Objectives

Understand the applications of NLP

Application

NLP

Understand the basic knowledge of Natural Language Processing (NLP)

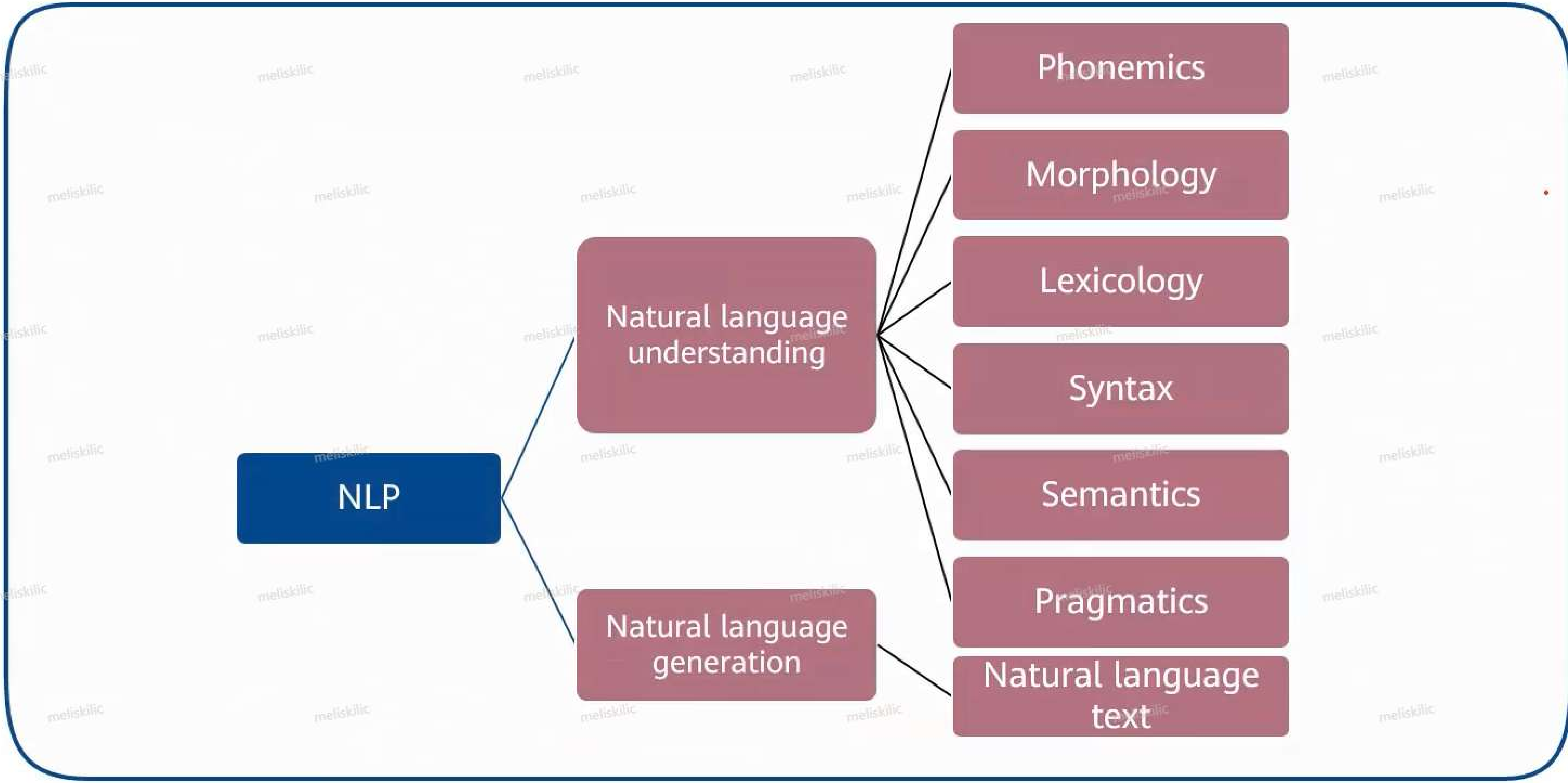
Master the key tasks of NLP

Tasks

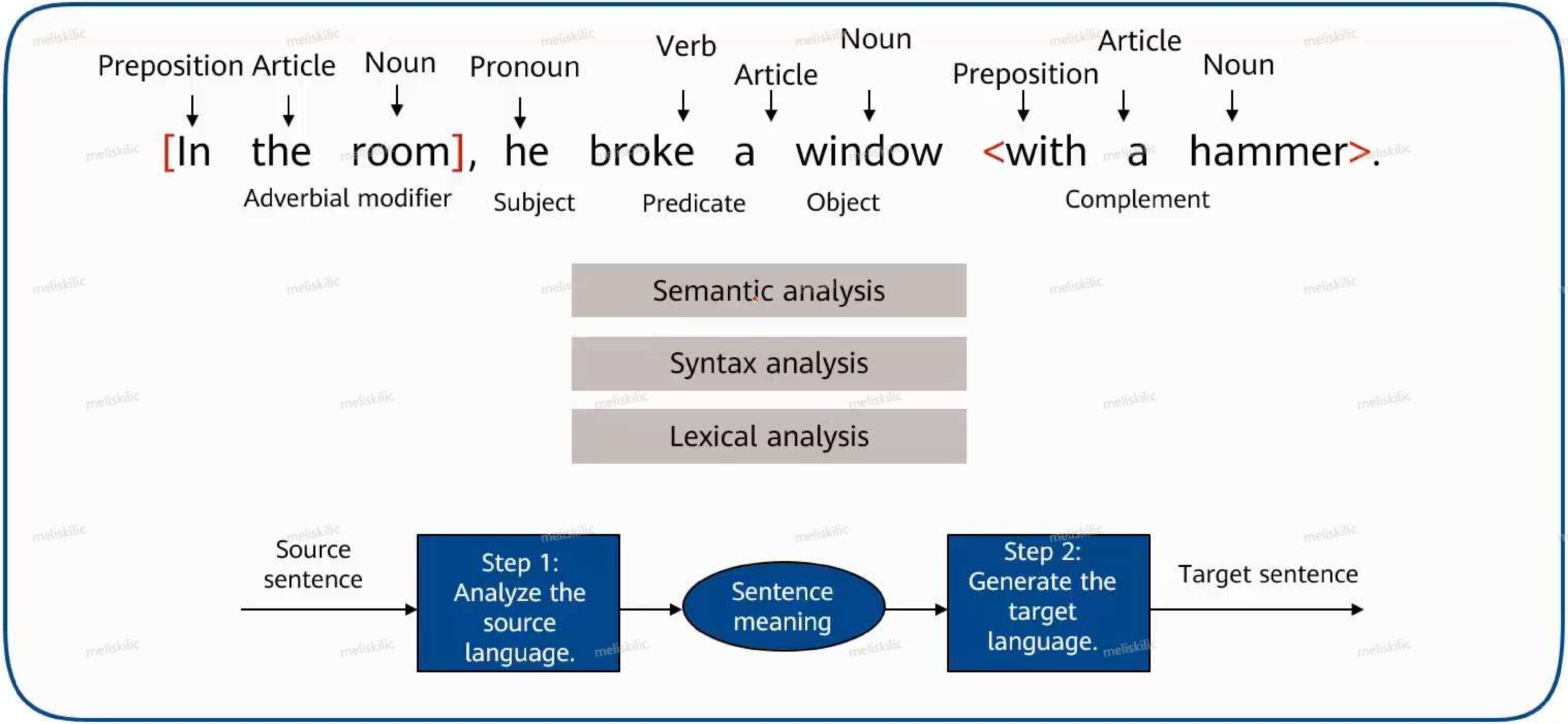
Models

Master the algorithms of the Recurrent Neural Network (RNN)

