}_{\text{Tesgüino}} \cdot \vec{C}_q}}$$

Normalization is too expensive!

- Cost function for all training samples

$$J = -\frac{1}{T} \sum_{1}^{T} \log p(c|w)$$

word2vec (SkipGram) with Negative Sampling

- word2vec an efficient and effective algorithm
- Instead of $p(c|w)$, word2vec measures $p(y = 1|w, c)$:
  - the probability that the co-occurrence of (w, c) comes from a genuine probability distribution, estimated with sigmoid

$$p(y = 1|w, c) = \sigma(\overrightarrow{W_w}, \vec{C_c}) = \frac{1}{1 + e^{-\overrightarrow{W_w} \cdot \vec{C_c}}}$$

- For a training sample (w, c), we aim to increase this probability.
- To contrast training samples (genuine ones) from the others, we select *k* negative samples of context words $\check{c}$
  - with random sampling from words distribution → why?
- We aim to decrease the probability of negative samples $p(y = 1|w, \check{c})$

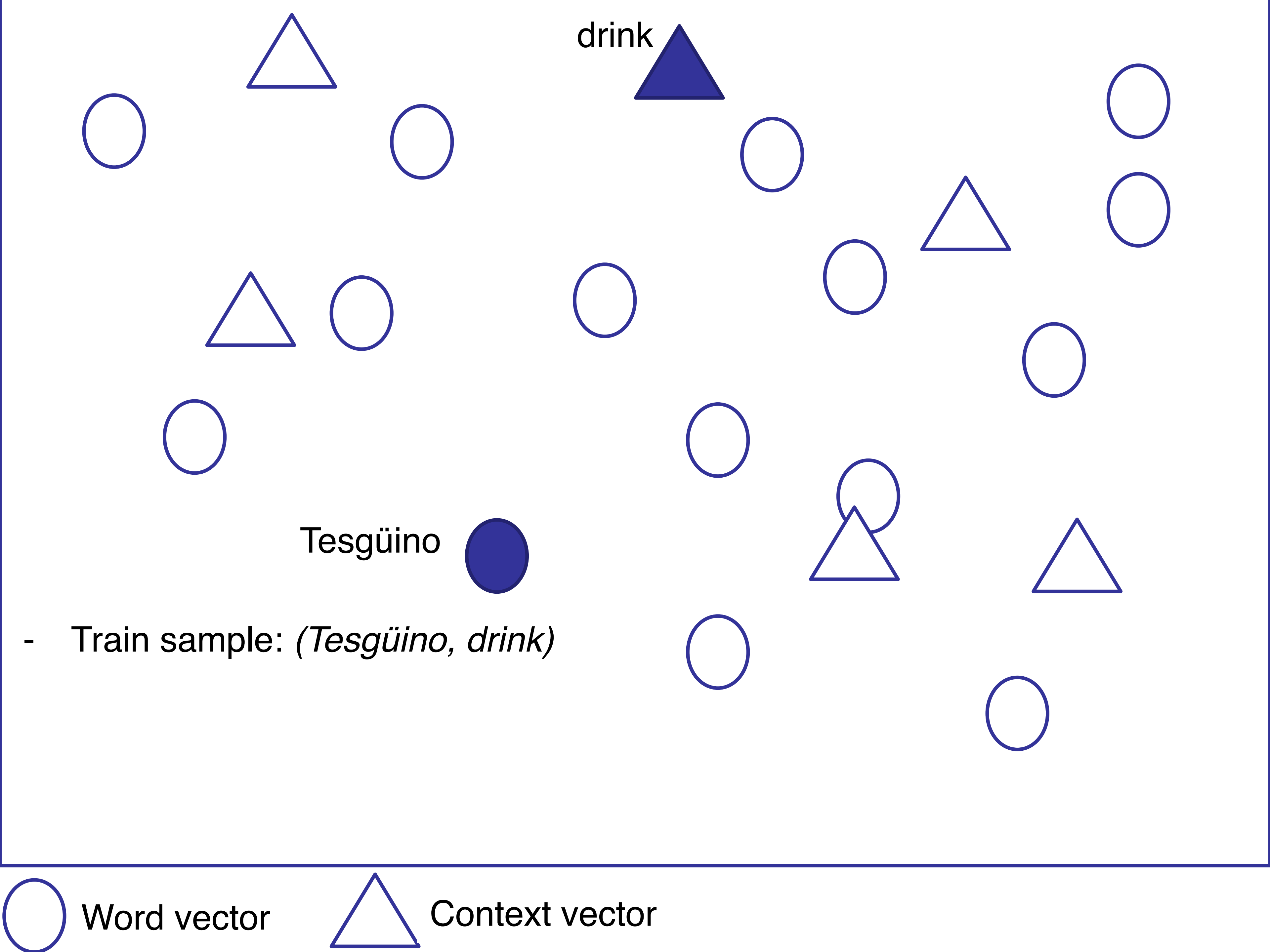
word2vec with Negative Sampling – Objective Function

