# An Evaluation Dataset for Intent Classification and Out-of-Scope Prediction

Petr Marek

# Outline

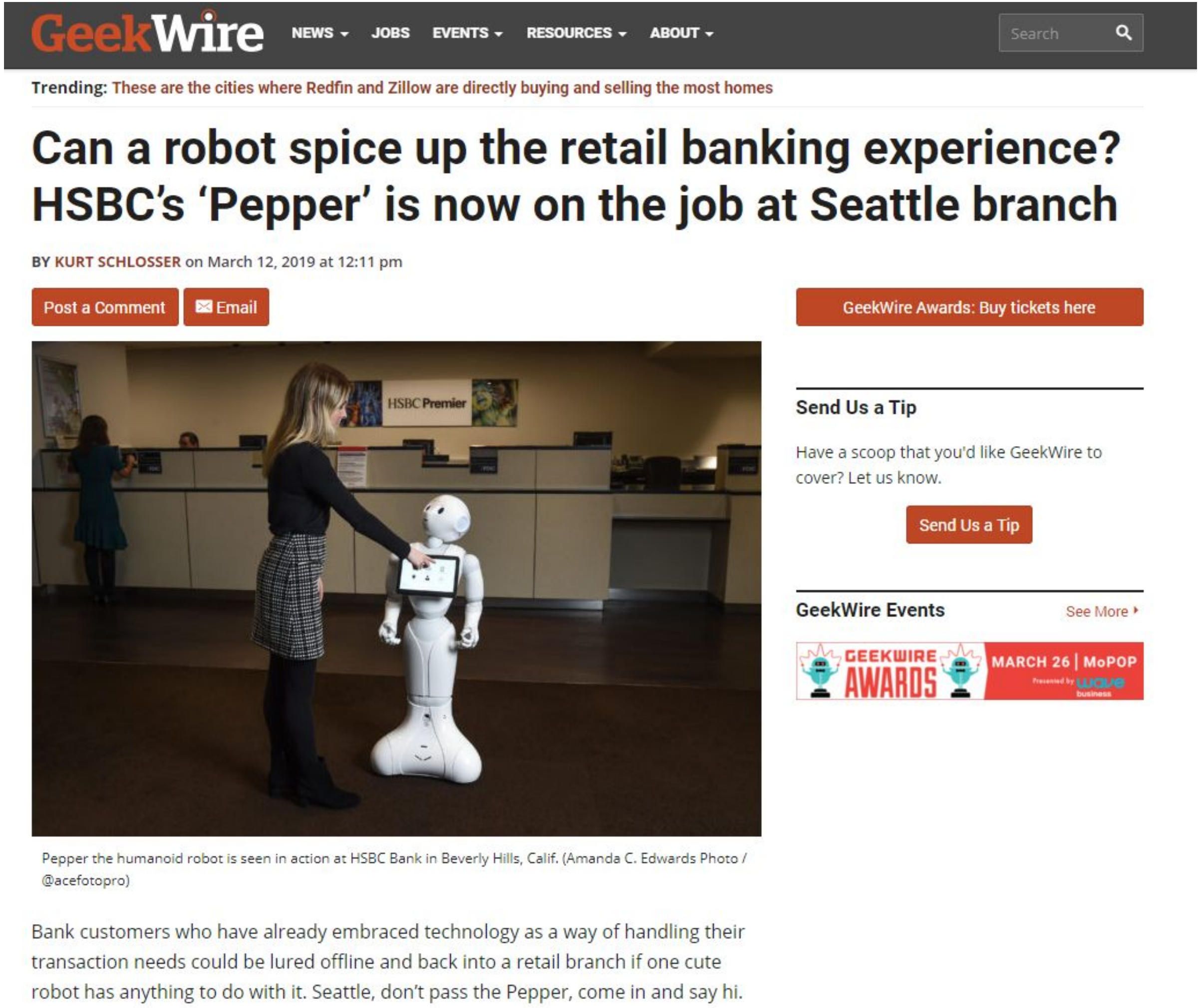
Problem statement

New Dataset

Metrics

Baseline Models

My Initial Work



Trending: These are the cities where Redfin and Zillow are directly buying and selling the most homes

# Can a robot spice up the retail banking experience? HSBC's 'Pepper' is now on the job at Seattle branch

BY KURT SCHLOSSER on March 12, 2019 at 12:11 pm

Pepper the humanoid robot is seen in action at HSBC Bank in Beverly Hills, Calif. (Amanda C. Edwards Photo / @acefotopro)

Bank customers who have already embraced technology as a way of handling their transaction needs could be lured offline and back into a retail branch if one cute robot has anything to do with it. Seattle, don't pass the Pepper, come in and say hi.

## Send Us a Tip

Have a scoop that you'd like GeekWire to cover? Let us know.