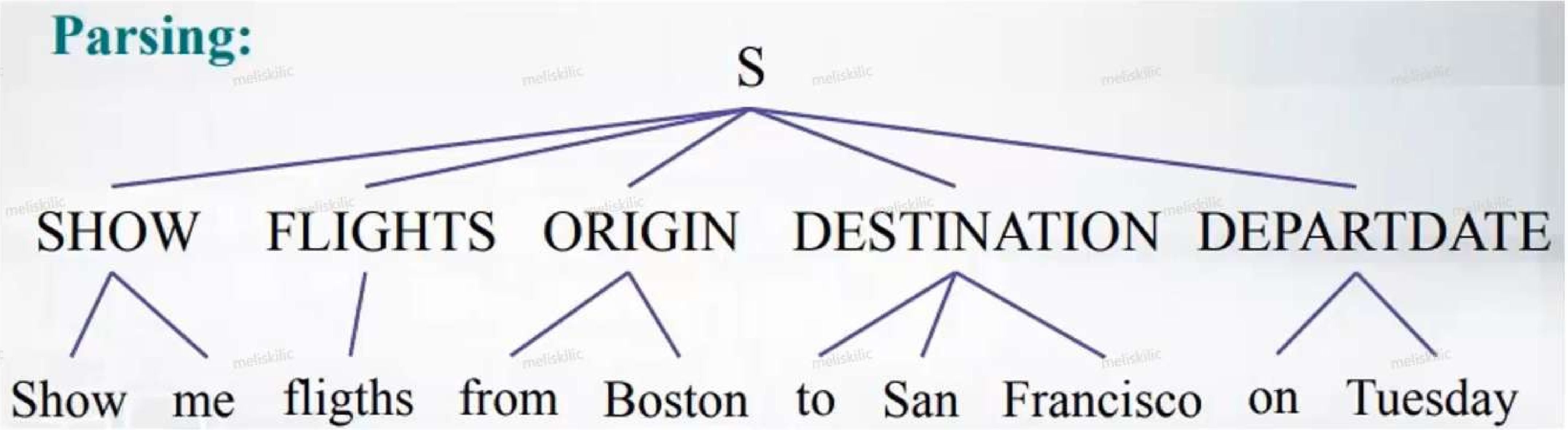
NLP Research Direction

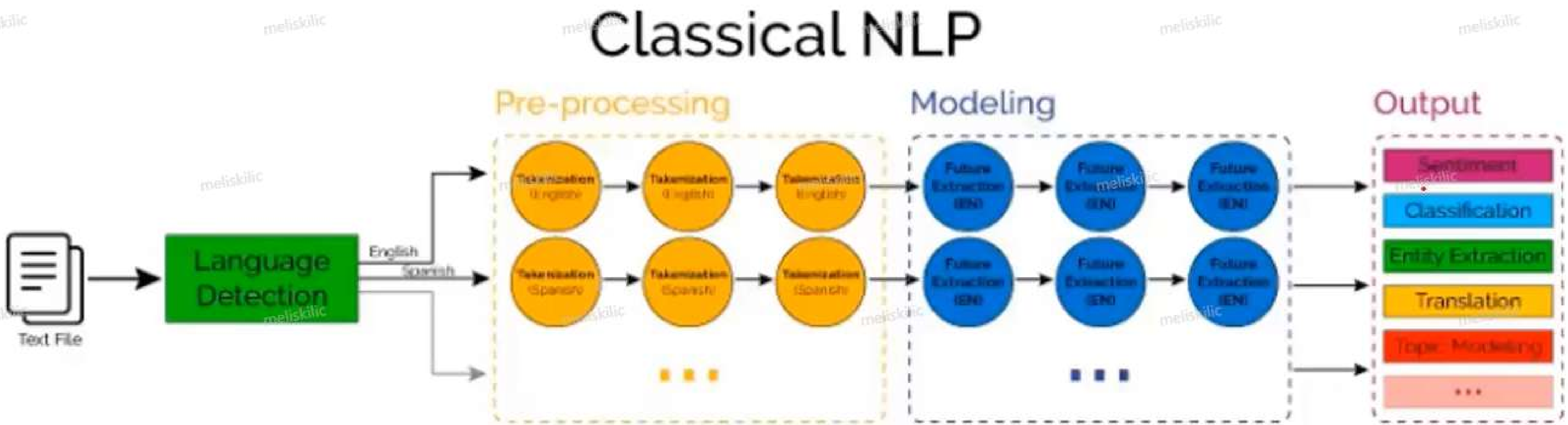


Three Levels of NLP



Basic Methods of NLP (1)



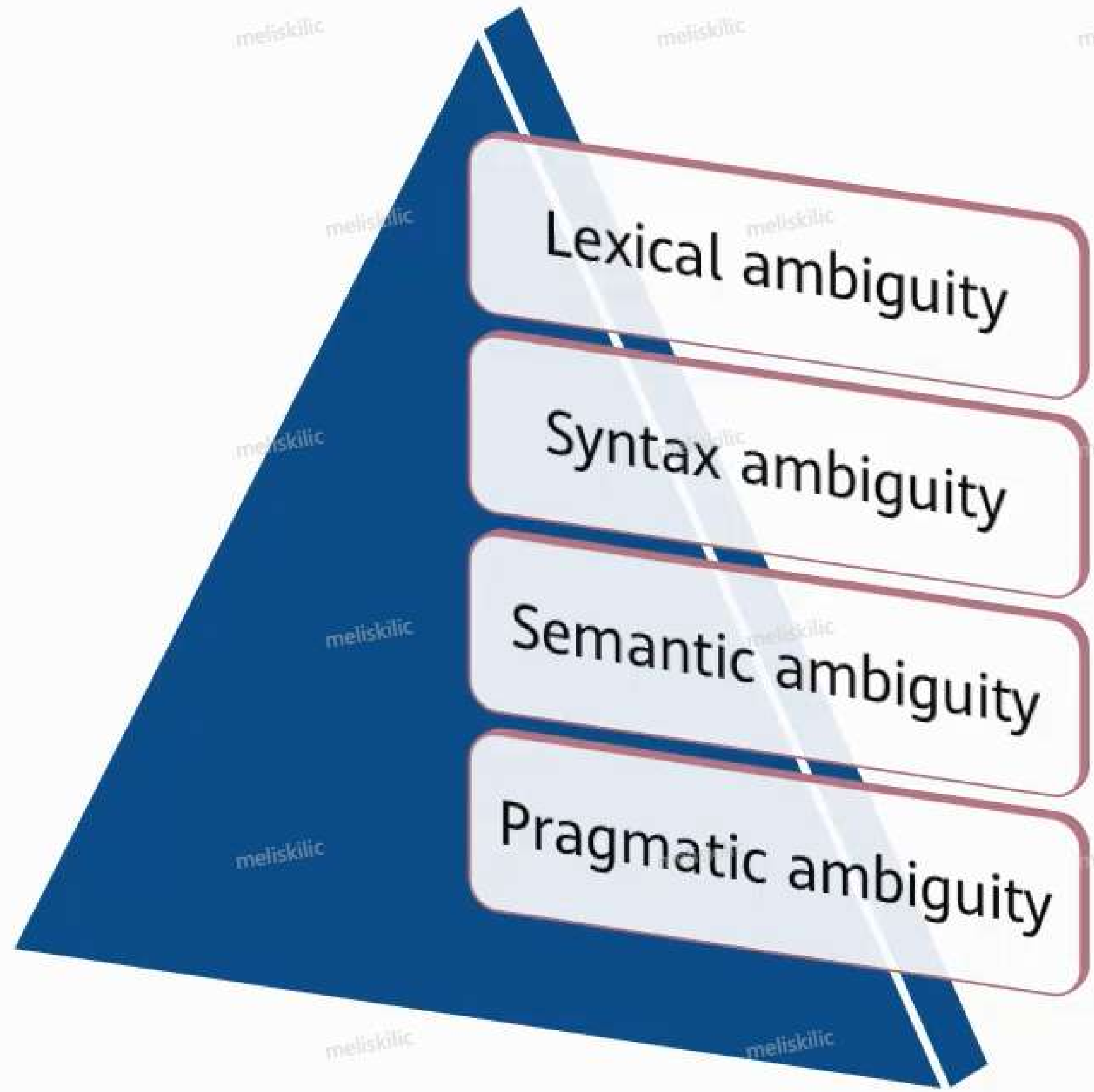


Basic Methods of NLP (3)

Application model

- base on different languages processing applications.
- to learn complex and extensive language structures, and uses methods
- The "empirical" language model
- Modeling steps:
  - obtain statistics
  - calculate statistics of higher-level language units

Challenges in NLP



x y
this that

Lexical ambiguity

Word segmentation

- English is easier to segment than other languages.

Part-of-speech tagging

- I plan/v to take the postgraduate
- I have completed the plan/n

Named entity recognition

- Apple

Syntax ambiguity

John likes Jane more than Adam.

- Compared to Adam and Jane, John likes Jane more.
- Compared to Adam's liking of Jane, John likes Jane more.

Alex saw a man on the hill with a telescope.

The government asks us to save soap and waste paper.

Semantic ambiguity

At last, a computer that understands you like your mother.

- Meaning 1: A computer understands you as your mother does.
- Meaning 2: A computer understands that you like your mother.
- Meaning 3: A computer understands you as it understands your mother.

Meredith is in a terrible state.

- “state”: condition of something.
- “state”: a country or part of a country.

Pragmatic ambiguity

"You are so bad"

- When this sentence is said to an adult who has done bad things
- When a mother says it to her naughty son
- When a girl in love says it to her boyfriend

Development Status

- A number of influential language databases have been developed
- Corpus of Peking University and HowNet
- Many new research directions merge
  - Reading comprehension
  - image (video) understanding
  - simultaneous interpretation of speech
- Unresolved
  - Problems of unregistered-word recognition, ambiguity elimination, and semantic understanding
  - Lack of a complete and systematic theoretical framework

What is a Language Model

A language model is an abstract correspondence established based on objective language facts.

Problems

- Spelling correction: P(about fifteen minutes from) > P(about fifteenminuets from)
- Question answering system, ...
- The above questions can be expressed as follows according to the chain rule:

$$P(w_1, w_2, \ldots, w_m) = P(w_1)P(w_2|w_1) \ldots P(w_i|w_1, w_2, \ldots, w_{i-1}) \ldots P(w_m|w_1, w_2, \ldots, w_{m-1})$$

N - gram Language Model

When N-gram model is used to estimate the conditional probability, the preceding words at a distance greater than or equal to n is ignored.

Therefore, the conditional probability can be calculated from frequency counts:

$$P(w_i|w_1, w_2, \dots, w_{i-1}) \approx P\left(w_i \middle| w_{i-(n-1)}, w_{i-(n-2)}, \dots, w_{i-1}\right) = \frac{P(w_{i-(n-1)}, w_{i-(n-2)}, \dots, w_{i-1}, w_i)}{P(w_{i-(n-1)}, w_{i-(n-2)}, \dots, w_{i-1})}$$

$$= \frac{count(w_{i-(n-1)}, w_{i-(n-2)}, \dots, w_{i-1}, w_i)}{count(w_{i-(n-1)}, w_{i-(n-2)}, \dots, w_{i-1})}$$

N - gram Language Model

Unigram model:

$$P(w_1, w_2, \dots, w_m) = P(w_1)P(w_2)\dots P(w_m)$$

Bigram model:

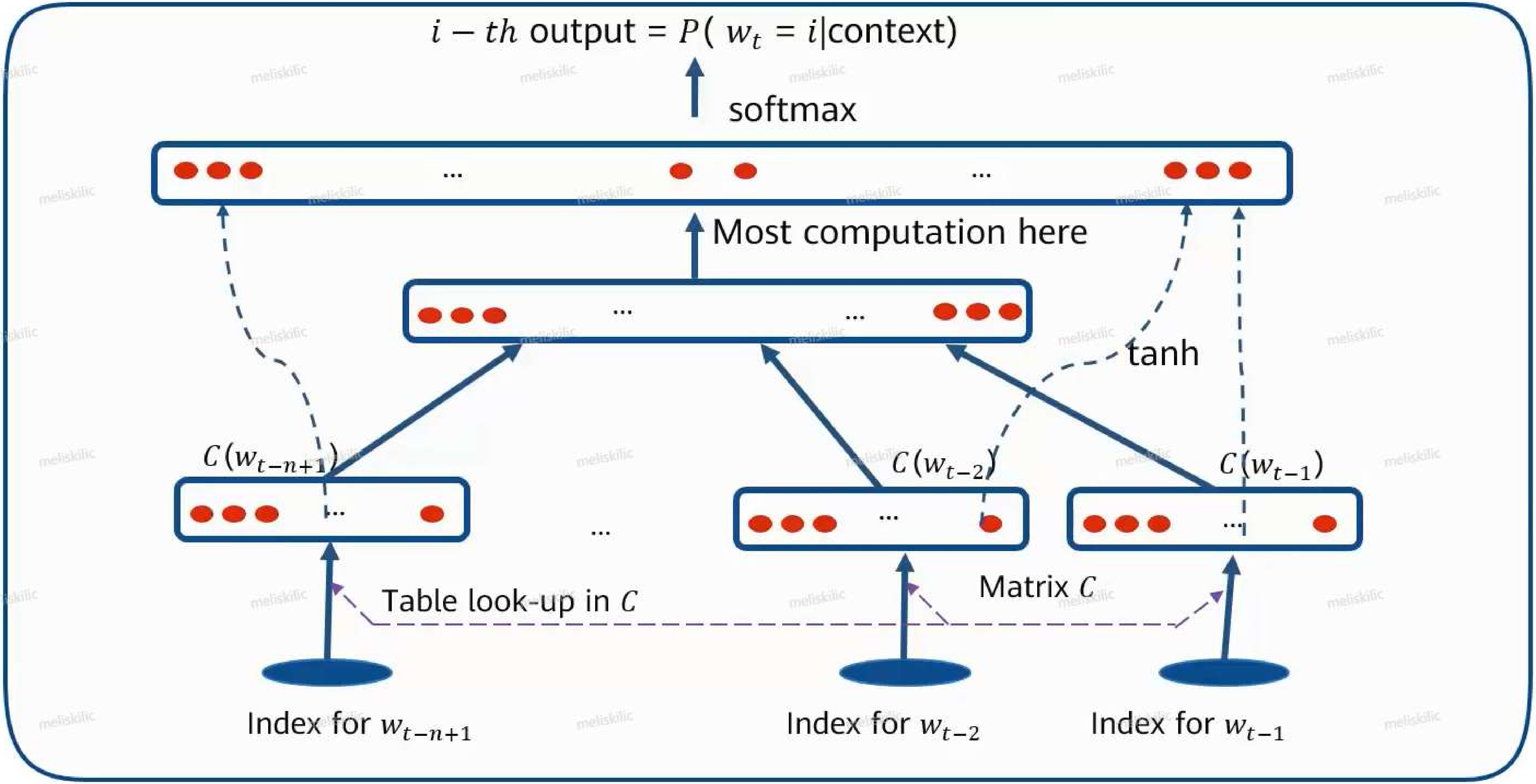
$$P(w_1, w_2, \dots, w_m) = P(w_1)P(w_2|w_1)\dots P(w_m|w_{m-1})$$

For example

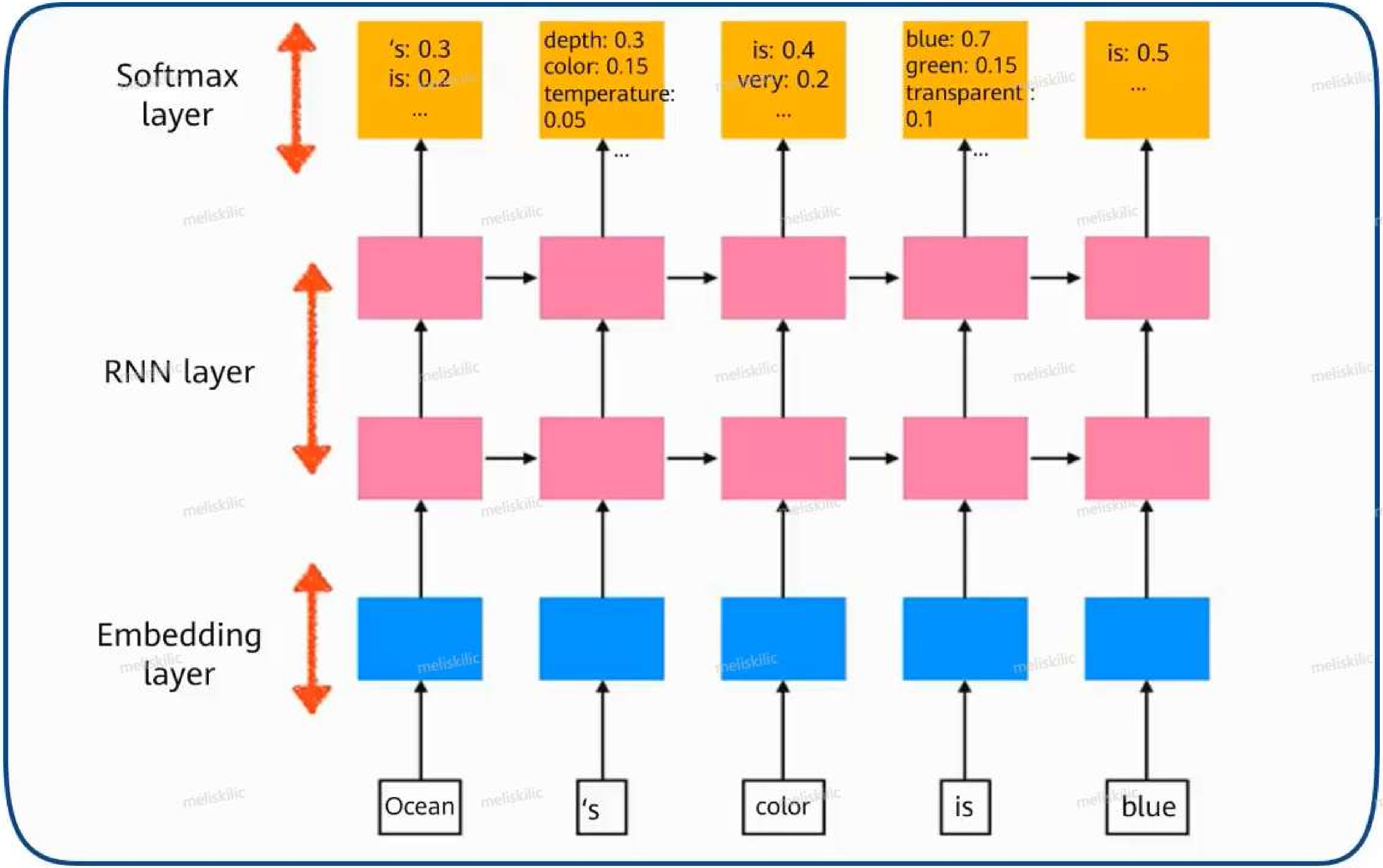
<s> I am Lily </s>
<s> Lily I am </s>
<s> I do not like green eggs and ham</s>

p(I | <s>) = 2/3 = 0.667
p(am | I) = 2/3 = 0.667
p(</s> | Lily) = 1/2 = 0.5
...

Neural Network Language Model (1)



Neural Network Language Model (2)



Comparison

Similarity:

A sentence as a word sequence

Differences:

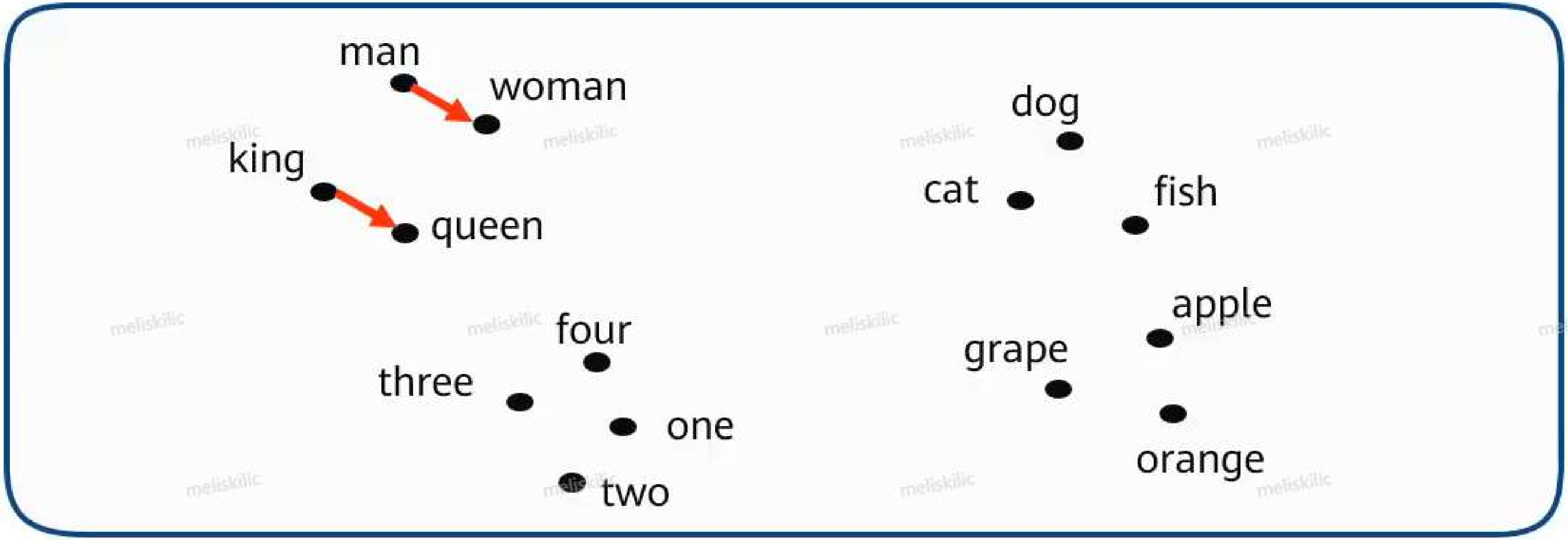
- Manner of probability calculation:
  - N-gram model：Only the first n words
  - NNLM： The context of the whole sentence.
- Manner of model training:
  - N-gram：maximum likelihood estimation
  - NNLM：RNN optimization method
- RNNs can store context information of any length in a hidden state, not subject to the window limit in the N-gram model.

