SkillFactory

Introduction
Word vector representation

Sidorov Nikita

MLE(NLP) in Sber

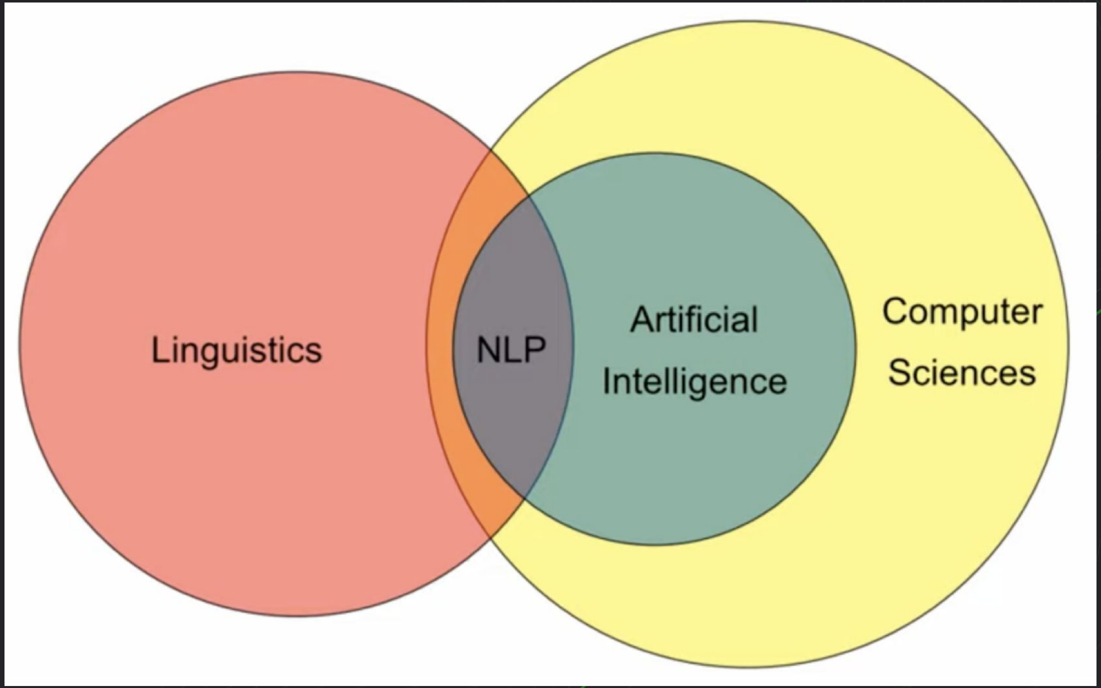
Lecture content

1. What is NLP
2. History and tasks
3. NLP toolkit
4. Words representations
5. Word2vec (overview)

NLP

Natural language processing is a subfield of linguistics, computer science, and artificial intelligence concerned with the interactions between computers and human language, in particular how to program computers to process and analyze large amounts of natural language data. The goal is a computer capable of "understanding" the contents of documents, including the contextual nuances of the language within them. (Wikipedia)

Related fields



Why this is complicated?

Jane went to store.

example from Graham Neubig lectures

Why this is complicated?

Jane went to store.

Store went to Jane.

Why this is complicated?

Jane went to store.

Store went to Jane.

Jane went store.

Why this is complicated?

Jane went to store.

Store went to Jane.

Jane went store.

Jane goed to store.

Why this is complicated?

Jane went to store.

Store went to Jane.

Jane went store.

Jane goed to store.

The store went to store.

Why this is complicated?

Jane went to store.

Store went to Jane.

Jane went store.

Create a grammar of the language

Jane goed to store.

Consider morphology and exceptions

The store went to store.