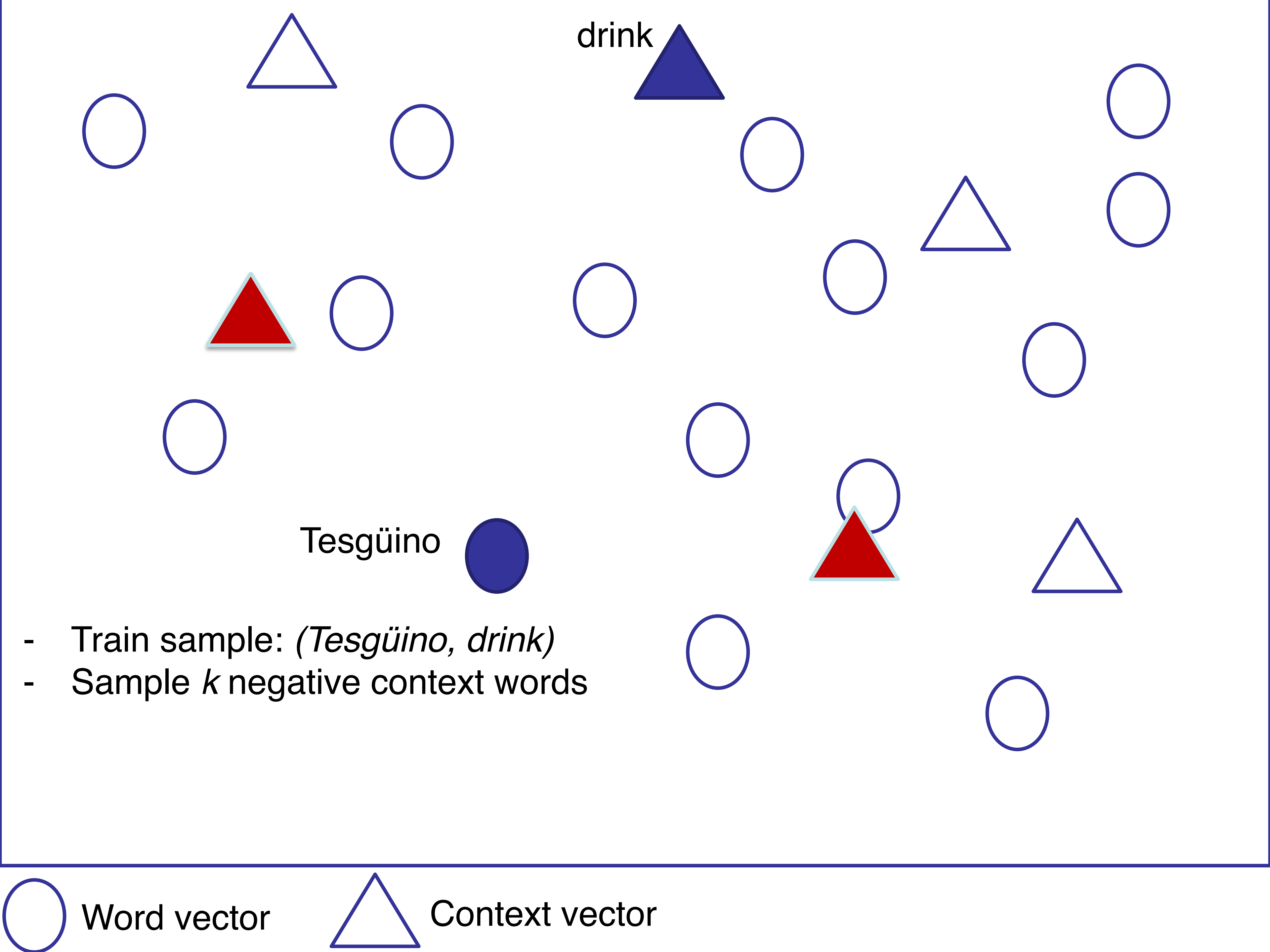
- $k \sim$ 2-10

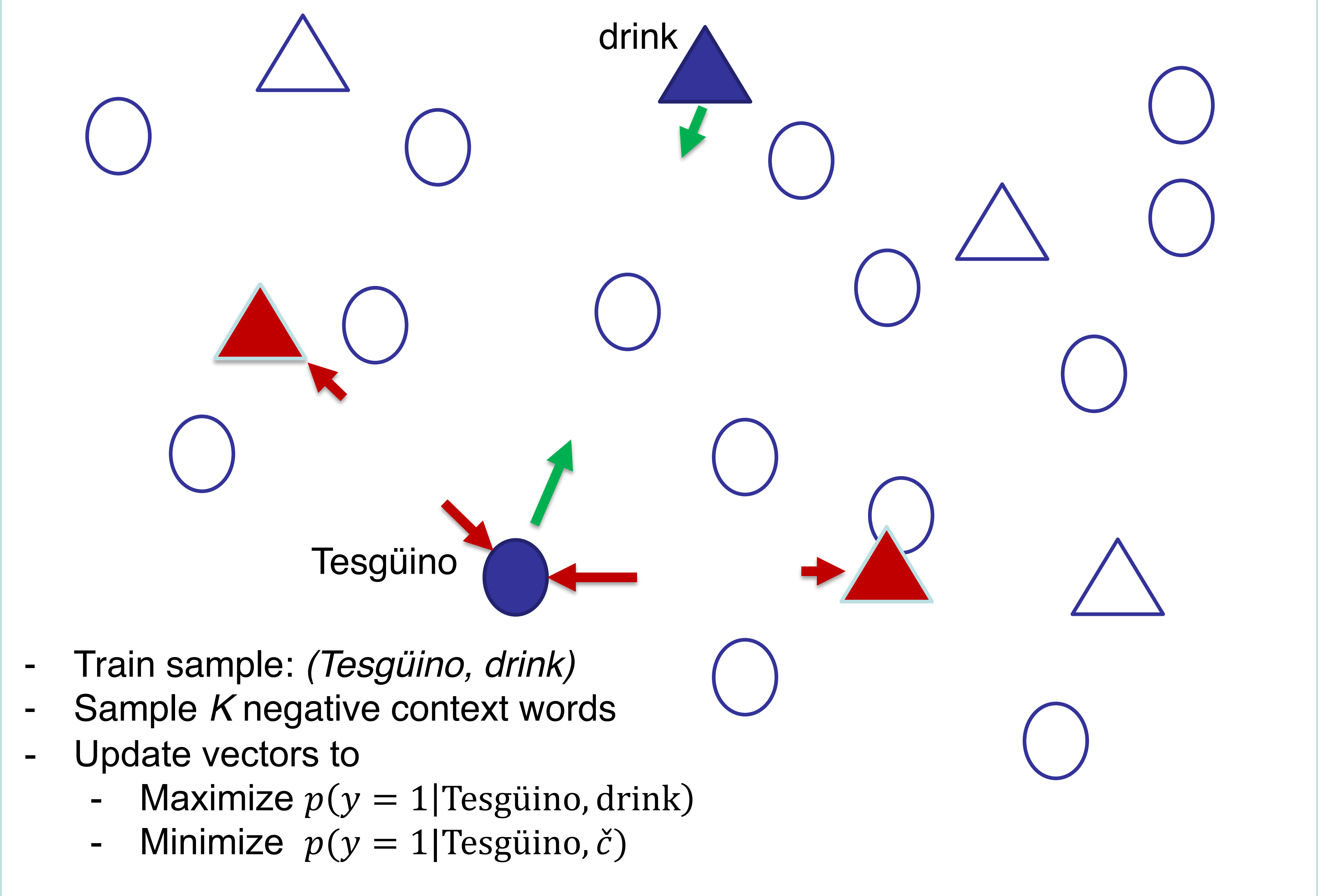
$$J = -\frac{1}{T}\sum_{1}^{T}\left[\log p(y = 1|w, c) - \sum_{i=1}^{k}\log p(y = 1|w, \check{c})\right]$$

Training Samples

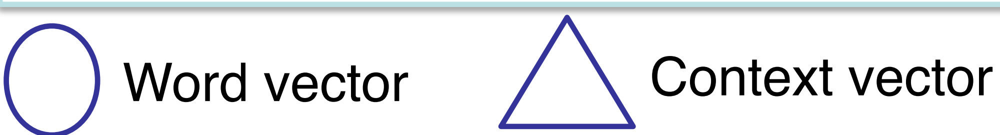
Negative Samples


drink
Tesgüino
- Train sample: *(Tesgüino, drink)*
Word vector
Context vector


drink
Tesgüino
- Train sample: (Tesgüino, drink)
- Sample k negative context words
Word vector
Context vector



- Train sample: *(Tesgüino, drink)*
- Sample K negative context words
- Update vectors to
  - Maximize $p(y = 1|\text{Tesgüino}, \text{drink})$
  - Minimize $p(y = 1|\text{Tesgüino}, \check{c})$



Word Embedding - Evaluation

- Intrinsic
  - Given a list of pairs and their relatedness, judged by human, what is the correlation of similarity values, when generated by a word embedding model.
- Extrinsic
  - The effect of using a word embedding model in another task such as sentiment analysis, document classification, document retrieval, etc..