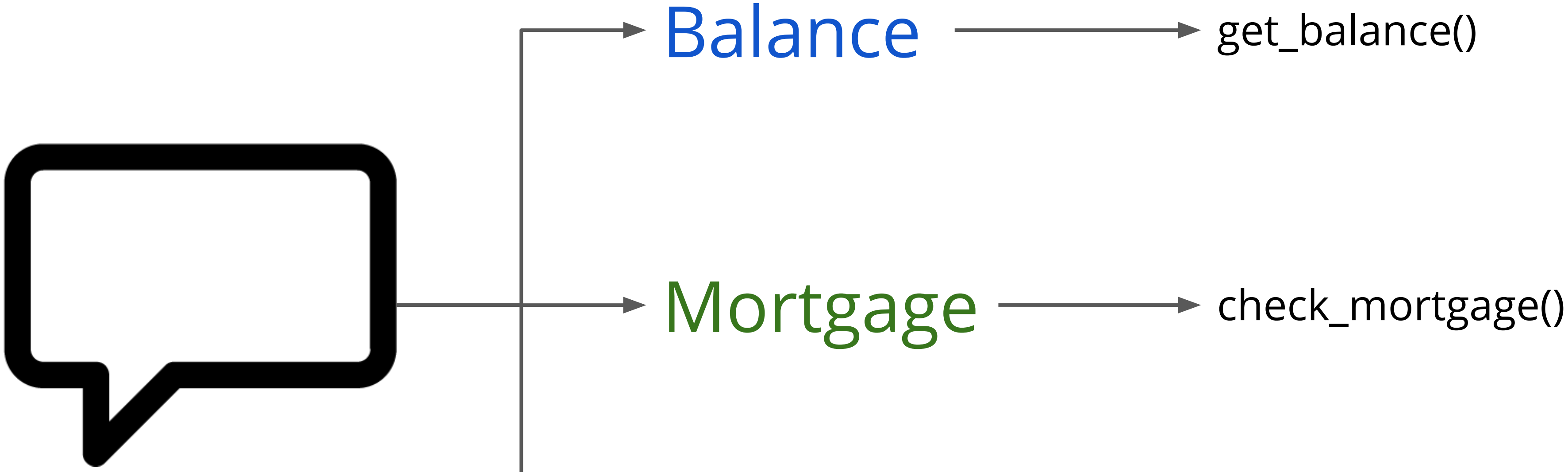
## GeekWire Events

What is my balance?

Can you give a mortgage?

I would like to open an account.

# Intent Classification



Balance → get_balance()

Mortgage → check_mortgage()

Account → open_accont()

| Balance |
|---|
| What is my balance? |
| Can you tell me how much money do I have left on my account? |
| How much money do I have? |
| Are there any money left on my account? |
| What is the balance of my account? |

| Mortgage |
|---|
| Can I have a mortgage? |
| I need a mortgage! |
| Can you assist me with a mortgage? |
| How can I get a mortgage? |
| I would like a mortgage! |

| Account |
|---|
| I want a new account! |
| Can you open me an account? |
| I would like to become your customer. |
| I would like to open an account. |
| New account, I need it now! |

● Mortgage

● Account

● Balance

| | |
|---|---|
| What is my balance? | Balance |
| Can you give a mortgage? | Mortgage |
| I would like to open an account. | Account |

Do you like movies?

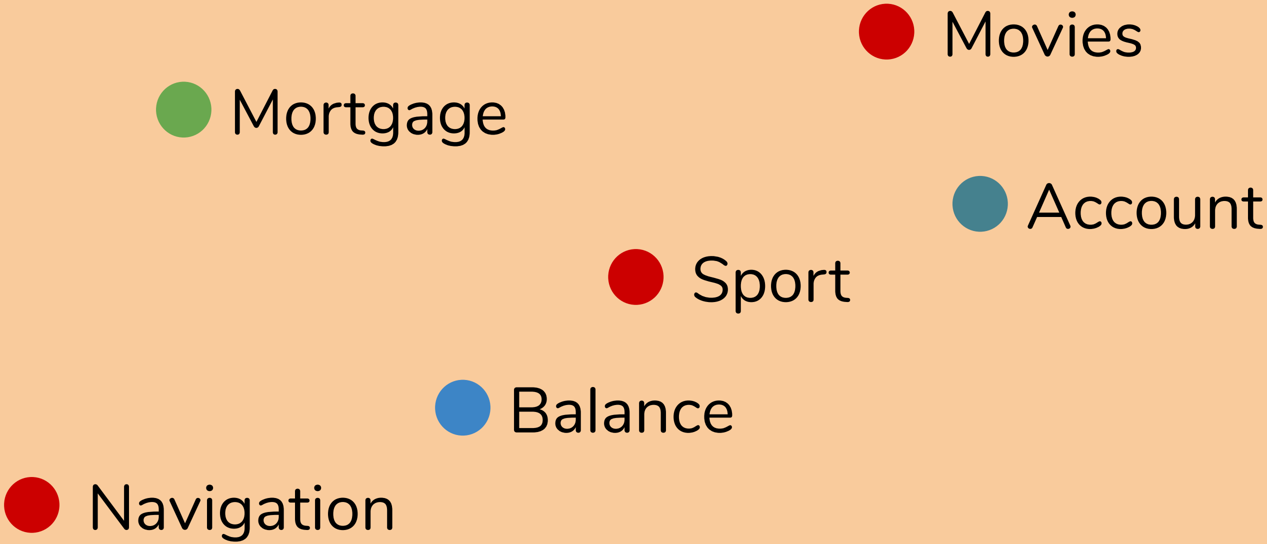
How are my sport teams doing?

How to get from Paris to London?