Semantic categories and exceptions

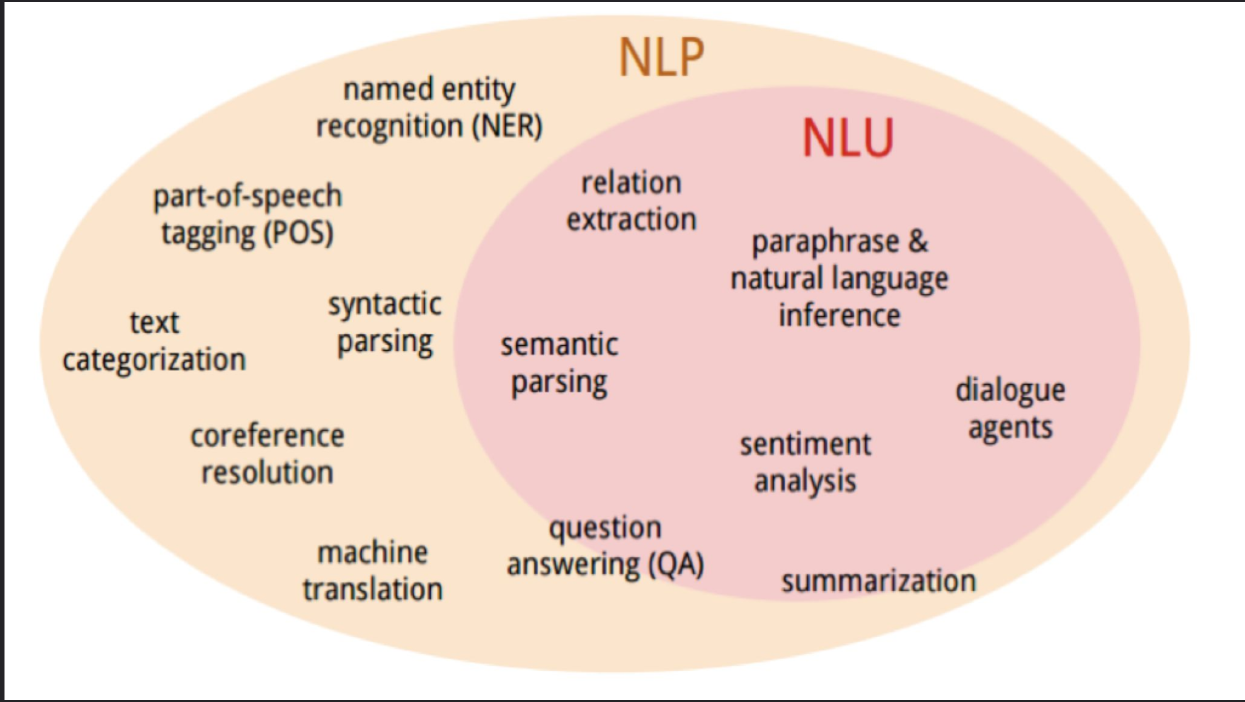
The food truck went to Jane.