"In general word2vec provides a strong and consistent baseline and a good starting point for task-specific representation learning"

Agenda

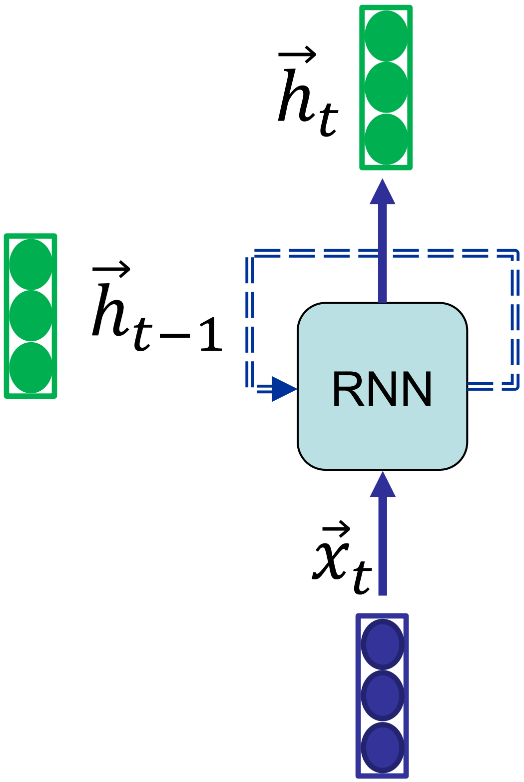
- Crash course (1)
  - Vector representation
  - Neural Networks
- Word Representation Learning
  - Neural word representation
  - word2vec

---Break---

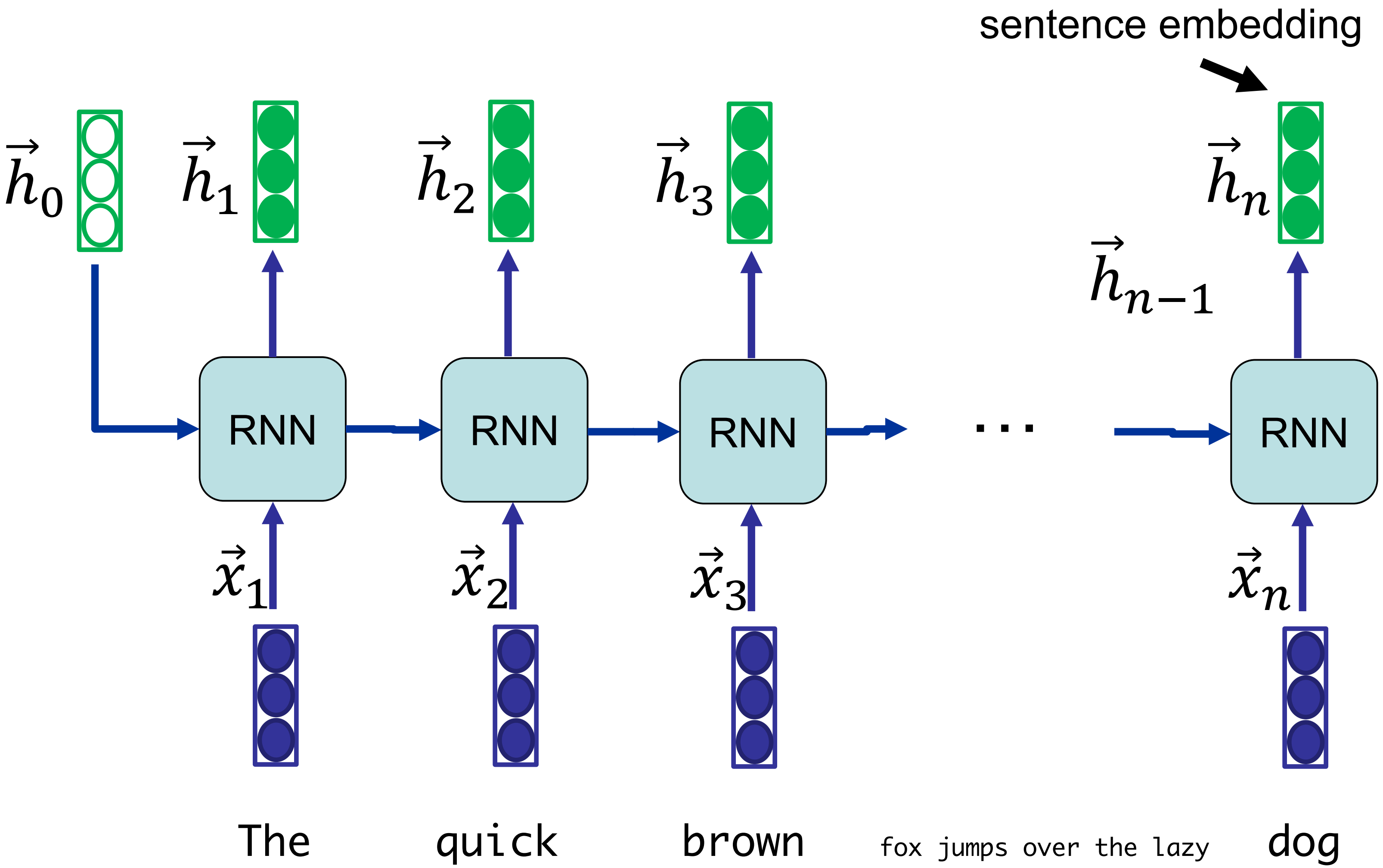
- Crash course (2)
  - Recurrent Neural Networks
  - Attention Mechanism
- Document Classification
- Applications and challenges

Recurrent Neural Networks

- Encodes sequences of entities
  - Sequence of words
  - Time series
- The output in every step uses the output of the previous step
- It carries a memory of past entities