# Other

- Mortgage
- Account
- Balance

# Other



● Movies

● Mortgage

● Account

● Sport

● Balance

● Navigation

$$y^* = \text{argmin}_{\hat{y}} \sum_{y \in Y} L(y, \hat{y}) \Pr[y|x]$$

Zero-one Loss

$$y^* = \text{argmax}_{y} \Pr[y|x]$$

# Other

🔴 Movies

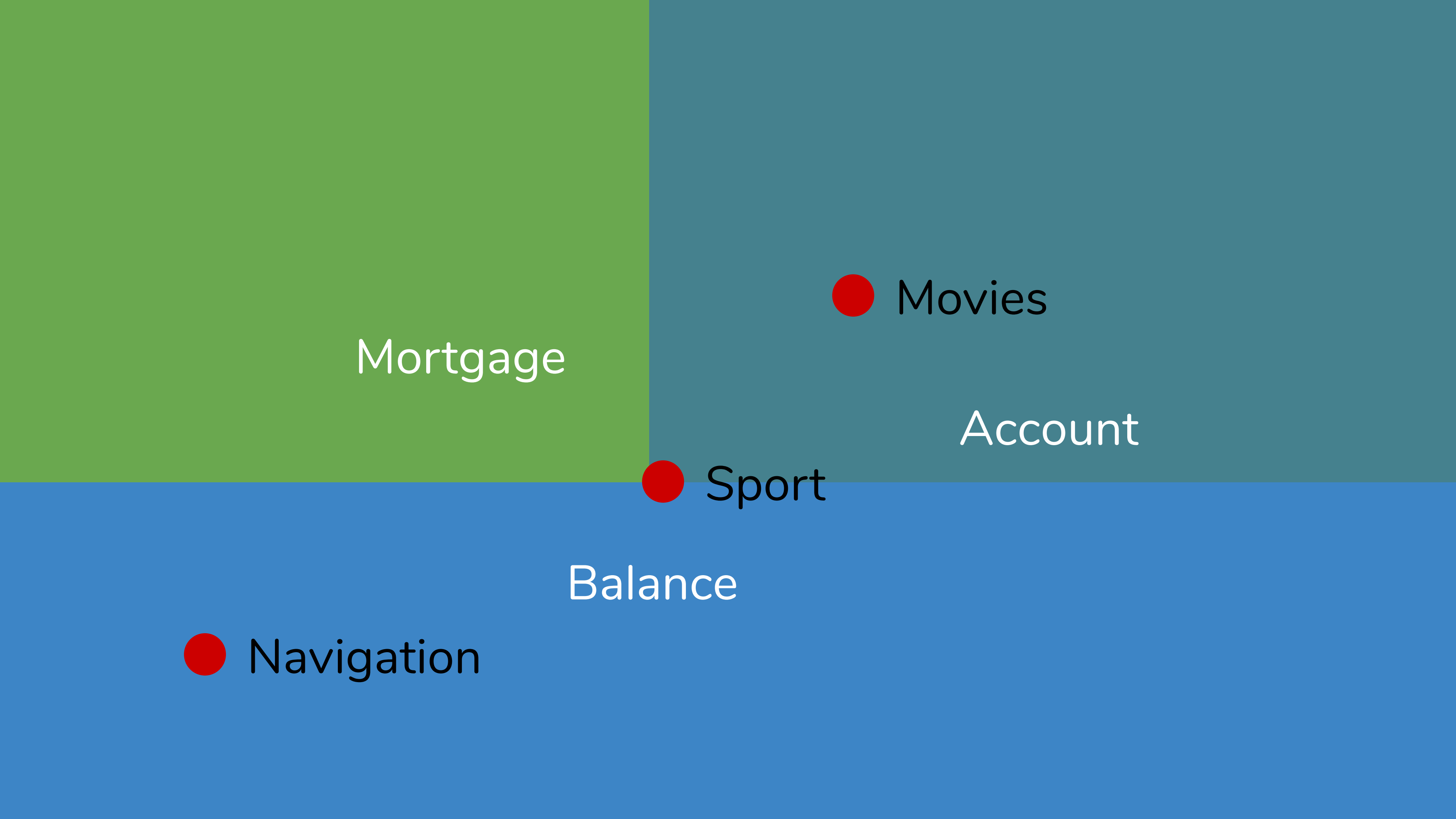
🟢 Mortgage

🔵 Account

🔴 Sport

🔵 Balance

🔴 Navigation



Movies

Mortgage

Account

Sport

Balance

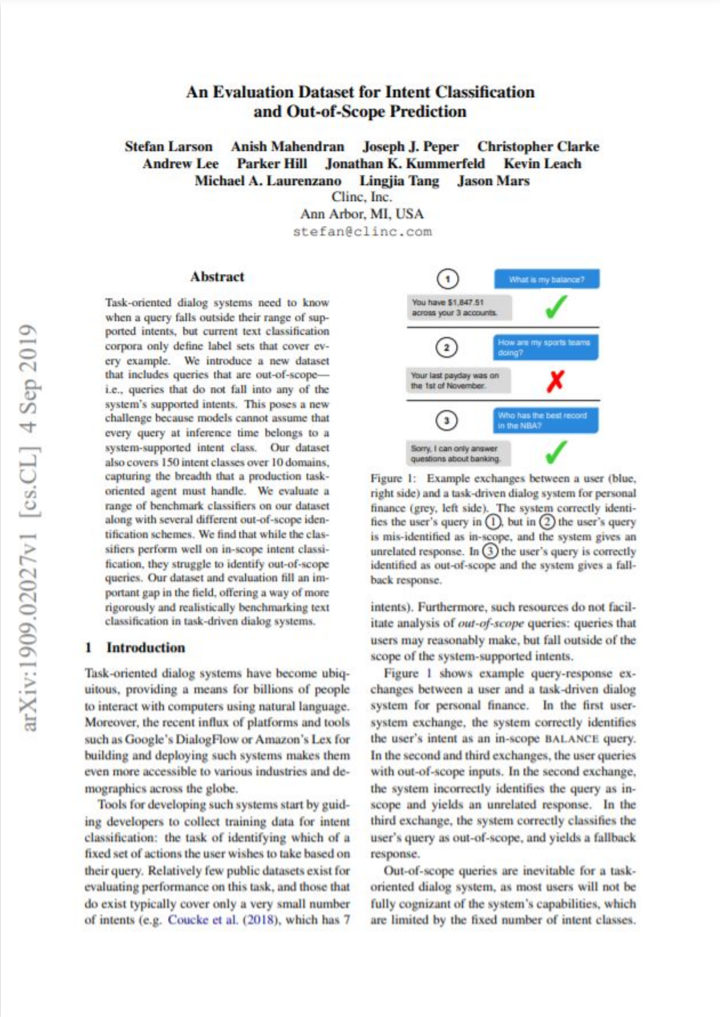
Navigation

# What is the most similar intent?

What is the most similar intent?

AND

Isn’t it out of scope?