Features in NLP

- Directly observable features
  - Word features
    - Prefixes\Suffixes: un, dis, er, sion...
    - Capitalization: Jane, Huawei, UN...
  - Context features
- Inferred linguistic features
- Core features and combination features
- N-gram features
- Distribution features

Text Vectorization (2)

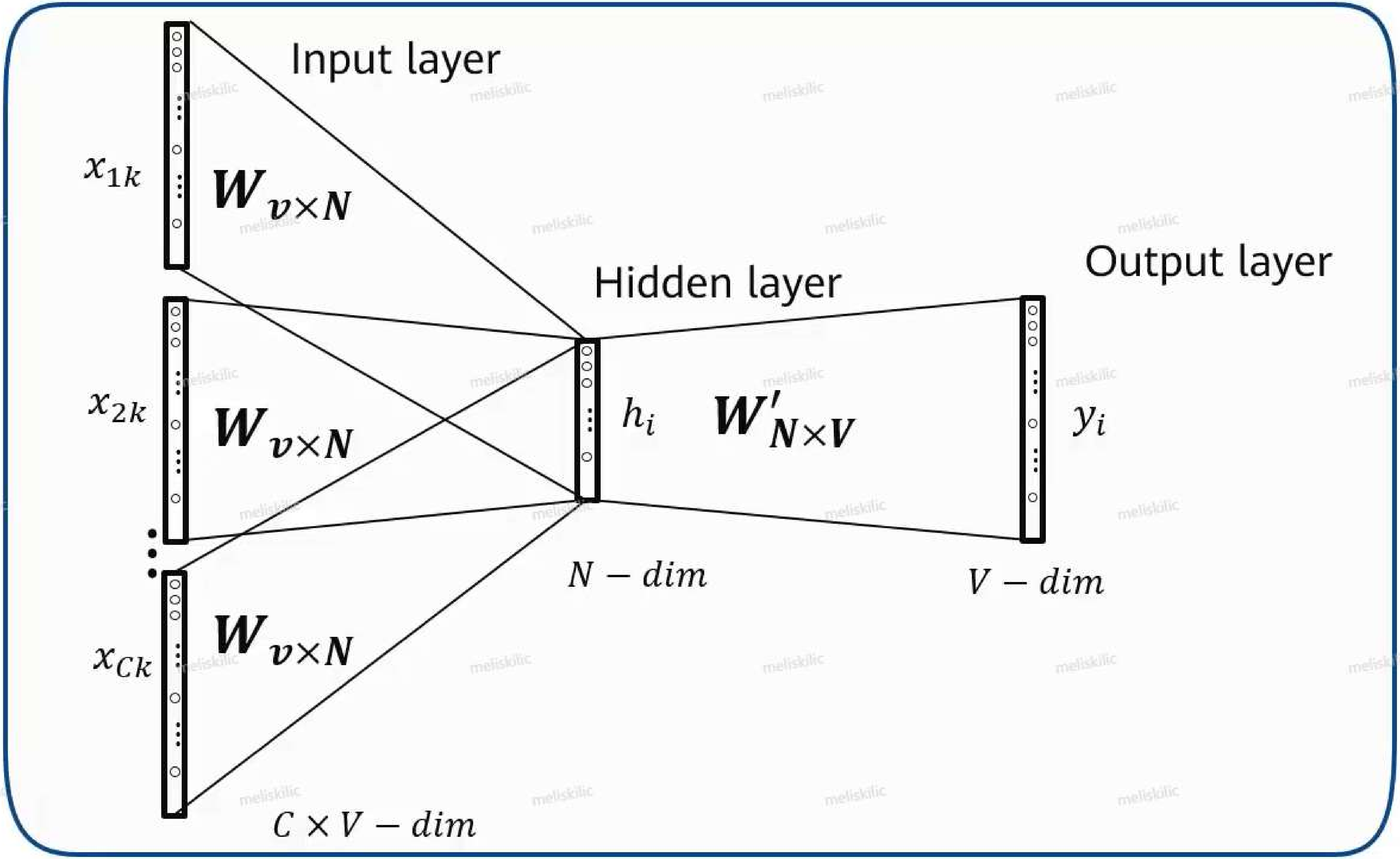
	Man (5391)	Woman (9853)	King (4914)	Queen (7157)	Apple (456)	Orange (6257)
Gender	-1	1	-0.95	0.97	0.00	0.01
Royal	0.01	0.02	0.93	0.95	-0.01	0.00
Age	0.03	0.02	0.70	0.69	0.03	-0.02
Food	0.09	0.01	0.02	0.01	0.95	0.97



Text Vectorization (1)

- word2vec
  - CBOW model
  - Skip-gram model
- doc2vec/str2vec
  - Distributed Memory (DM)
  - Distributed Bag of Words (DBOW)