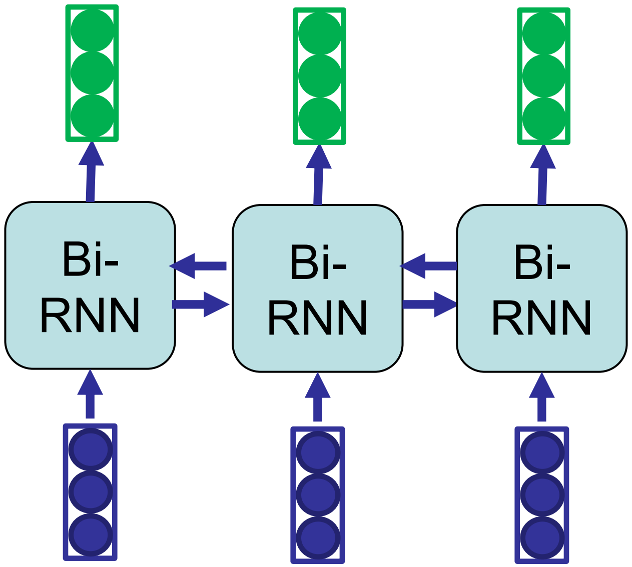
Recurrent Neural Networks



Recurrent Neural Networks

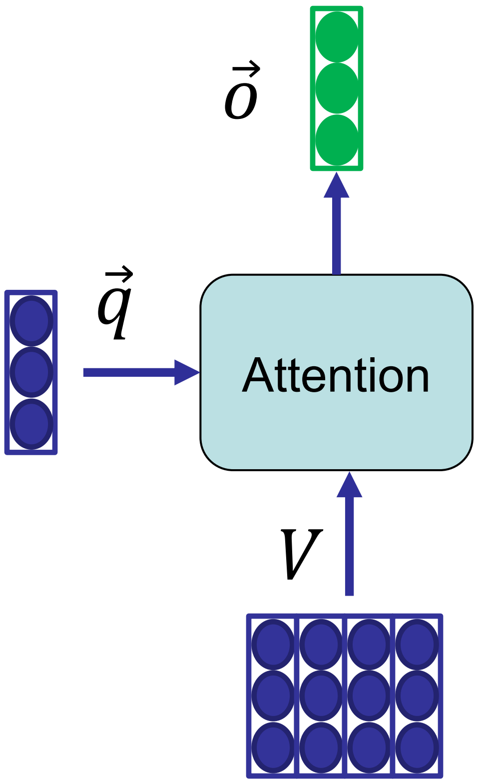
- Various types
  - Standard (Elman) RNN
  - Long Short Term Memory (LSTM)
  - Gated Recurrent Unit (GRU)

- Bi-directional RNN
  - Two RNNs: the first reads from beginning to end, the second from end to beginning
  - Output at each time step is the sum of the hidden states of both RNNs

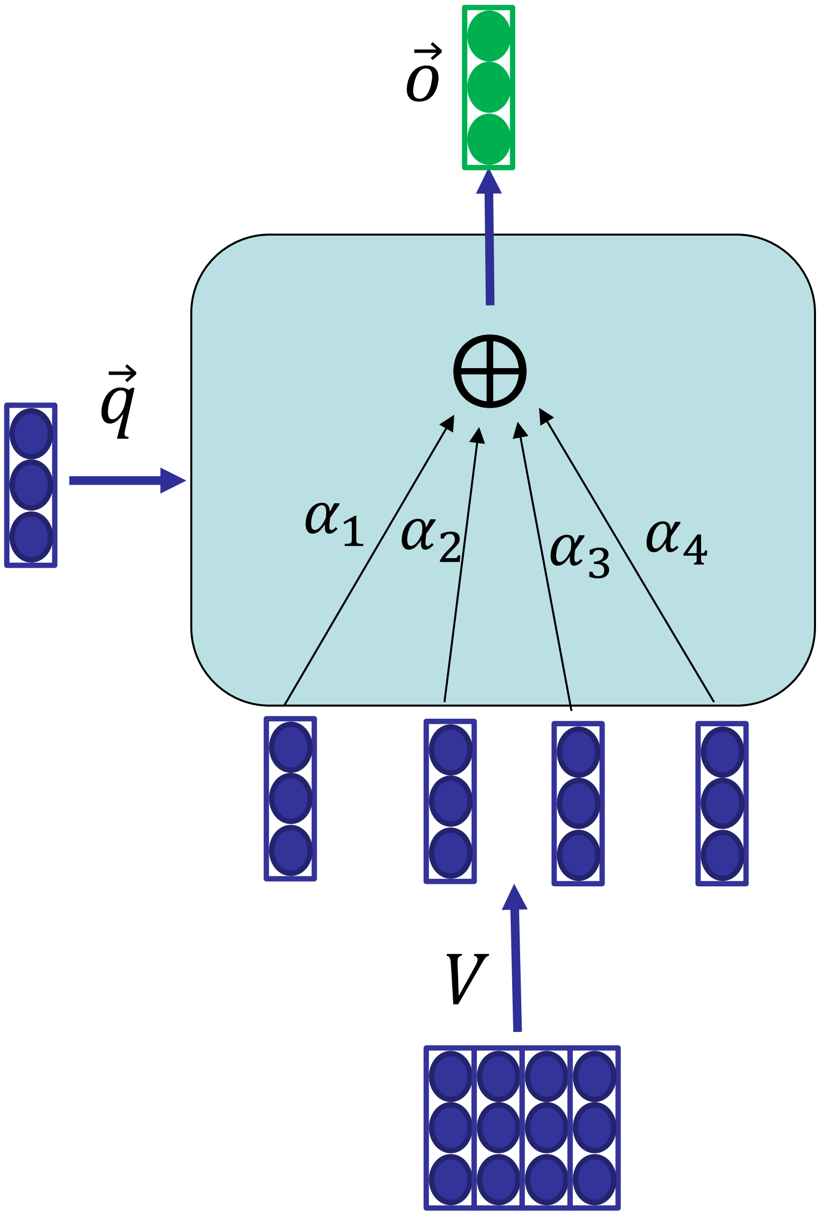


Attention Mechanism (simplified)

- Given a query $\vec{q}$ and a values matrix V, it “looks up” a vector output $\vec{o}$
- It learns to put different degrees of attentions on the entities of the values matrix
- Output is the weighted sum of the V vectors



Attention Mechanism



$$\alpha_i = f(\vec{q}, \vec{v_i})$$

$$\sum_{i=1}^{n} \alpha_i = 1$$

$$\vec{o} = \sum_{i=1}^{n} \alpha_i . \vec{v_i}$$

Document Classification - Recap

- Regrouping documents with similar concepts (classes)

- Prepare a document representation, e.g. using
  - TF-IDF
  - Latent Semantic Indexing (LSI)
  - Latent Dirichlet Allocation (LDA)

- Supervised Classification
  - Given training data, learn a discriminator model on document representations to separate the classes
  - Use the discriminator model to classify the test-set documents
  - Evaluate the accuracy of prediction

Document Classification - Recap

- Sample test collections to evaluate the effectiveness of methods

Data set	classes	documents	average #s	max #s	average #w	max #w	vocabulary
Yelp 2013	5	335,018	8.9	151	151.6	1184	211,245
Yelp 2014	5	1,125,457	9.2	151	156.9	1199	476,191
Yelp 2015	5	1,569,264	9.0	151	151.9	1199	612,636
IMDB review	10	348,415	14.0	148	325.6	2802	115,831
Yahoo Answer	10	1,450,000	6.4	515	108.4	4002	1,554,607
Amazon review	5	3,650,000	4.9	99	91.9	596	1,919,336

Table 1: Data statistics: #s denotes the number of sentences (average and maximum per document), #w denotes the number of words (average and maximum per document).