And their exceptions

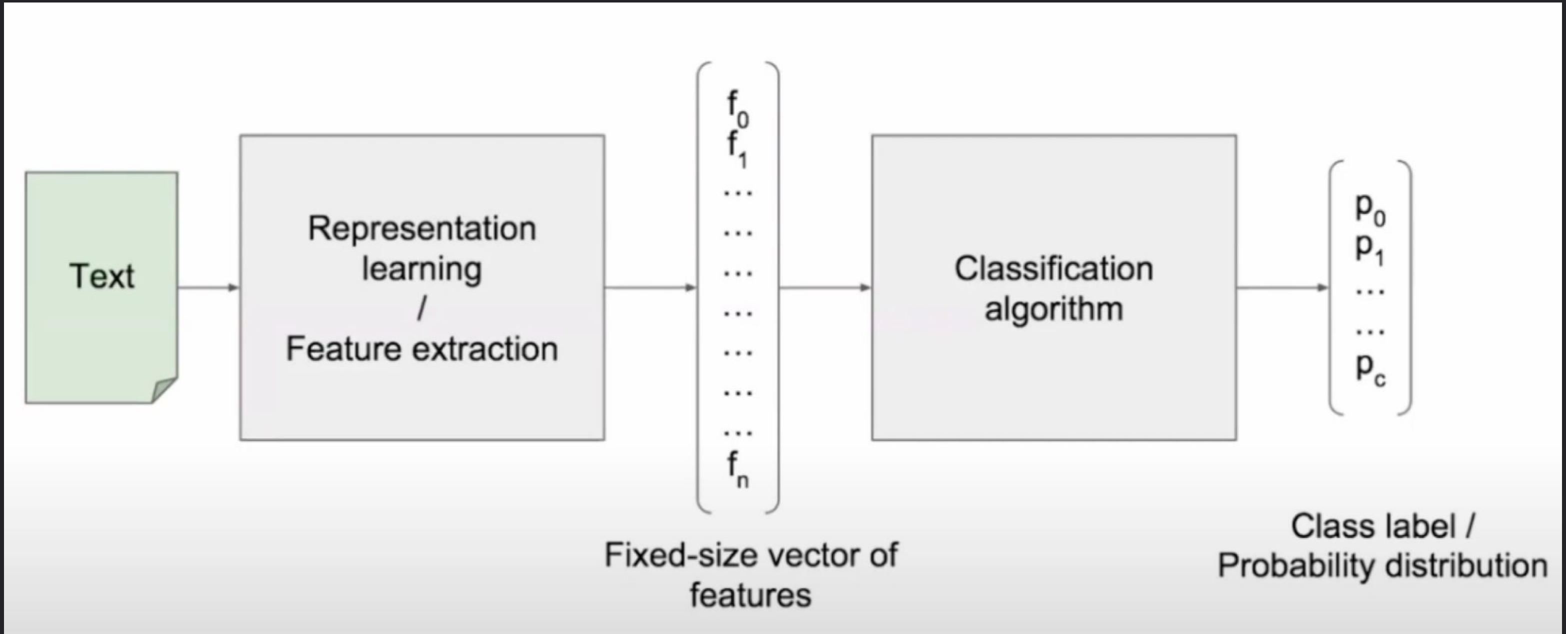
Stages in NLP

1. Rule-based methods (1950-1990s)
2. Statistical approaches (1990-2010s)
3. Deep Learning approaches (2010s-present)

Tasks in NLP

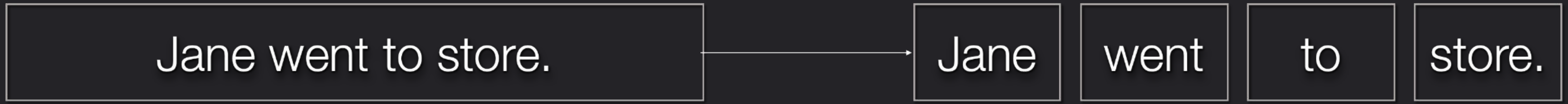

NLP
NLU
named entity recognition (NER)
part-of-speech tagging (POS)
text categorization
syntactic parsing
coreference resolution
machine translation
relation extraction
semantic parsing
question answering (QA)
paraphrase & natural language inference
sentiment analysis
dialogue agents
summarization

Common pipeline



Tokenization

Tokenization - splitting text into tokens.



Properties of word vectors

- fixed size
- contains the meaning of the word
- formed automatically

Bag of Words (BOW)

Cat can meow.

Dog can bark.

	Cat	can	meow	Dog	bark
Sentence 1	1	1	1	0	0
Sentence 2	0	1	0	1	1

Bag of Words (BOW)

Cat can meow.

Dog can bark and can not meow.

	Cat	can	meow	Dog	bark	and	not
Sentence 1	1	1	1	0	0	0	0
Sentence 2	0	2	0	1	1	1	1

One hot encoding (OHE)

Cat can meow.

Dog can bark.

→

Cat = [1, 0, 0, 0, 0]

can = [0, 1, 0, 0, 0]

meow = [0, 0, 1, 0, 0]

Dog = [0, 0, 0, 1, 0]

bark = [0, 0, 0, 0, 1]

Distributional semantics

Do you now word *tezgüino*?

Count-based methods

1. A bottle of _____ is on the table.
2. Everybody likes _____.
3. Don`t have _____ before you drive.
4. We make _____ out of corn.

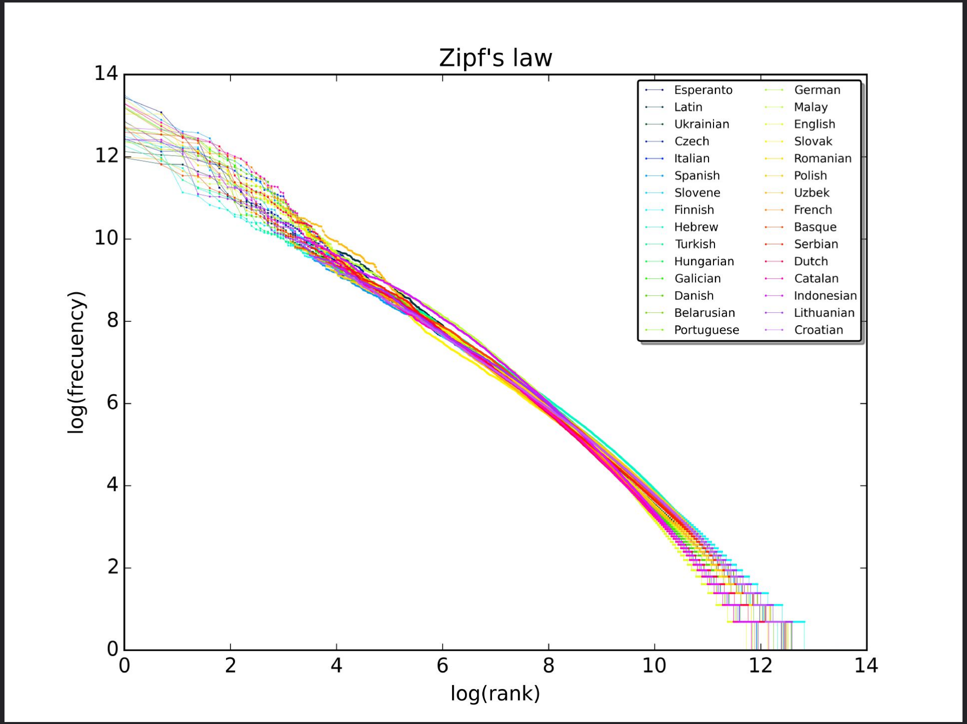
example from Jacob Eisenstein's NLP notes