# An Evaluation Dataset for Intent Classification and Out-of-Scope Prediction

**Stefan Larson   Anish Mahendran   Joseph J. Peper   Christopher Clarke**
**Andrew Lee   Parker Hill   Jonathan K. Kummerfeld   Kevin Leach**
**Michael A. Laurenzano   Lingjia Tang   Jason Mars**
Clinc, Inc.
Ann Arbor, MI, USA
stefan@clinc.com

## Abstract

Task-oriented dialog systems need to know when a query falls outside their range of supported intents, but current text classification corpora only define label sets that cover every example. We introduce a new dataset that includes queries that are out-of-scope—i.e., queries that do not fall into any of the system's supported intents. This poses a new challenge because models cannot assume that every query at inference time belongs to a system-supported intent class. Our dataset also covers 150 intent classes over 10 domains, capturing the breadth that a production task-oriented agent must handle. We evaluate a range of benchmark classifiers on our dataset along with several different out-of-scope identification schemes. We find that while the classifiers perform well on in-scope intent classification, they struggle to identify out-of-scope queries. Our dataset and evaluation fill an important gap in the field, offering a way of more rigorously and realistically benchmarking text classification in task-driven dialog systems.

## 1 Introduction

Task-oriented dialog systems have become ubiquitous, providing a means for billions of people to interact with computers using natural language. Moreover, the recent influx of platforms and tools such as Google's DialogFlow or Amazon's Lex for building and deploying such systems makes them even more accessible to various industries and demographics across the globe.

Tools for developing such systems start by guiding developers to collect training data for intent classification: the task of identifying which of a fixed set of actions the user wishes to take based on their query. Relatively few public datasets exist for evaluating performance on this task, and those that do exist typically cover only a very small number of intents (e.g. Coucke et al. (2018), which has 7 intents). Furthermore, such resources do not facilitate analysis of *out-of-scope* queries: queries that users may reasonably make, but fall outside of the scope of the system-supported intents.

①
What is my balance?
You have $1,847.51 across your 3 accounts.
✓

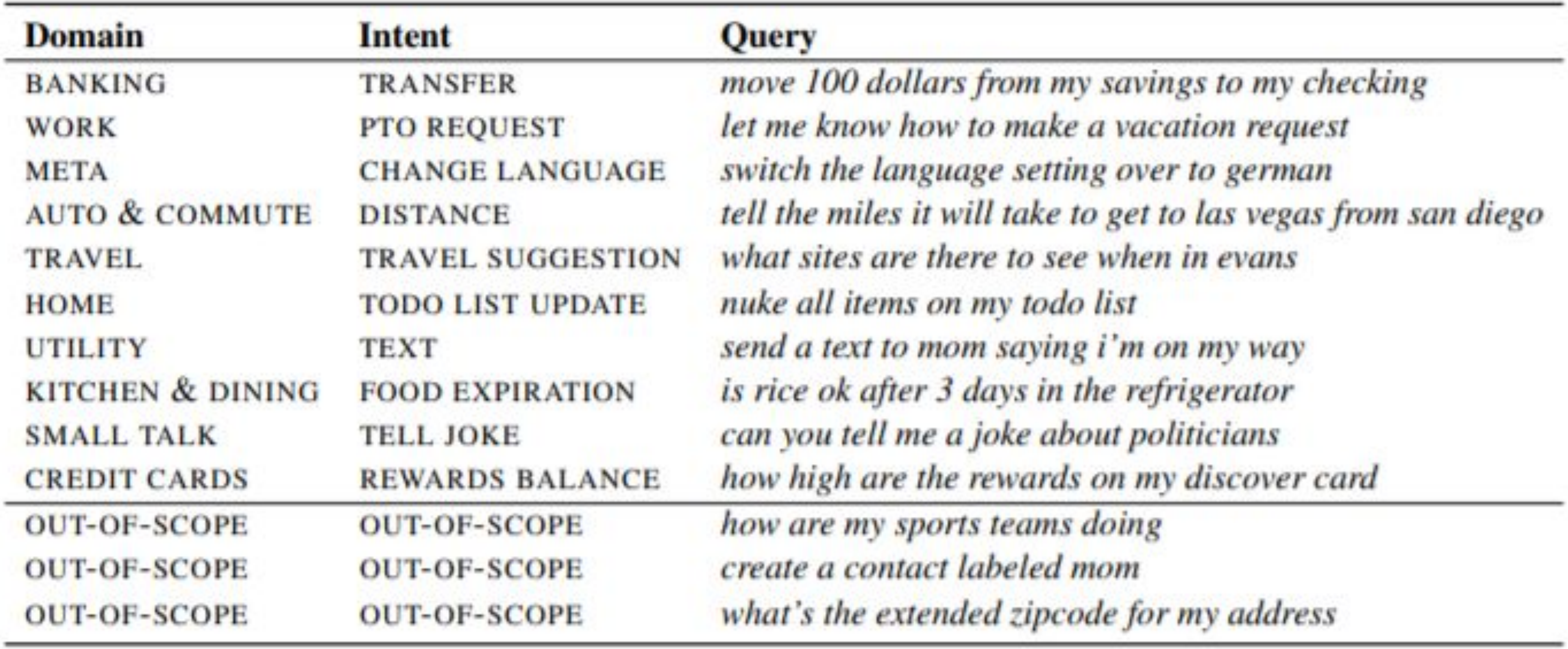
②
How are my sports teams doing?
Your last payday was on the 1st of November.
✗

③
Who has the best record in the NBA?
Sorry, I can only answer questions about banking.
✓

Figure 1: Example exchanges between a user (blue, right side) and a task-driven dialog system for personal finance (grey, left side). The system correctly identifies the user's query in ①, but in ② the user's query is mis-identified as in-scope, and the system gives an unrelated response. In ③ the user's query is correctly identified as out-of-scope and the system gives a fallback response.