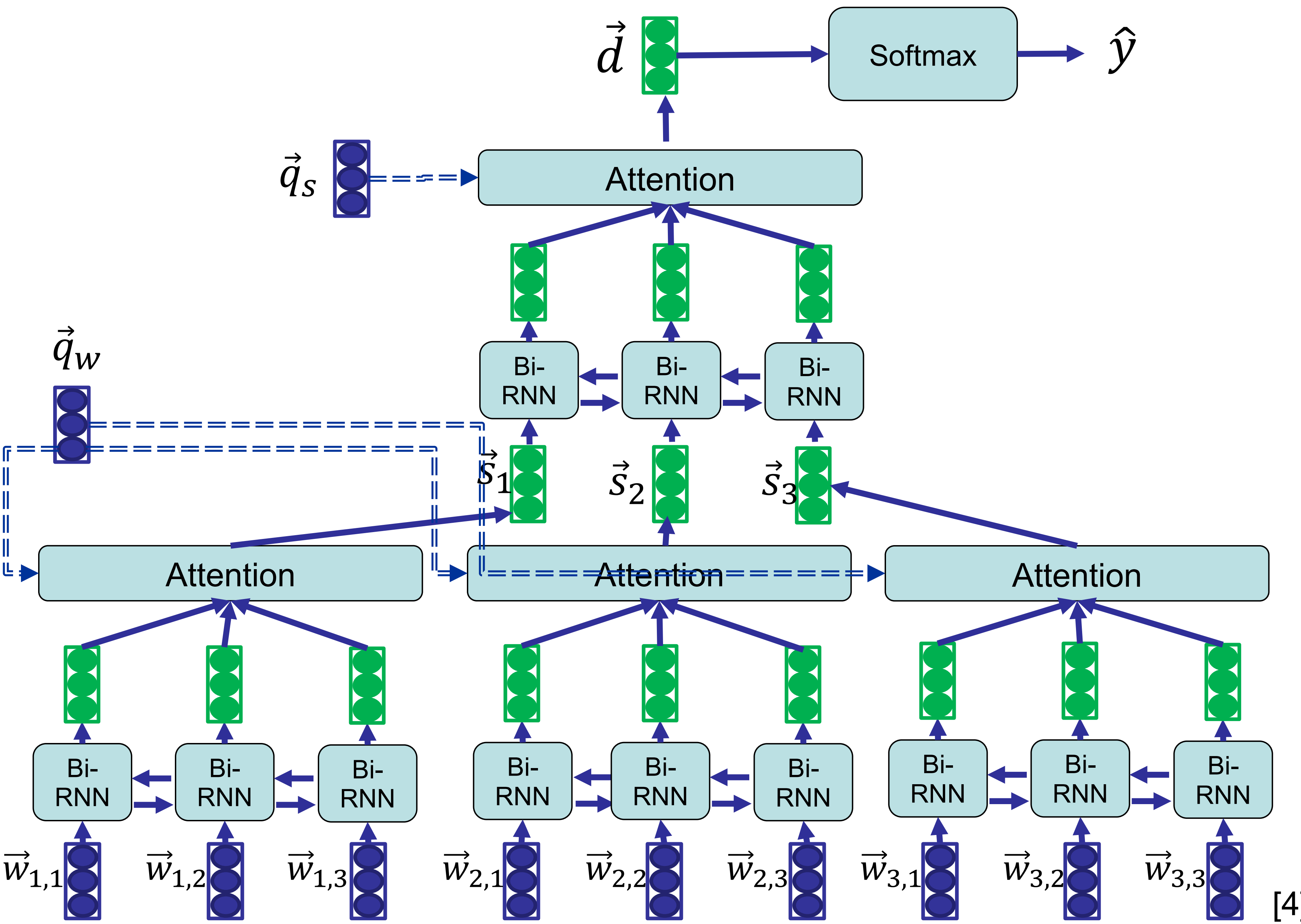
[4]

Document Representation

- Document representation is the key!
- With a better document representation, the classes can be more effectively separated

Two sample document representation sets, projected to two-dimensional spaces

source source



$\vec{d}$
Softmax
$\hat{y}$
$\vec{q}_s$
Attention
$\vec{q}_w$
Bi-RNN
$\vec{s}_1$
$\vec{s}_2$
$\vec{s}_3$
Attention
Attention
Attention
Bi-RNN
$\vec{w}_{1,1}$
$\vec{w}_{1,2}$
$\vec{w}_{1,3}$
$\vec{w}_{2,1}$
$\vec{w}_{2,2}$
$\vec{w}_{2,3}$
$\vec{w}_{3,1}$
$\vec{w}_{3,2}$
$\vec{w}_{3,3}$

[4]

Document Representation Learning

- $\vec{w}$ word embedding
- $\vec{q}_w$ query representation for words: *what is the informative word?*
- $\vec{s}$ high level sentence representation
- $\vec{q}_s$ query representation for sentences: *which sentence is informative?*
- $\vec{d}$ high level document representation
- $\hat{y}$ class prediction

Sample Evaluation Results

	Methods	Yelp'13	Yelp'14	Yelp'15	IMDB	Yahoo Answer	Amazon
Zhang et al., 2015	BoW	-	-	58.0	-	68.9	54.4
	BoW TFIDF	-	-	59.9	-	71.0	55.3
	ngrams	-	-	56.3	-	68.5	54.3
	ngrams TFIDF	-	-	54.8	-	68.5	52.4
	Bag-of-means	-	-	52.5	-	60.5	44.1
Tang et al., 2015	Majority	35.6	36.1	36.9	17.9	-	-
	SVM + Unigrams	58.9	60.0	61.1	39.9	-	-
	SVM + Bigrams	57.6	61.6	62.4	40.9	-	-
	SVM + TextFeatures	59.8	61.8	62.4	40.5	-	-
	SVM + AverageSG	54.3	55.7	56.8	31.9	-	-
	SVM + SSWE	53.5	54.3	55.4	26.2	-	-
Zhang et al., 2015	LSTM	-	-	58.2	-	70.8	59.4
	CNN-char	-	-	62.0	-	71.2	59.6
	CNN-word	-	-	60.5	-	71.2	57.6
Tang et al., 2015	Paragraph Vector	57.7	59.2	60.5	34.1	-	-
	CNN-word	59.7	61.0	61.5	37.6	-	-
	Conv-GRNN	63.7	65.5	66.0	42.5	-	-
	LSTM-GRNN	65.1	67.1	67.6	45.3	-	-
This paper	HN-AVE	67.0	69.3	69.9	47.8	75.2	62.9
	HN-MAX	66.9	69.3	70.1	48.2	75.2	62.9
	HN-ATT	**68.2**	**70.5**	**71.0**	**49.4**	**75.8**	**63.6**

[4]