Distributional semantics

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

J.Firth, 1957

Zipf law



A plot of the rank versus frequency for the first 10 million words in 30 Wikipedias (dumps from October 2015) in a log-log scale.

TF-IDF

TF-IDF = Term Frequency (TF) * Inverse Document Frequency (IDF)

- t — term (word)
- d — document (set of words)
- N — count of corpus
- corpus — the total document set

TF-IDF

tf(t,d) = count of t in d / number of words in d

df(t) = occurence of t in documents → idf(t) = N / df

idf(t) = log(N / (df + 1)) *

tf-idf(t,d) = tf(t,d) * log(N / (df + 1))

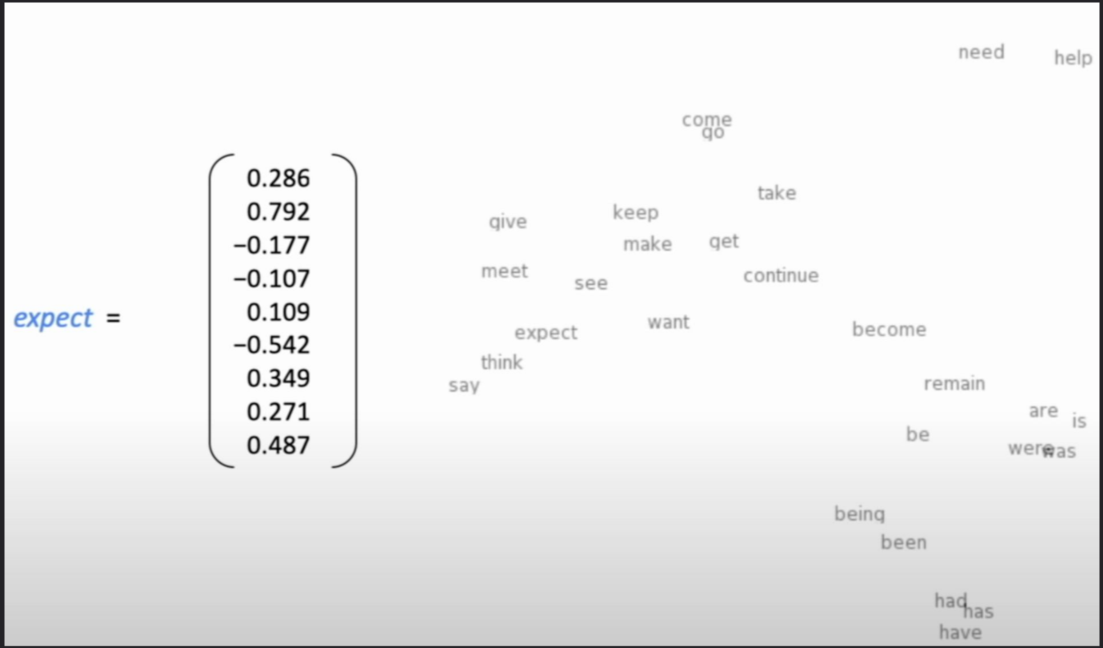
TF-IDF

A) The car is driven on the road.
B) The truck is driven on the highway.