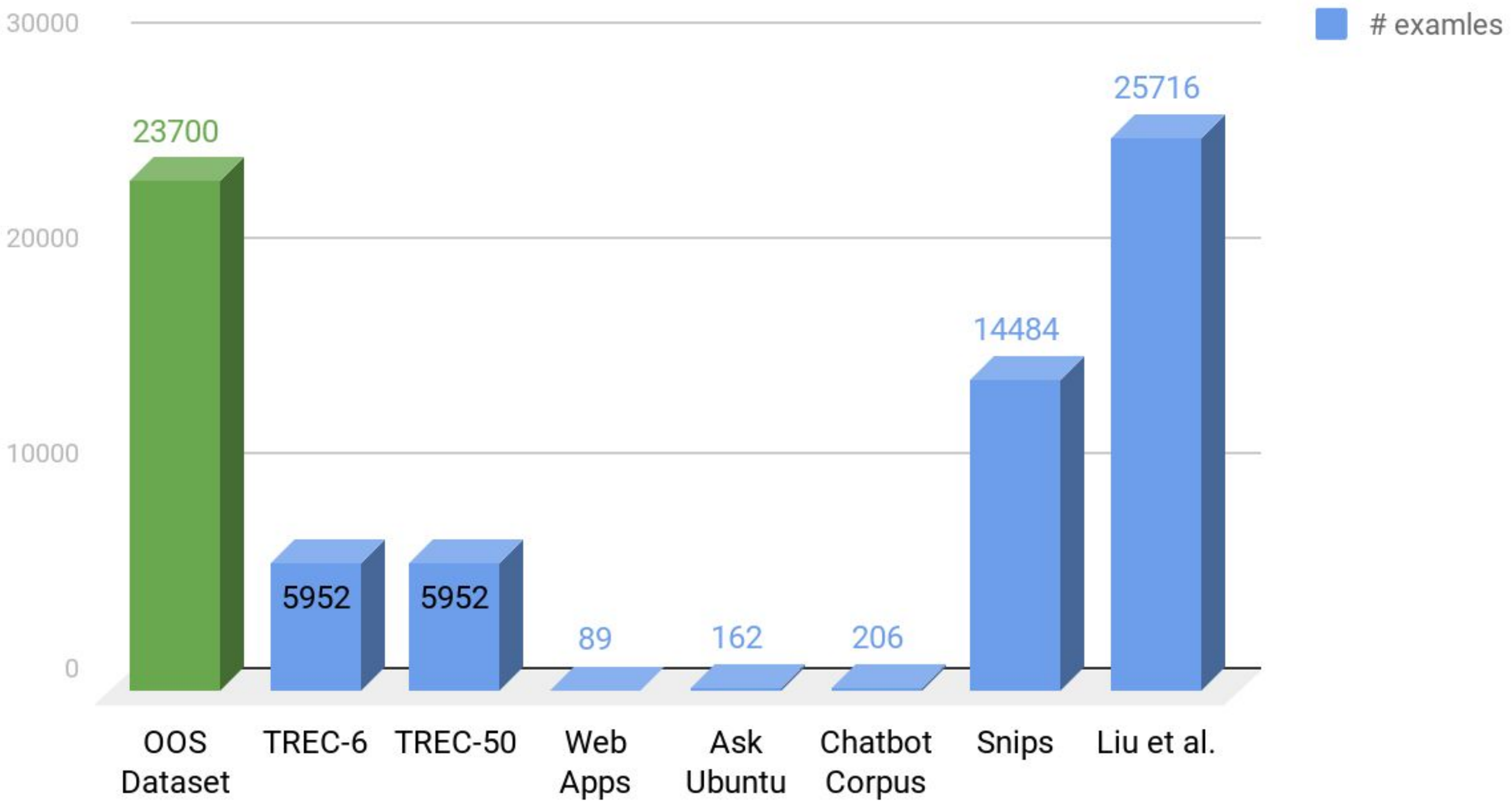
Figure 1 shows example query-response exchanges between a user and a task-driven dialog system for personal finance. In the first user-system exchange, the system correctly identifies the user's intent as an in-scope BALANCE query. In the second and third exchanges, the user queries with out-of-scope inputs. In the second exchange, the system incorrectly identifies the query as in-scope and yields an unrelated response. In the third exchange, the system correctly classifies the user's query as out-of-scope, and yields a fallback response.

Out-of-scope queries are inevitable for a task-oriented dialog system, as most users will not be fully cognizant of the system's capabilities, which are limited by the fixed number of intent classes.