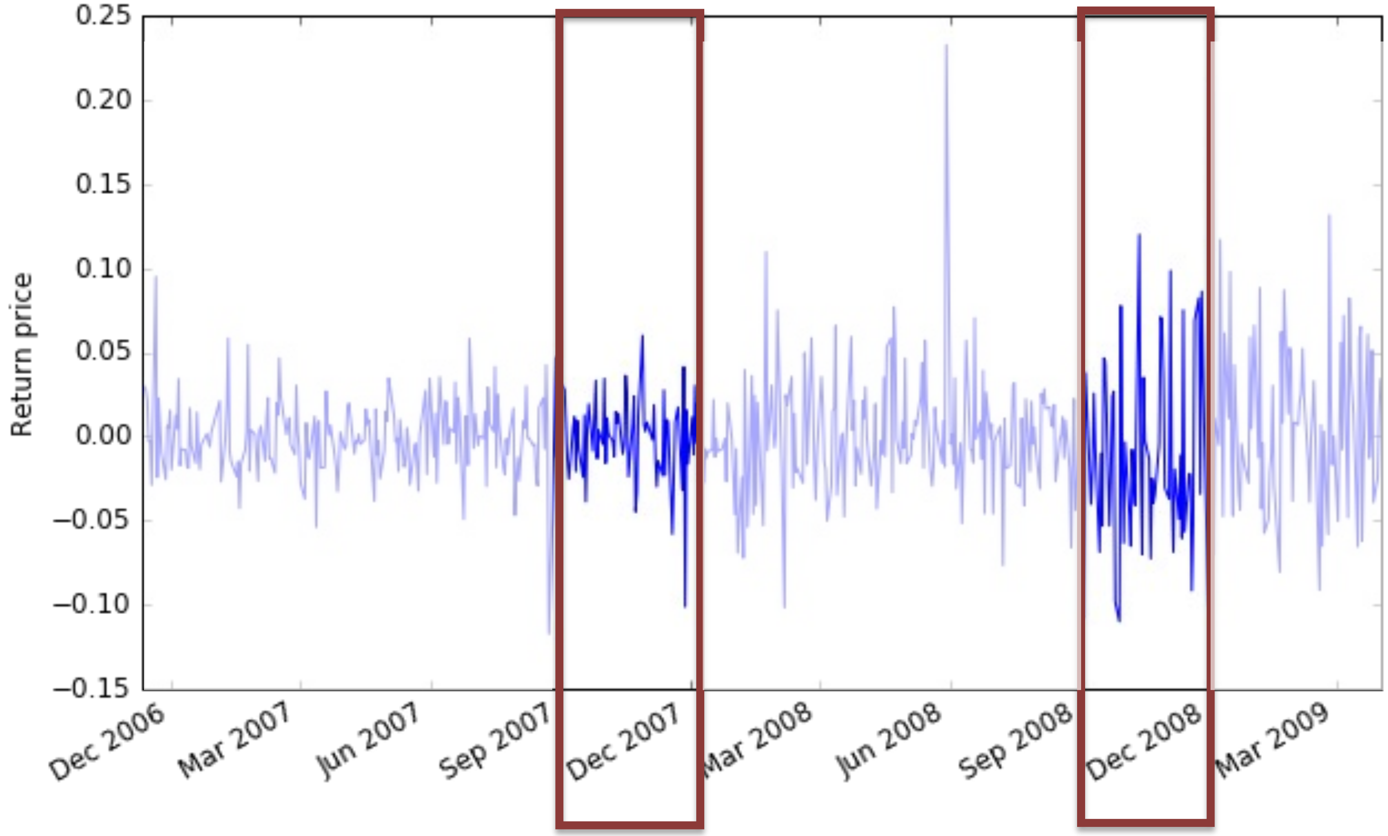
Agenda

- Crash course (1)
  - Vector representation
  - Neural Networks
- Word Representation Learning
  - Neural word representation
  - word2vec

---Break---

- Crash course (2)
  - Recurrent Neural Networks
  - Attention Mechanism
- Document Classification
- Applications and challenges

Volatility in Financial System



return price=(price$_{(t)}$ / price$_{(t-1)}$)-1

volatility=log(std(return prices))

Companies Annual Reports

UNITED STATES
SECURITIES AND EXCHANGE COMMISSION
Washington, D.C. 20549

FORM 10-K

☒ **ANNUAL REPORT PURSUANT TO SECTION 13 OR 15(d) OF THE SECURITIES EXCHANGE ACT OF 1934**

For the fiscal year ended May 31, 2011

OR

☐ **TRANSITION REPORT PURSUANT TO SECTION 13 OR 15(d) OF THE SECURITIES EXCHANGE ACT OF 1934**

For the transition period from ______ to ______
Commission file number: 000-51788

Oracle Corporation

(Exact name of registrant as specified in its charter)

Delaware	**54-2185193**
(State or other jurisdiction of incorporation or organization)	**(I.R.S. Employer Identification No.)**
500 Oracle Parkway Redwood City, California	**94065**
(Address of principal executive offices)	**(Zip Code)**

(650) 506-7000
(Registrant's telephone number, including area code)

Securities registered pursuant to Section 12(b) of the Act:

Title of each class	**Name of each exchange on which registered**
Common Stock, par value $0.01 per share	The NASDAQ Stock Market LLC

Securities registered pursuant to Section 12(g) of the Act:
None

Indicate by check mark if the registrant is a well-known seasoned issuer, as defined in Rule 405 of the Securities Act. YES ☒ NO ☐

Indicate by check mark if the registrant is not required to file reports pursuant to Section 13 or Section 15(d) of the Act. YES ☐ NO ☒

Indicate by check mark whether the registrant (1) has filed all reports required to be filed by Section 13 or 15(d) of the Securities Exchange Act of 1934 during the preceding 12 months (or for such shorter period that the registrant was required to file such reports), and (2) has been subject to such filing requirements for the past 90 days. YES ☒ NO ☐

Indicate by check mark whether the registrant has submitted electronically and posted on its corporate Website, if any, every Interactive Data File required to be submitted and posted pursuant to Rule 405 of Regulation S-T (§232.405 of this chapter) during the preceding 12 months (or for such shorter period that the registrant was required to submit and post such files). YES ☒ NO ☐

Indicate by check mark if disclosure of delinquent filers pursuant to Item 405 of Regulation S-K (§229.405 of this chapter) is not contained herein, and will not be contained, to the best of registrant's knowledge, in definitive proxy or information statements incorporated by reference in Part III of this Form 10-K or any amendment to this Form 10-K. ☐

Indicate by check mark whether the registrant is a large accelerated filer, an accelerated filer, a non-accelerated filer, or a smaller reporting company. See the definitions of "large accelerated filer," "accelerated filer" and "smaller reporting company" in Rule 12b-2 of the Exchange Act.

Large accelerated filer ☒	Accelerated filer ☐
Non-accelerated filer ☐ (Do not check if a smaller reporting company)	Smaller reporting company ☐

Indicate by check mark whether the registrant is a shell company (as defined in Rule 12b-2 of the Exchange Act). YES ☐ NO ☒

The aggregate market value of the voting stock held by non-affiliates of the registrant was $107,183,061,000 based on the number of shares held by non-affiliates of the registrant as of May 31, 2011, and based on the closing sale price of common stock as reported by the NASDAQ Global Select Market on November 30, 2010, which is the last business day of the registrant's most recently completed second fiscal quarter. This calculation does not reflect a determination that persons are affiliates for any other purposes.

Number of shares of common stock outstanding as of June 20, 2011: 5,065,515,000.

Documents Incorporated by Reference:

Portions of the registrant's definitive proxy statement relating to its 2011 annual stockholders' meeting are incorporated by reference into Part III of this Annual Report on Form 10-K where indicated.

manufacturing, professional services, public sector, retail, travel, transportation and utilities. For example, we offer the banking and financial services sector a suite of applications addressing cash management, trade, treasury, payments, lending, private wealth management, asset management, compliance, enterprise risk and business analytics, among others. We offer the retail sector software solutions designed to provide unified and actionable data among store, merchandising and financial operations. Our applications for consumer goods manufacturers are designed to provide them with the ability to build their brand against retail private label programs by engaging directly with the consumer. Our ability to offer applications to address industry-specific complex processes provides us an opportunity to expand our customers' knowledge of our broader product offerings and address customer specific technology challenges.