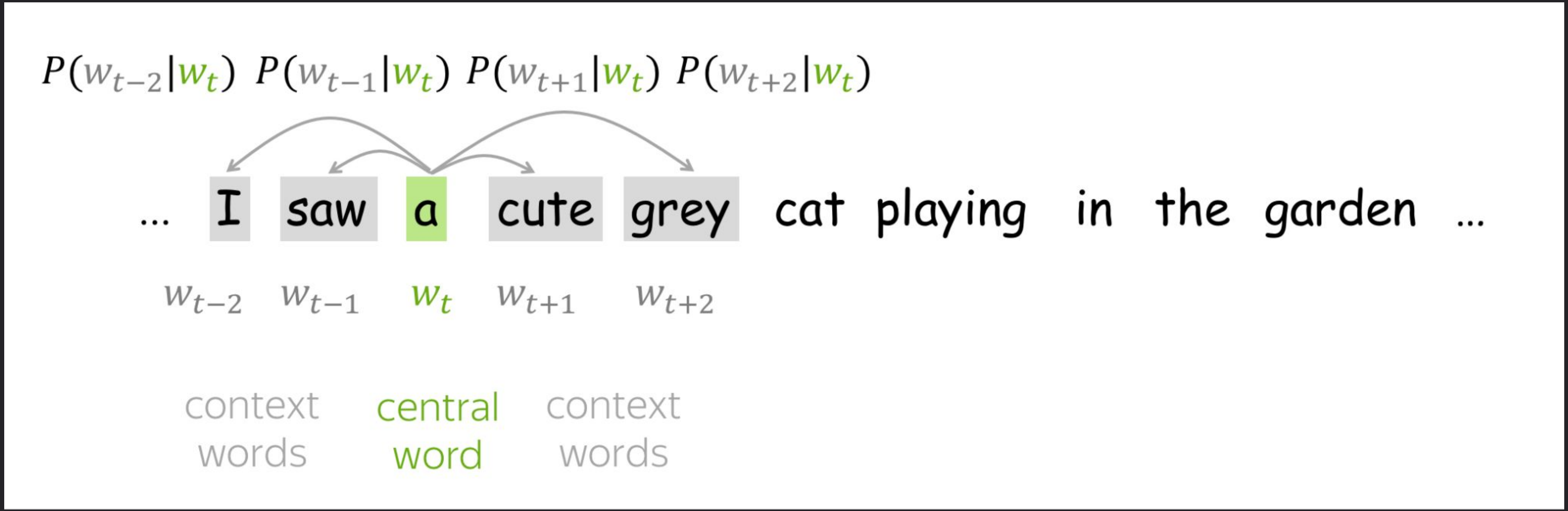
Word	TF(word,A)	TF(word,B)	IDF	TF*IDF(word, A)	TF*IDF(word,B)
the	1/7	1/7	log(2/2) = 0	0	0
car	1/7	0	log(2/1) = 0.3	0.043	0
is	1/7	1/7	log(2/2) = 0	0	0
driven	1/7	1/7	log(2/2) = 0	0	0
on	1/7	1/7	log(2/2) = 0	0	0
road	1/7	0	log(2/1) = 0.3	0.043	0
truck	0	1/7	log(2/1) = 0.3	0	0.043
highway	0	1/7	log(2/1) = 0.3	0	0.043

Word2vec

IDEA -> Put information about context into vector.