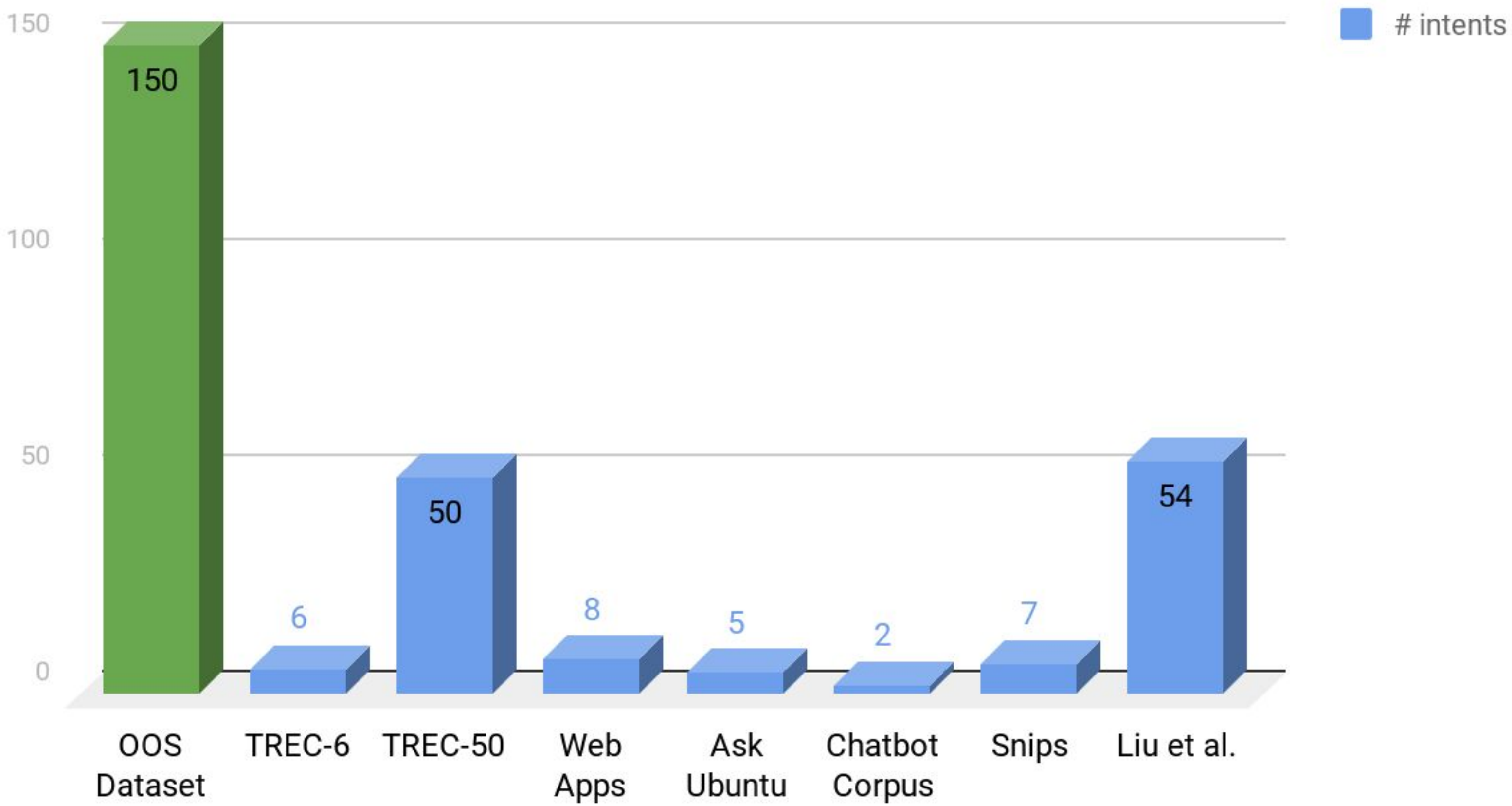


| **Domain** | **Intent** | **Query** |
|---|---|---|
| BANKING | TRANSFER | *move 100 dollars from my savings to my checking* |
| WORK | PTO REQUEST | *let me know how to make a vacation request* |
| META | CHANGE LANGUAGE | *switch the language setting over to german* |
| AUTO & COMMUTE | DISTANCE | *tell the miles it will take to get to las vegas from san diego* |
| TRAVEL | TRAVEL SUGGESTION | *what sites are there to see when in evans* |
| HOME | TODO LIST UPDATE | *nuke all items on my todo list* |
| UTILITY | TEXT | *send a text to mom saying i'm on my way* |
| KITCHEN & DINING | FOOD EXPIRATION | *is rice ok after 3 days in the refrigerator* |
| SMALL TALK | TELL JOKE | *can you tell me a joke about politicians* |
| CREDIT CARDS | REWARDS BALANCE | *how high are the rewards on my discover card* |
| OUT-OF-SCOPE | OUT-OF-SCOPE | *how are my sports teams doing* |
| OUT-OF-SCOPE | OUT-OF-SCOPE | *create a contact labeled mom* |
| OUT-OF-SCOPE | OUT-OF-SCOPE | *what's the extended zipcode for my address* |

# Dataset Size



# Number of Intents



| Dataset | Out-of-Scope Utterances |
|---|---|
| OOS Dataset | ✓ |
| TREC-6 | ✗ |
| TREC-50 | ✗ |
| Web Apps | ✗ |
| Ask Ubuntu | ✗ |
| Chatbot Corpus | ✗ |
| Snips | ✗ |
| Liu et al. | ✗ |

## OOS Train

Train Out-of-Scope as 151st intent

## OOS Threshold

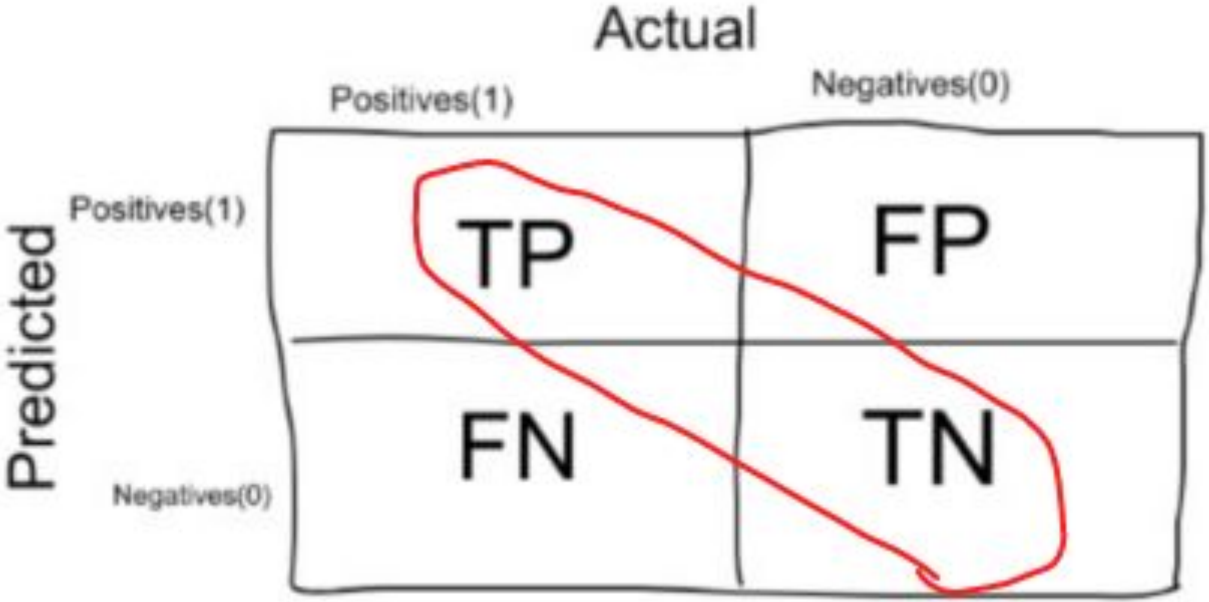
A threshold on the classifier’s probability estimate