Software License Updates and Product Support

We seek to protect and enhance our customers' current investments in Oracle software by offering proactive and personalized support services, including our Lifetime Support policy, and unspecified product enhancements and upgrades. Software license updates provide customers with rights to unspecified software product upgrades and maintenance releases and patches released during the term of the support period. Product support includes internet and telephone access to technical support personnel located in our global support centers, as well as internet access to technical content through "My Oracle Support." Software license updates and product support contracts are generally priced as a percentage of the net new software license fees. Substantially all of our customers purchase software license updates and product support contracts when they acquire new software licenses and renew their software license updates and product support contracts annually. Our software license updates and product support revenues represented 42%, 49% and 50% of our total revenues in fiscal 2011, 2010 and 2009, respectively.

Hardware Systems Business

As a result of our acquisition of Sun in January 2010, we entered into the hardware systems business. Our hardware systems business consists of two operating segments: hardware systems products and hardware systems support.

Hardware Systems Products

Our customers demand a broad set of hardware systems solutions to manage growing amounts of data and computational requirements, to meet increasing compliance and regulatory demands, and to reduce energy, space, and operational costs. To meet these demands, we have a wide variety of innovative hardware systems offerings, including servers and storage products, networking components, operating systems and other hardware-related software. Our hardware systems component products are designed to be "open," or to work in customer environments that may include other Oracle or non-Oracle hardware or software components. We have also engineered our hardware systems products to create performance and operational cost advantages for customers when our hardware and software products are combined as engineered systems, as with Oracle Exadata and Oracle Exalogic Elastic Cloud. By combining our server and storage hardware with our software, our open, integrated products better address customer requirements for performance, scalability, reliability, security, ease of management, and lower total cost of ownership. Our hardware systems products represented 12% and 6% of our total revenues in fiscal 2011 and 2010, respectively.

Servers

We offer a wide range of server systems using our SPARC microprocessor. Our SPARC servers are differentiated by their reliability, security and scalability; and by the customer environments that they target (general purpose or specialized systems). Our midsize and large servers are designed to offer greater performance and lower total cost of ownership than mainframe systems for business critical applications and for customers having more computationally intensive needs. Our SPARC servers run the Oracle Solaris operating system and are designed for the most demanding mission critical enterprise environments at any scale. We have a long-standing relationship with Fujitsu Limited for the development, manufacturing and marketing of certain of our SPARC server components and products.

13

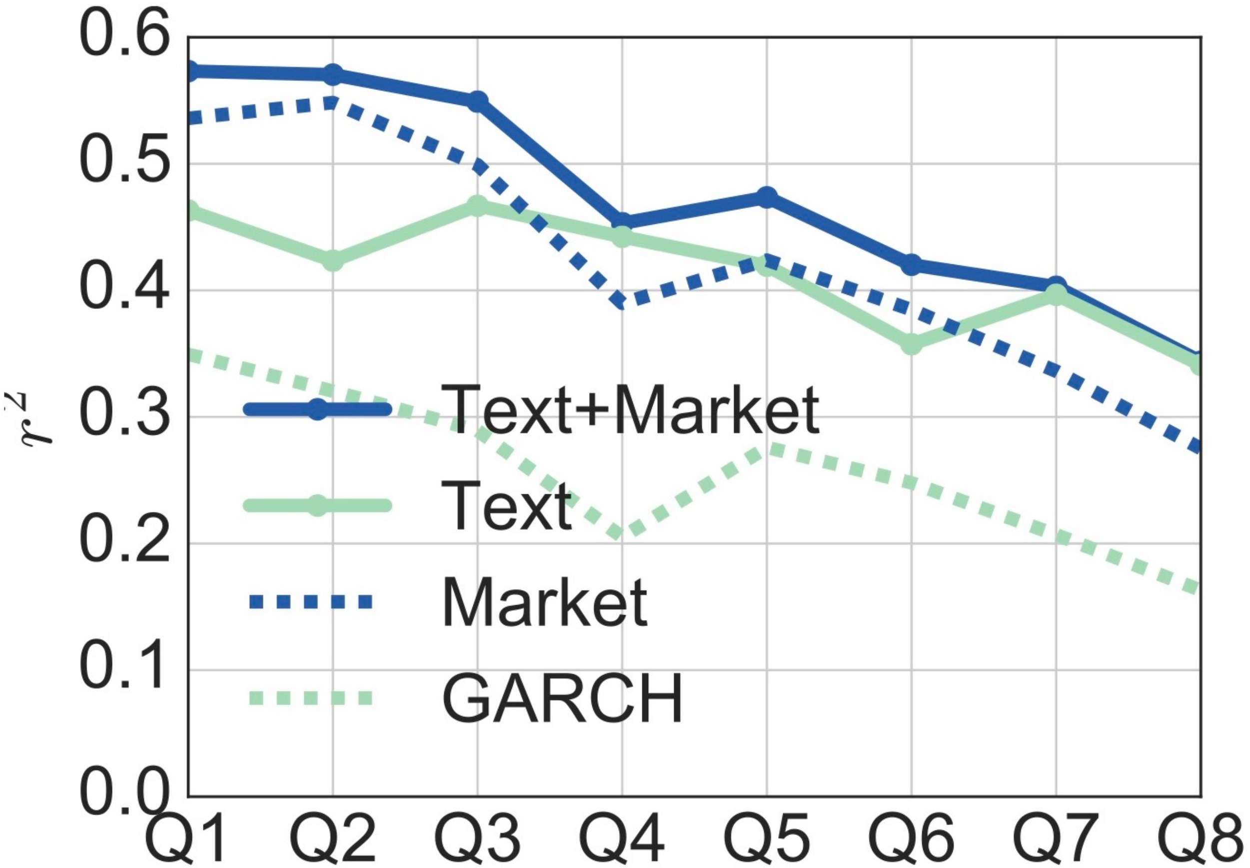
Servers

We offer a wide range of server systems using our SPARC microprocessor. Our SPARC servers are differentiated by their reliability, security and scalability; and by the customer environments that they target (general purpose or specialized systems). Our midsize and large servers are designed to offer greater performance and lower total cost of ownership than mainframe systems for business critical applications and for customers having more computationally intensive needs. Our SPARC servers run the Oracle Solaris operating system and are designed for the most demanding mission critical enterprise environments at any scale. We have a long-standing relationship with Fujitsu Limited for the development, manufacturing and marketing of certain of our SPARC server components and products.