word2vec - CBOW Model

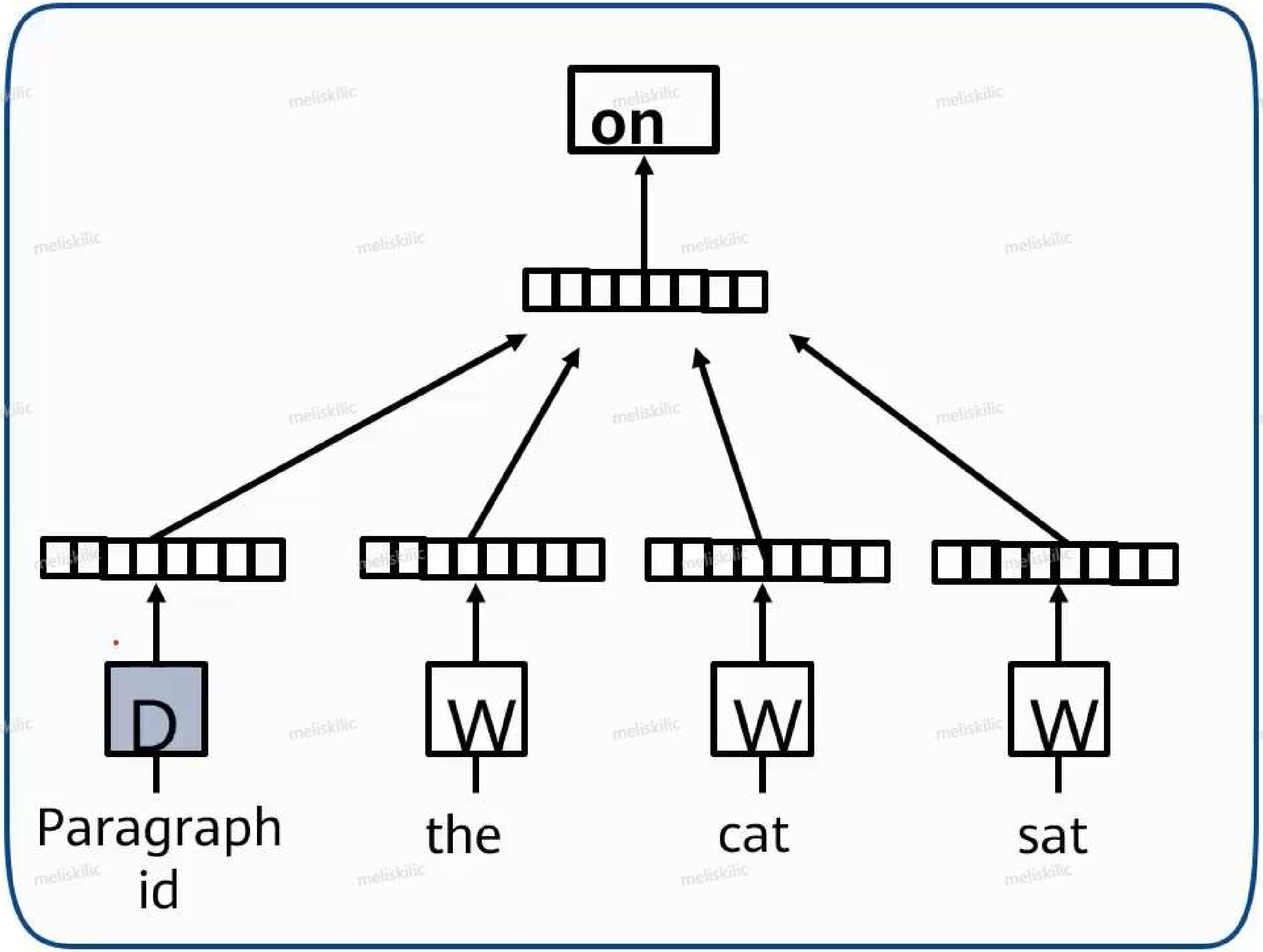


doc2vec - DM Model

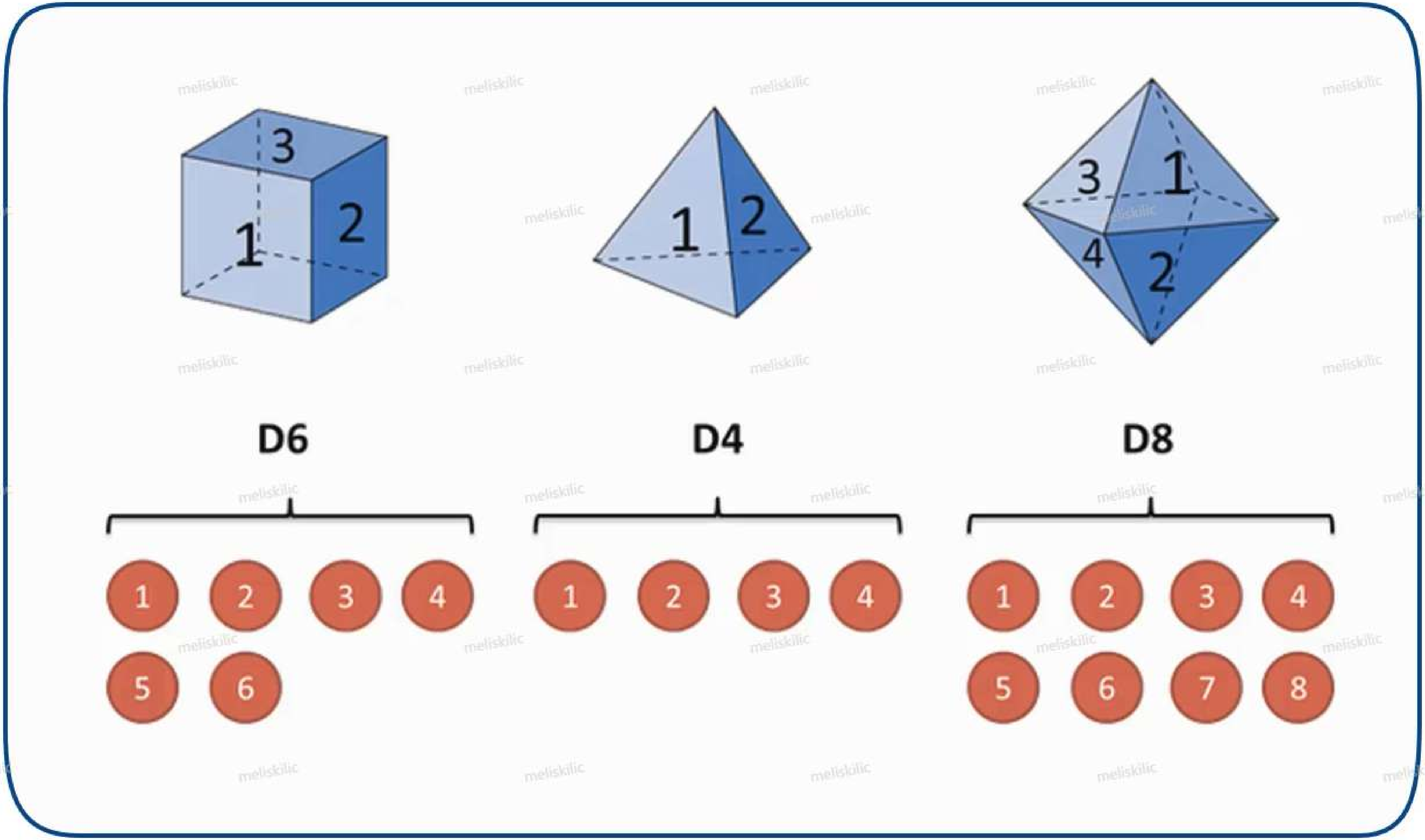
Classifier

Average/Concatenate

Paragraph Matrix

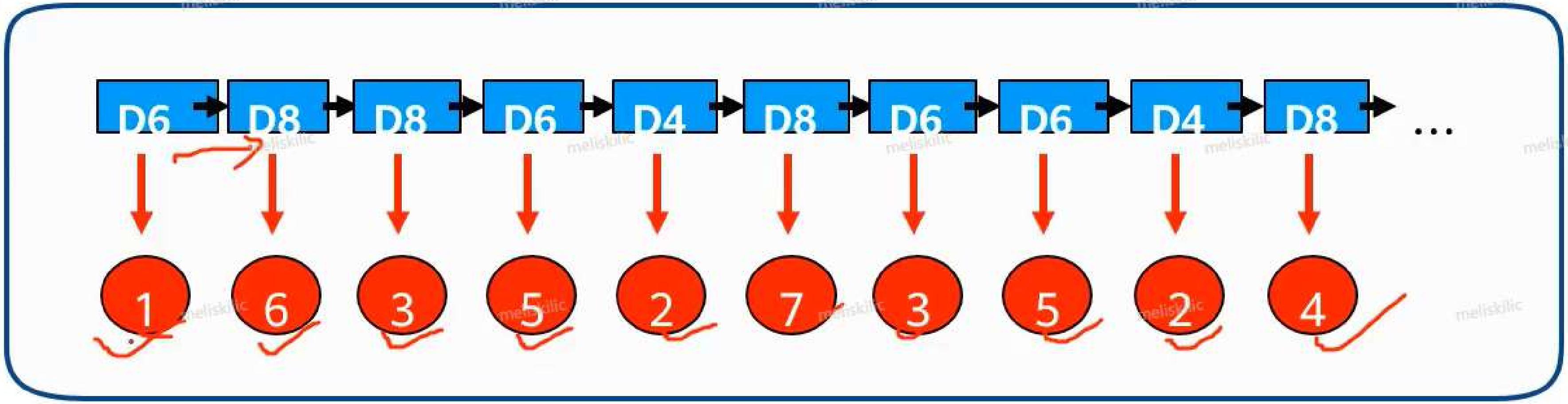


HMM (1)



HMM (2)

Schematic diagram of the Hidden Markov Model (HMM)



D6 A hidden state

From a hidden state to the next hidden state

1 An observed state

Output from a hidden state to an observed state

HMM (3)

$$maxP(w) = P(w_1, w_2, \ldots, w_m) = P(w_1)P(w_2|w_1) \ldots P(w_i|w_1, w_2, \ldots, w_{i-1}) \ldots P(w_m|w_1, w_2, \ldots, w_{m-1})$$