## OOS Binary

Two-stage process where we

first classify a query as in- or out-of-scope,

then classify it into one of the 150 intents if classified as in-scope

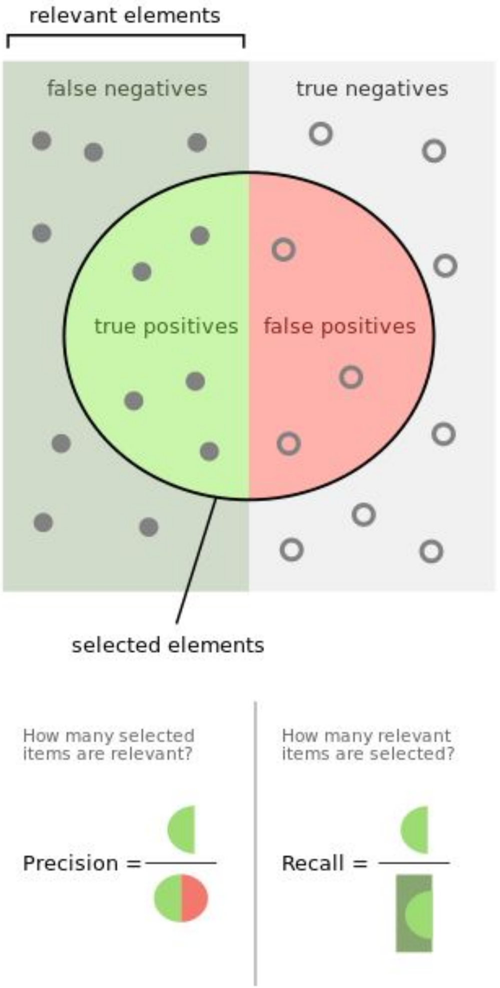


| | | Actual | |
|---|---|---|---|
| | | Positives(1) | Negatives(0) |
| Predicted | Positives(1) | TP | FP |
| | Negatives(0) | FN | TN |

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + FN + TN}$$

|  |  | Actual |  |
|---|---|---|---|
|  |  | Positives(1) | Negatives(0) |
| Predicted | Positives(1) | TP | FP |
|  | Negatives(0) | FN | TN |

$$\text{Recall} = \frac{TP}{TP + FN}$$



relevant elements

false negatives

true negatives

true positives

false positives

selected elements

How many selected items are relevant?

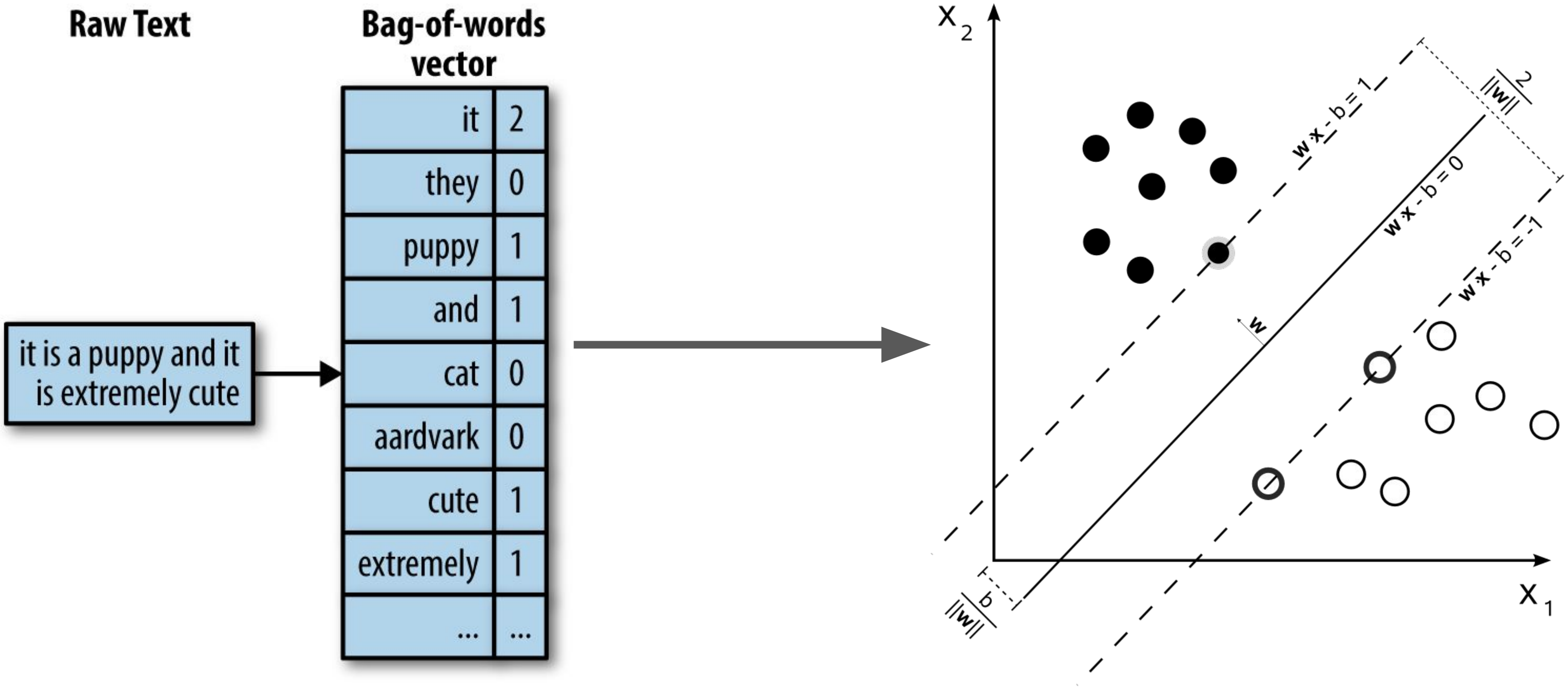
Precision =

How many relevant items are selected?