13

Volatility Prediction with Sentiment Analysis

- Using sentiment analysis (Text) to predict the volatility of upcoming quartiles
- Text data significantly improves prediction



[5]

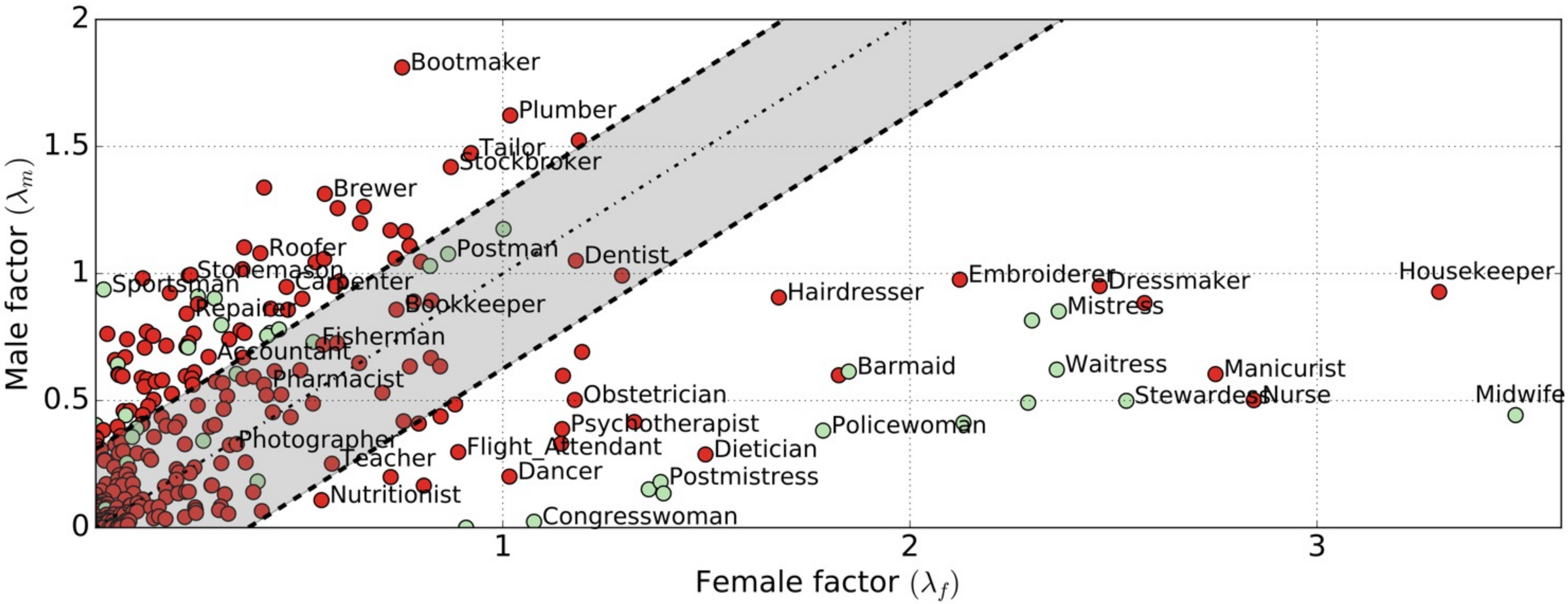
Semantic Change

Figure 1: Two-dimensional visualization of semantic change in English using SGNS vectors.[2] **a**, The word *gay* shifted from meaning "cheerful" or "frolicsome" to referring to homosexuality. **b**, In the early 20th century *broadcast* referred to "casting out seeds"; with the rise of television and radio its meaning shifted to "transmitting signals". **c**, *Awful* underwent a process of pejoration, as it shifted from meaning "full of awe" to meaning "terrible or appalling" (Simpson et al., 1989).

[6]

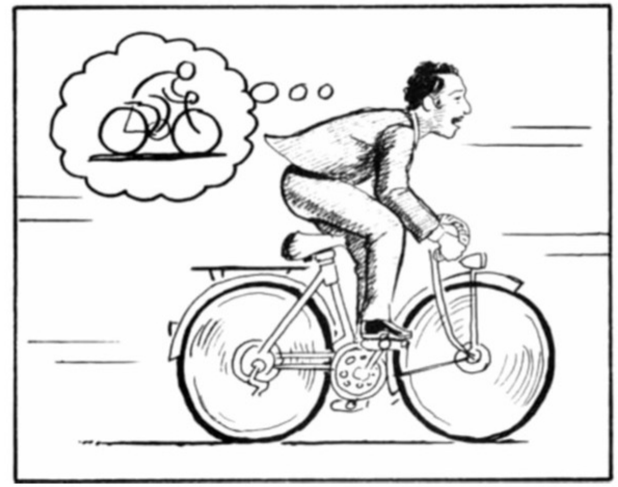
Quantification of Gender Bias

- The tendencies of various jobs to genders, based on the representation of words, trained on the Wikipedia.
- It indicates the ethical bias in data.



Challenges

- Semantics of language and the world



[8]

"The image of the world around us, which we carry in our head, is just a model. Nobody in his[/her] head imagines all the world, government or country. He[/She] has only selected concepts, and relationships between them, and uses those to represent the real system." Mental Model [7]

- Representation Learning
  - Abstract representation of spatial and temporal aspects of information
  - Various granularities → abstraction level
  - Task-specific, domain specific → transfer learning
  - Commonalities among languages → multilingual models

Challenges

- Understanding information contents
  - Information Retrieval
  - Summarization

- Aspects of information tailoring and provision
  - Personalization vs. De-personalization
  - Controversy detection
  - Multiple points of view

Challenges

- Exploring aspects of society
  - Computational Social Science with NLP
    - Economy
    - Sociology
    - Psychology
- Ethics, fairness, and transparency
  - implications of the new technology on the society
  - Ownership of data and models, laws, etc.
  - Ethical bias in data and algorithms
  - Interpretability of the models


FINANCIAL TIMES myFT

Special Report **Modern Workplace: Ethnic Diversity**

Workplace diversity + Add to myFT

AI risks replicating tech's ethnic minority bias across business

Diverse workforce essential to combat new danger of 'bias in, bias out'

References

[1] Jurafsky, Dan, and James H. Martin. *Speech and language processing*. Vol. 3. London: Pearson, 2014.

[2] Kulkarni, Vivek, et al. "Statistically significant detection of linguistic change." *Proceedings of the 24th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 2015.

[3] Mikolov, Tomas, et al. "Distributed representations of words and phrases and their compositionality." *Advances in neural information processing systems*. 2013.

[4] Yang, Zichao, et al. "Hierarchical attention networks for document classification." Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2016.

[5] Volatility Prediction using Financial Disclosures Sentiments with Word Embedding-based IR Models. Navid Rekabsaz, Mihai Lupu, Artem Baklanov, Allan Hanbury, Alexander Duer, Linda Anderson. In Proceedings of the Annual Meeting of the Association for Computational Linguistics

[6] Hamilton, William L., Jure Leskovec, and Dan Jurafsky. "Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change." Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Vol. 1. 2016.

[7] Forrester, Jay Wright. Counterintuitive behavior of so- cial systems, 1971. URL https://en.wikipedia. org/wiki/Mental_model. [Online; accessed 01- Nov-2017].

[8] Ha, David, and Jürgen Schmidhuber. "World Models." arXiv preprint arXiv:1803.10122 (2018)