Add hidden variable h

$$maxP(h|w)$$

Bayesian formula

$$max\frac{P(w|h)P(h)}{P(w)}$$

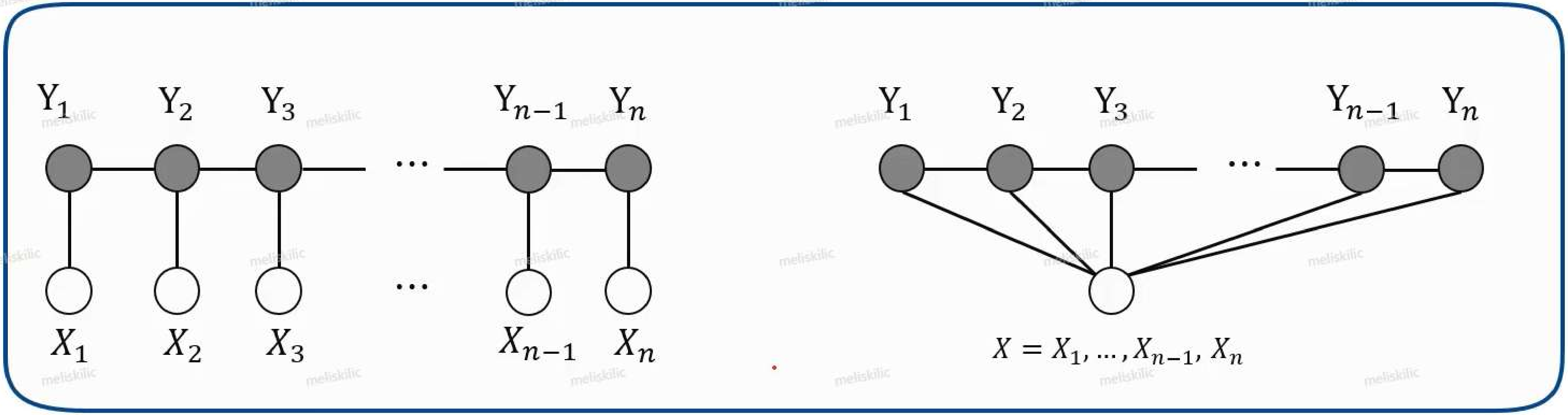
Constant $P(w)$

$$maxP(w|h)P(h)$$

Observation independence hypothesis, chain rule

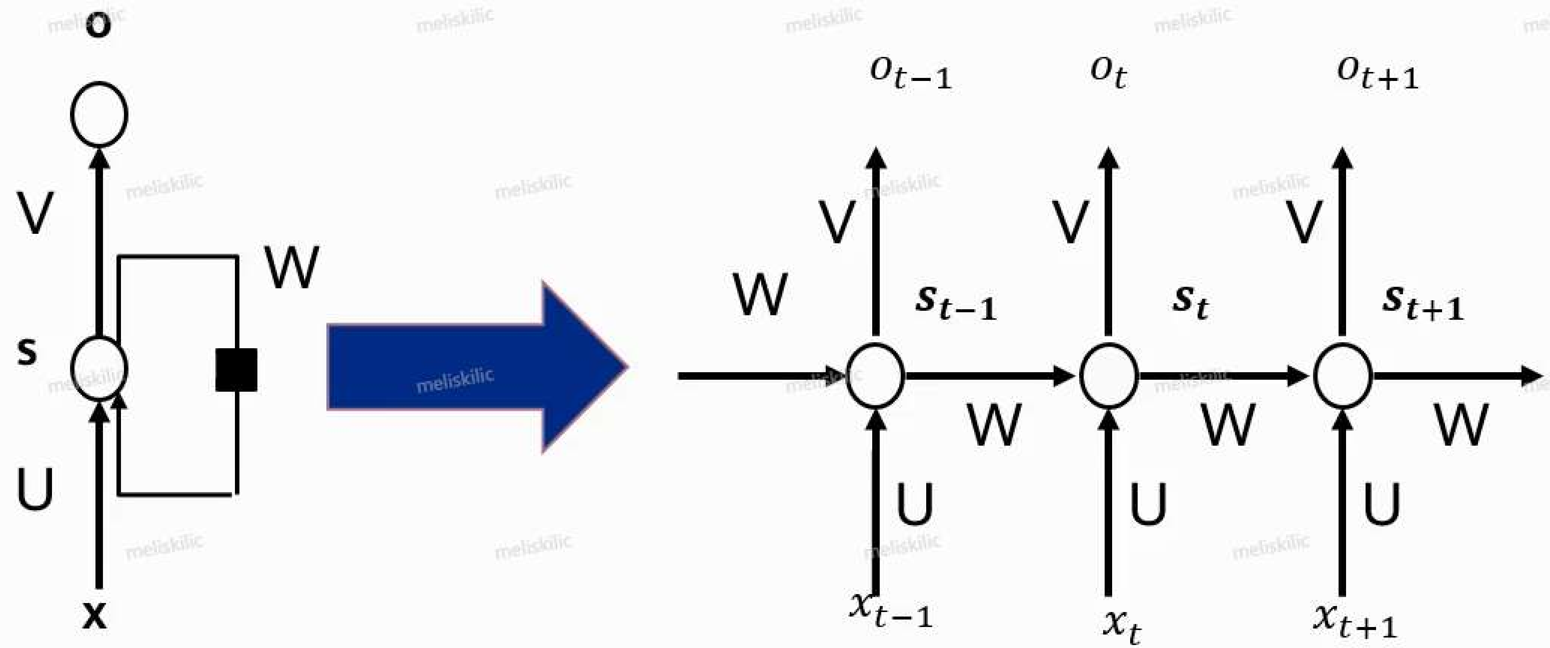
$$maxP(w_1|h_1)P(w_2|h_2) \ldots P(w_n|h_n)P(h_1)P(h_2|h_1)P(h_3|h_1, h_2) \ldots P(h_n|h_1, h_2, \ldots, h_{n-1})$$

Conditional Random Field

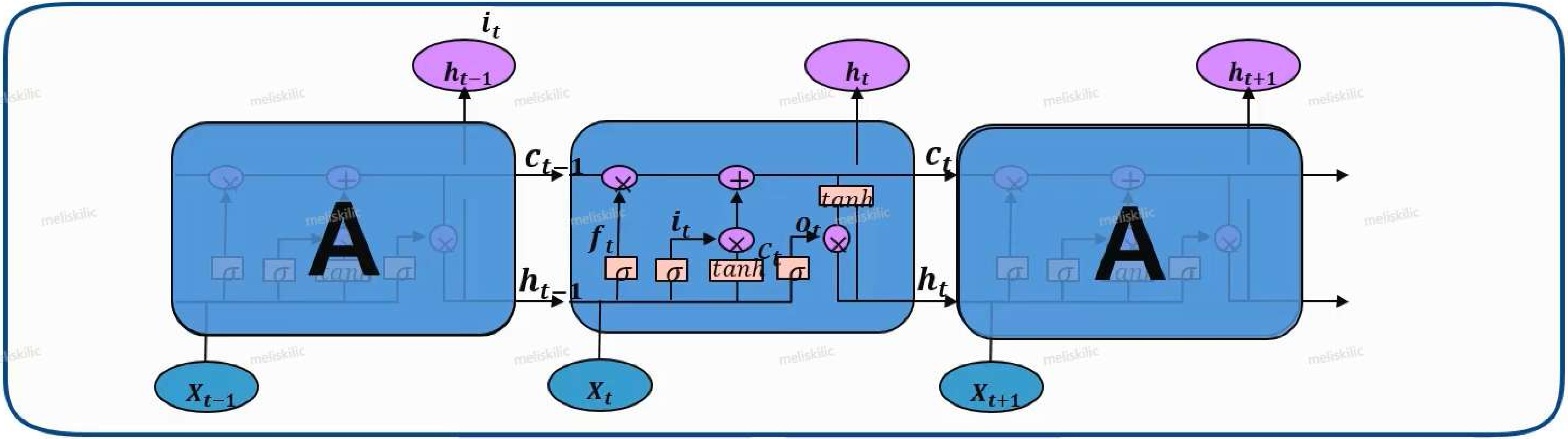


Linear chain CRF

RNN



LSTM



Colah, 2015, Understanding LSTMs Networks

$$i_t = \sigma(W^{(i)}x_t + U^{(i)}h_{t-1}) \quad \text{(Input gate)}$$

$$f_t = \sigma(W^{(f)}x_t + U^{(f)}h_{t-1}) \quad \text{(Forget gate)}$$

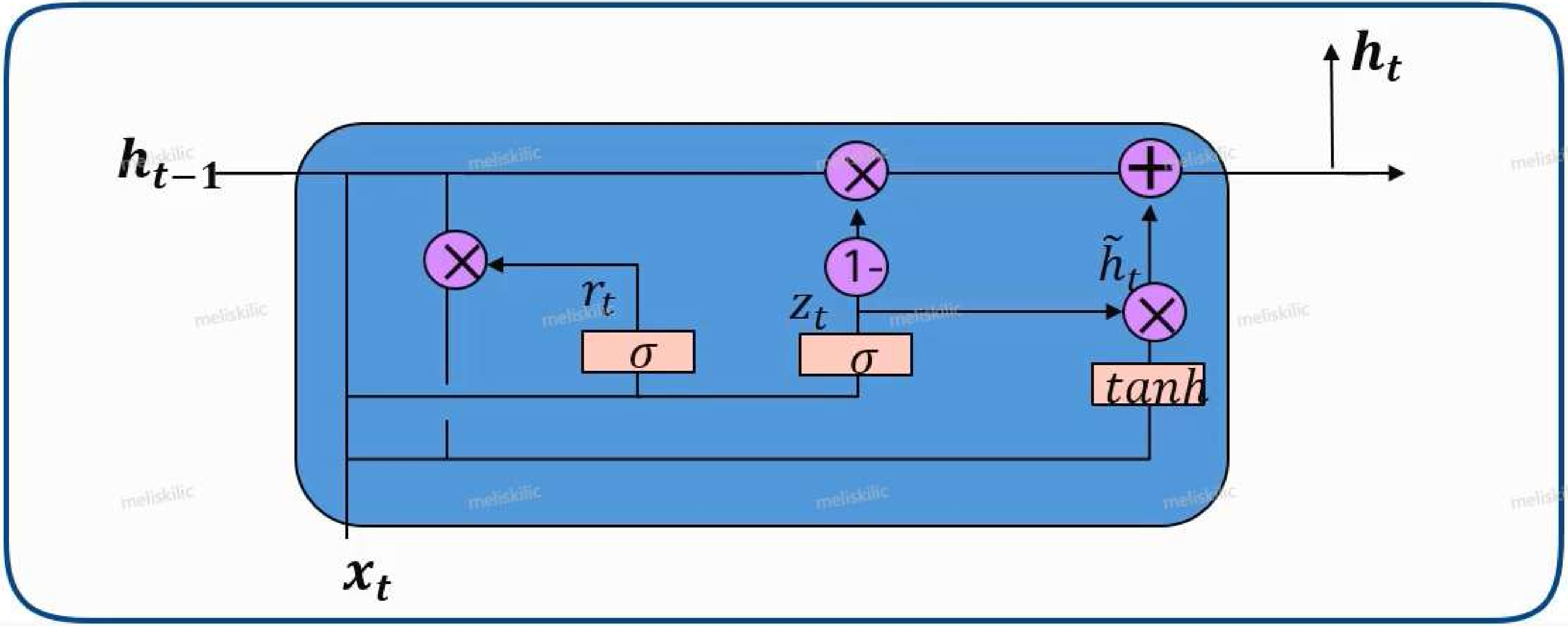
$$o_t = \sigma(W^{(o)}x_t + U^{(o)}h_{t-1}) \quad \text{(Output/Exposure gate)}$$

$$\tilde{c}_t = \tanh(W^{(c)}x_t + U^{(c)}h_{t-1}) \quad \text{(New memory cell)}$$

$$c_t = f_t \circ c_{t-1} + i_t \circ \tilde{c}_t \quad \text{(Final memory cell)}$$