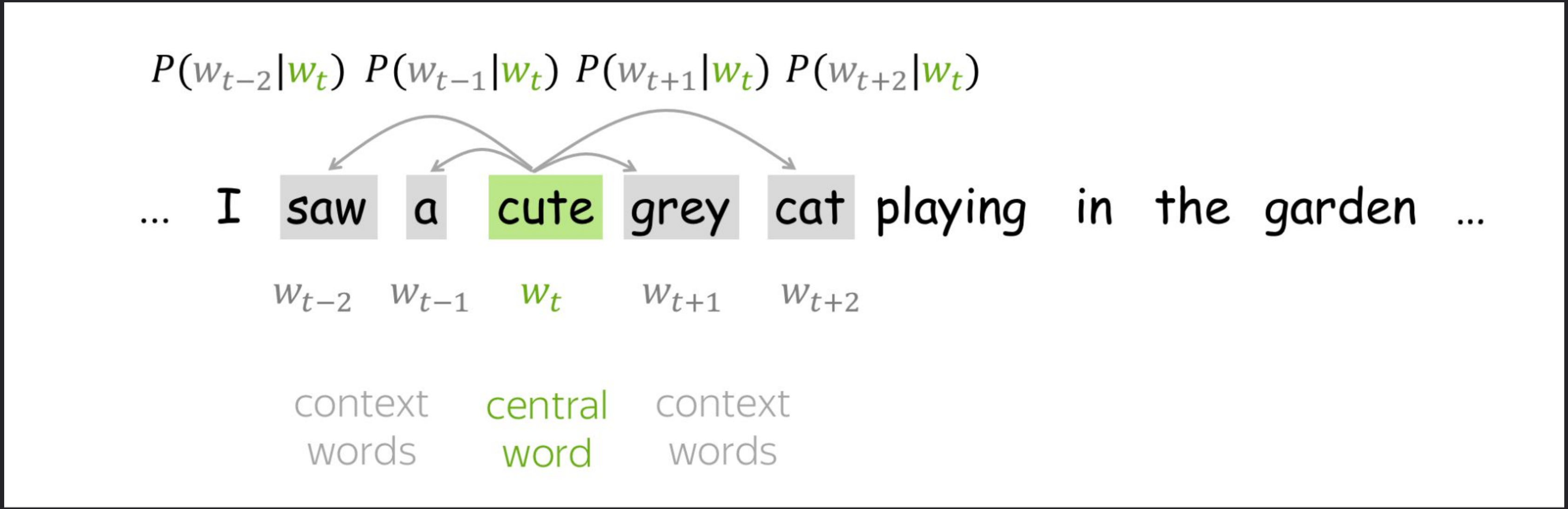


Word2vec



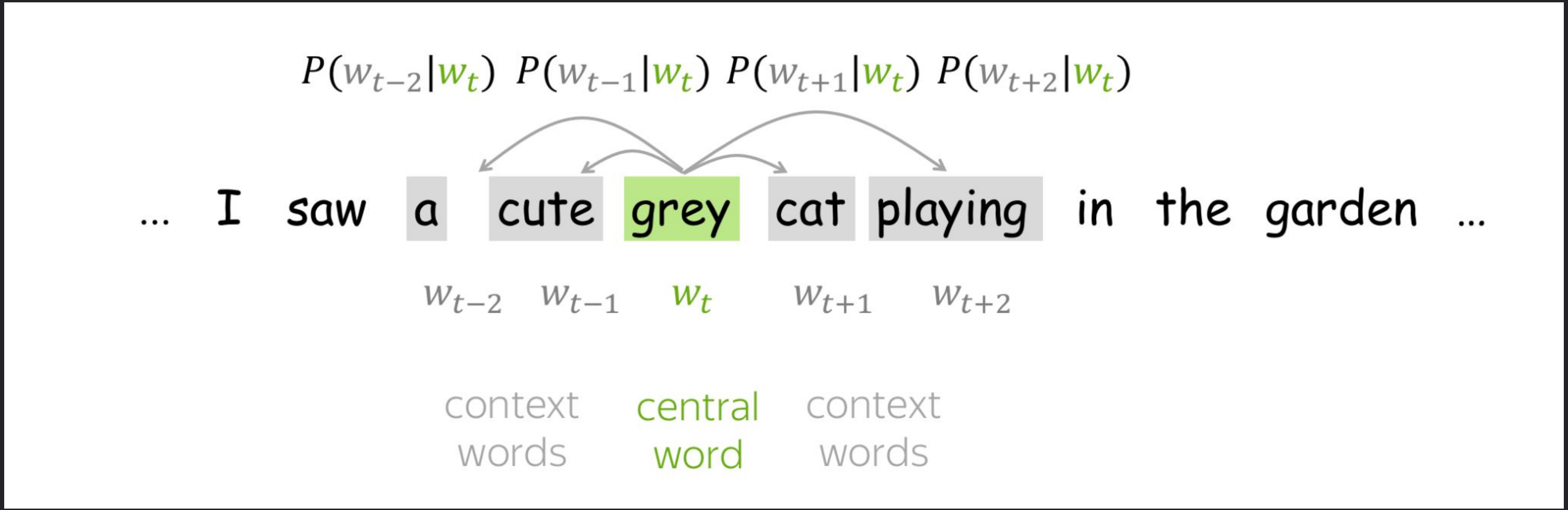
example from [Lena Voita NLP course]

Word2vec


$P(w_{t-2}|w_t)$ $P(w_{t-1}|w_t)$ $P(w_{t+1}|w_t)$ $P(w_{t+2}|w_t)$
... I saw a cute grey cat playing in the garden ...
w_{t-2} w_{t-1} w_t w_{t+1} w_{t+2}
context words
central word
context words

Introduction.
Word vector representation

Word2vec



Introduction.
Word vector representation

Word2vec