$$h_t = o_t \circ \tanh(c_t)$$

GRU



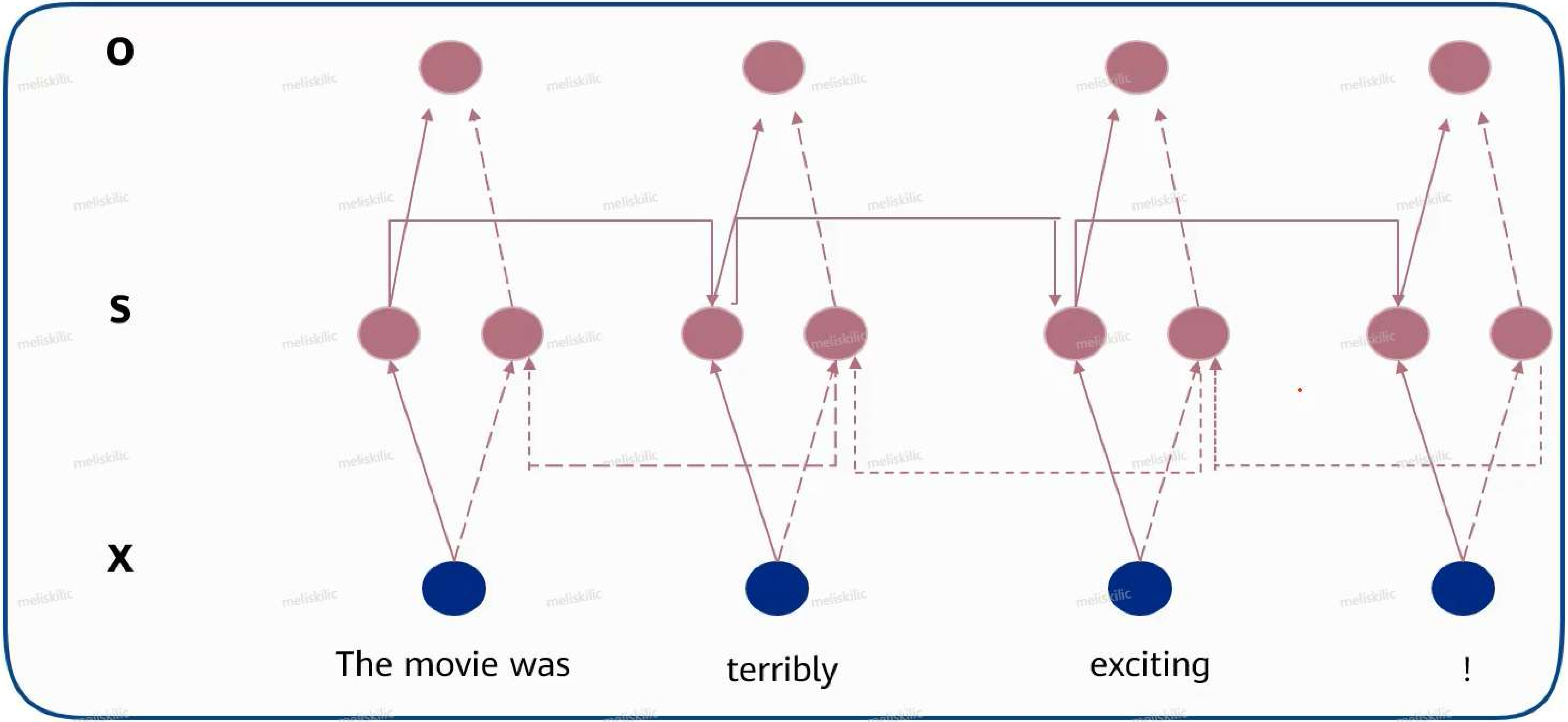
$$z_t = \sigma\left(W_z \cdot [h_{t-1}, x_t]\right)$$

$$r_t = \sigma\left(W_r \cdot [h_{t-1}, x_t]\right)$$

$$\tilde{h}_t = \tanh\left(W \cdot [r_t * h_{t-1}, x_t]\right)$$

$$h_t = (1 - z_t) * h_{t-1} + z_t * \tilde{h}_t$$

BiRNN



Part-of-Speech Tagging

Part-of-speech tagging

process of tagging a correct part of speech

For example:

They refuse to permit us to obtain the refuse permit.

Part of speech: a basic syntax attribute of a word

Purpose

Methods:

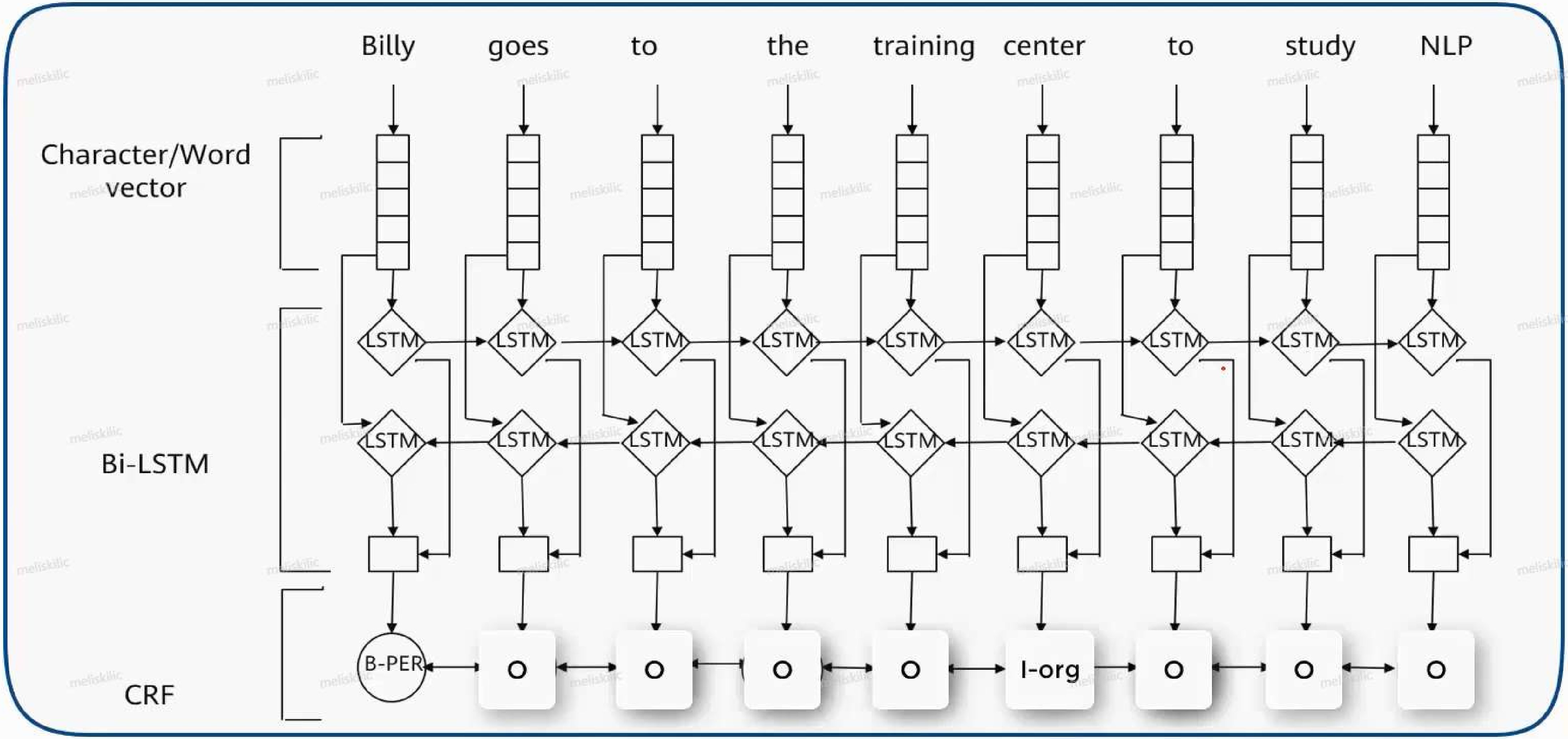
- Rule-based
- Statistics-based
- deep learning-based methods

Named Entity Recognition

Named Entities Recognition (NER):

- For example:
  - metallurgy/n, ministry of industry/n, Hongkong/n, fireproofing material/l, and research institute/n
- Classification:
  - categories (entity, time, and number)
  - seven subcategories (person name, place name, institution name, time, date, currency, and percent).
- Steps:
  - Recognize the entity boundary.
  - Determine the entity category (such as person name, place name, or institution name).

Deep Learning NER



Difficulties

- There are a large number of various named entities.
- The composition of named entities is complex.
- Entities are embedded and complex.
- The entity length is uncertain.

TF - IDF Algorithm (1)

For example:

On the World Blood Donor Day, school groups and blood donation service volunteers can go to the blood center to visit the inspection process. We will publicize the test results, and the price of blood will also be publicized.

TF - IDF Algorithm (2)

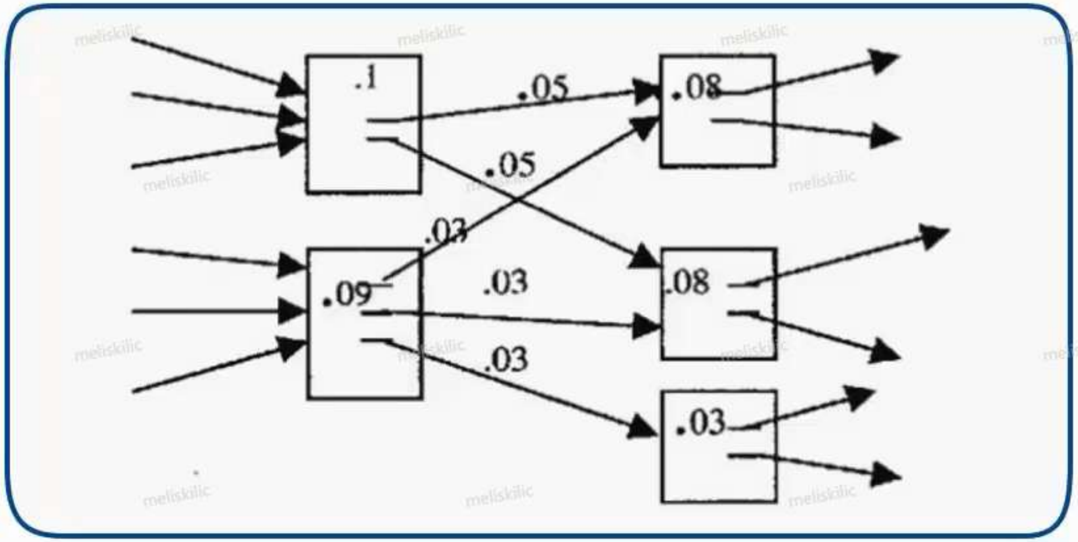
$$tf_{ij} = \frac{n_{ij}}{\sum_k n_{kj}} = \frac{\text{Number of times that a word appears in the document}}{\text{Number of total words in the document}}$$

$$idf_i = \log\left(\frac{|D|}{1 + |D_i|}\right)$$

TextRank Algorithm (1)

Google's PageRank algorithm

- Google founder Larry Page and Sergey Brin
- Evaluate the importance of a web page in the search system.
- Basic ideas:
  - Link quantity
  - Link quality



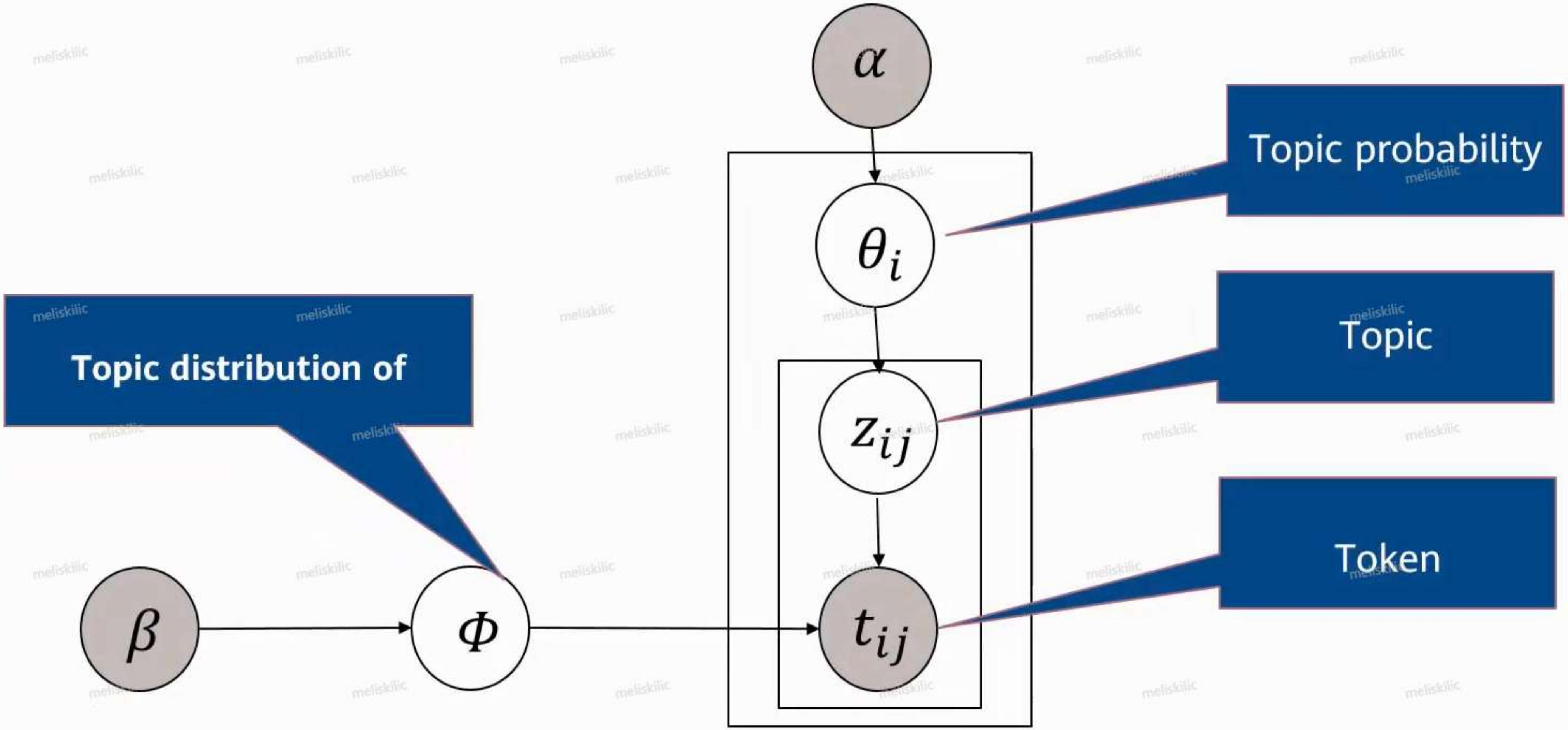
TextRank Algorithm (2)

$$S(V_i) = \sum_{j \in In(V_i)} \left(\frac{1}{|Out(V_j)|} \times S(V_j)\right)$$

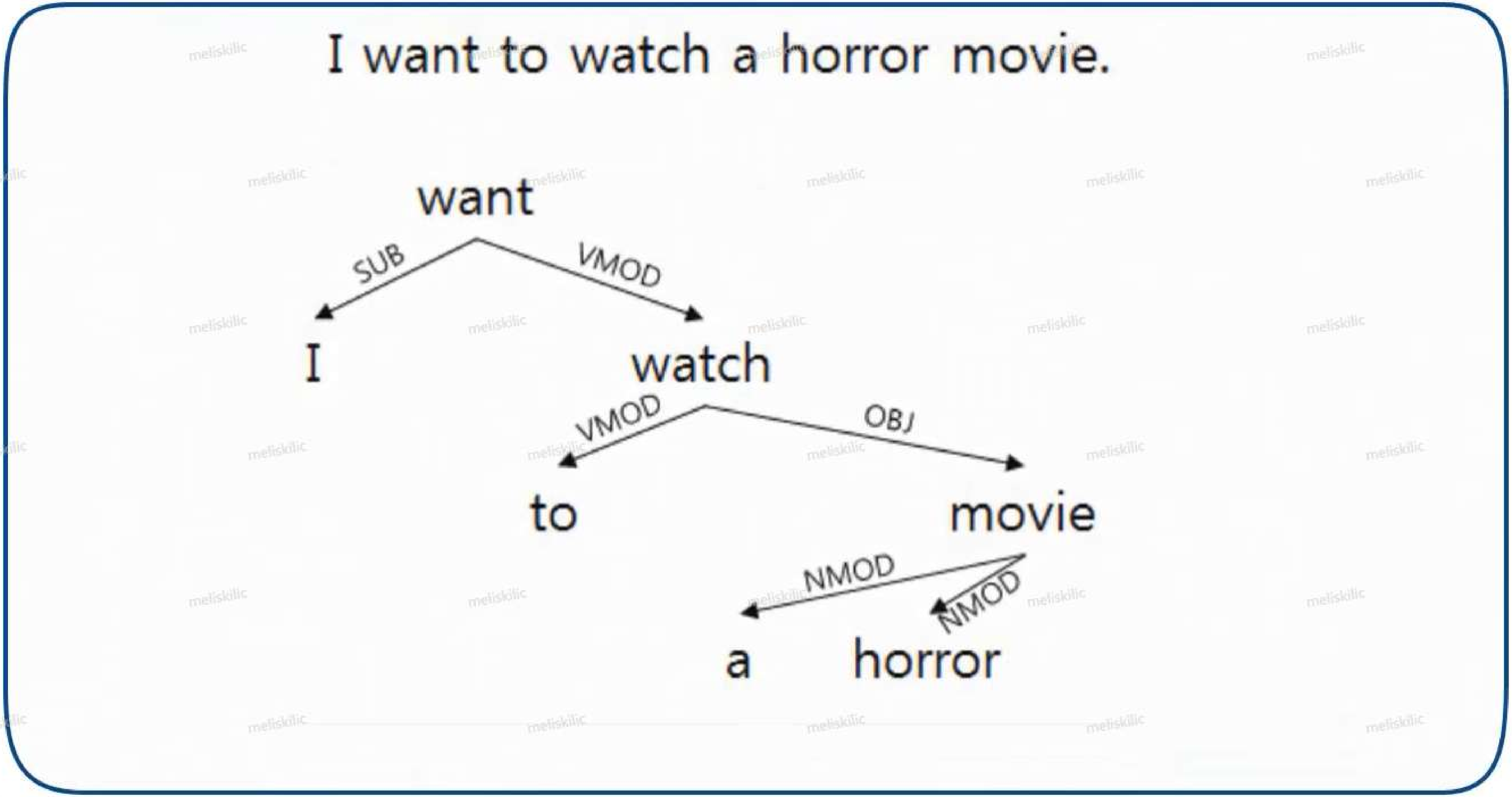
$$S(V_i) = (1 - d) + d \times \sum_{j \in In(V_i)} \left(\frac{1}{|Out(V_j)|} \times S(V_j)\right)$$

$$WS(V_i) = (1 - d) + d \times \sum_{V_j \in In(V_i)} \left(\frac{1}{\sum V_k \in Out(V_j) w_{jk}} \times WS(V_j)\right)$$

LDA Algorithm (2)



Syntax Analysis



Importance of Semantic Analysis

Syllogism

All men are mortal

Socrates is a man

Therefore, Socrates is mortal

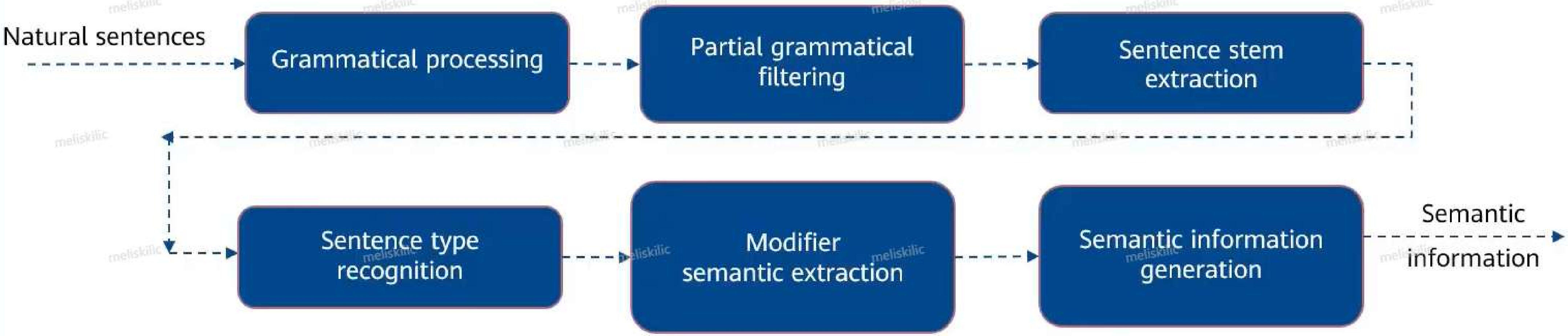
Inference:

All plants die.

All men die.

Men are plants.

Semantic Analysis



Semantic analysis process

Applications

Text classification

Automatic abstract

Text clustering

Information Extraction (IE)

Machine translation

Public opinion analysis

Question answering system

Machine writing