Recall =

| Baseline Models |
|---|
| Support Vector Machine |
| Multilayer Perceptron |
| FastText |
| Convolutional Neural Network |
| Bert |
| Google’s DialogFlow |
| Rasa NLU |

# Support Vector Machine



**Raw Text**

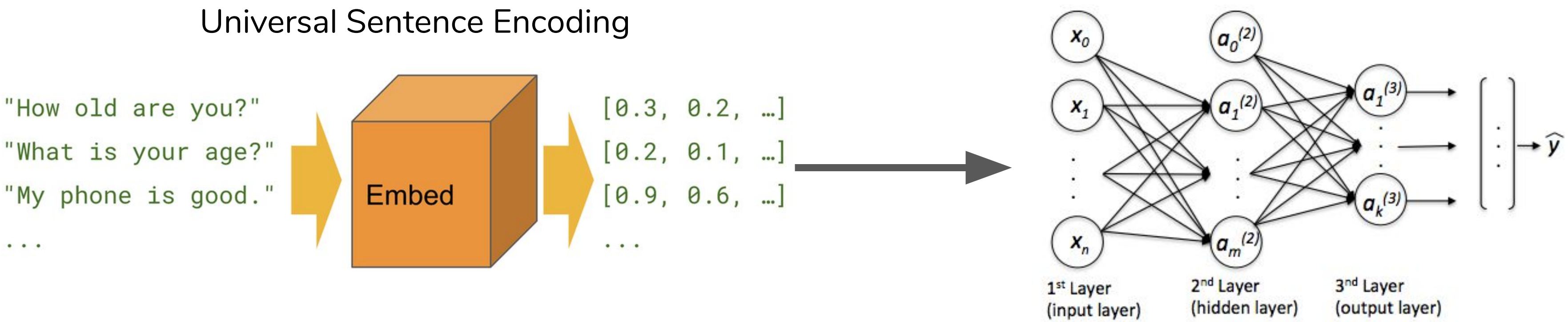
it is a puppy and it is extremely cute

**Bag-of-words vector**

| | |
|---|---|
| it | 2 |
| they | 0 |
| puppy | 1 |
| and | 1 |
| cat | 0 |
| aardvark | 0 |
| cute | 1 |
| extremely | 1 |
| ... | ... |

$X_2$
$X_1$
$\mathbf{w}\cdot\mathbf{x} - b = 1$
$\mathbf{w}\cdot\mathbf{x} - b = 0$
$\mathbf{w}\cdot\mathbf{x} - b = -1$
$\frac{2}{\|\mathbf{w}\|}$
$\frac{b}{\|\mathbf{w}\|}$
$\mathbf{w}$

# Multilayer Perceptron

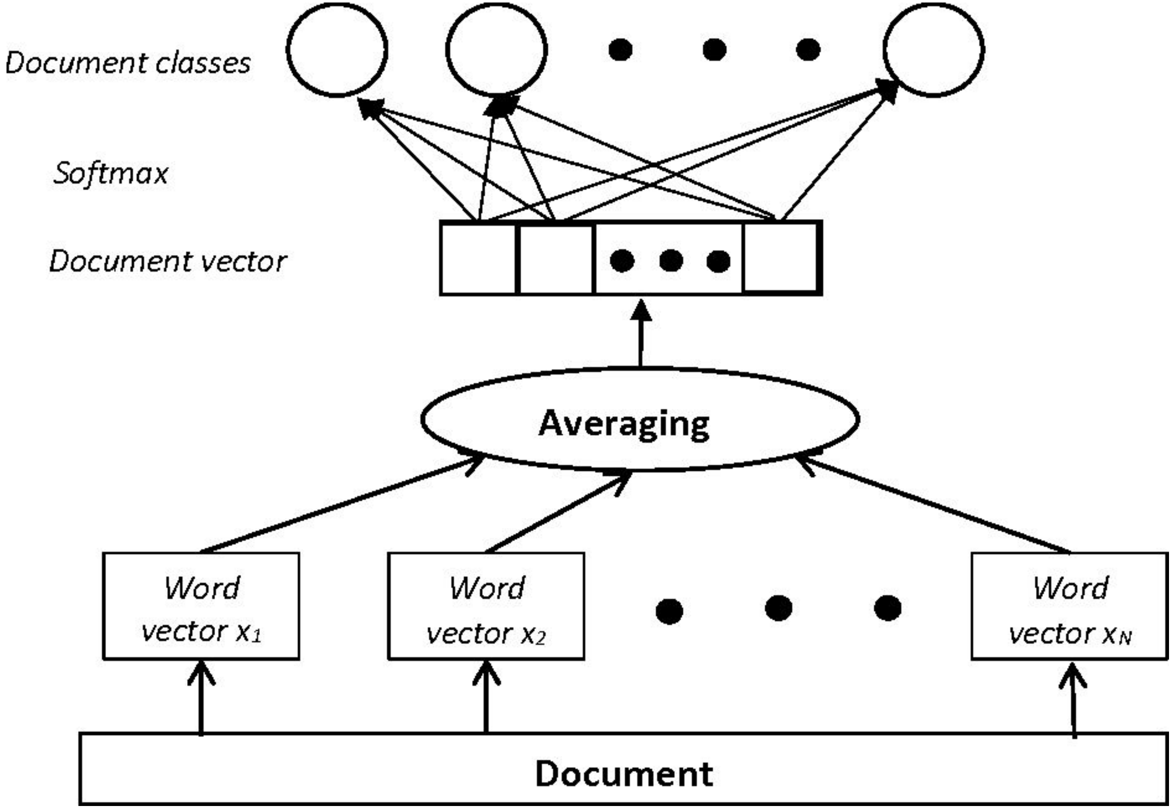


Universal Sentence Encoding

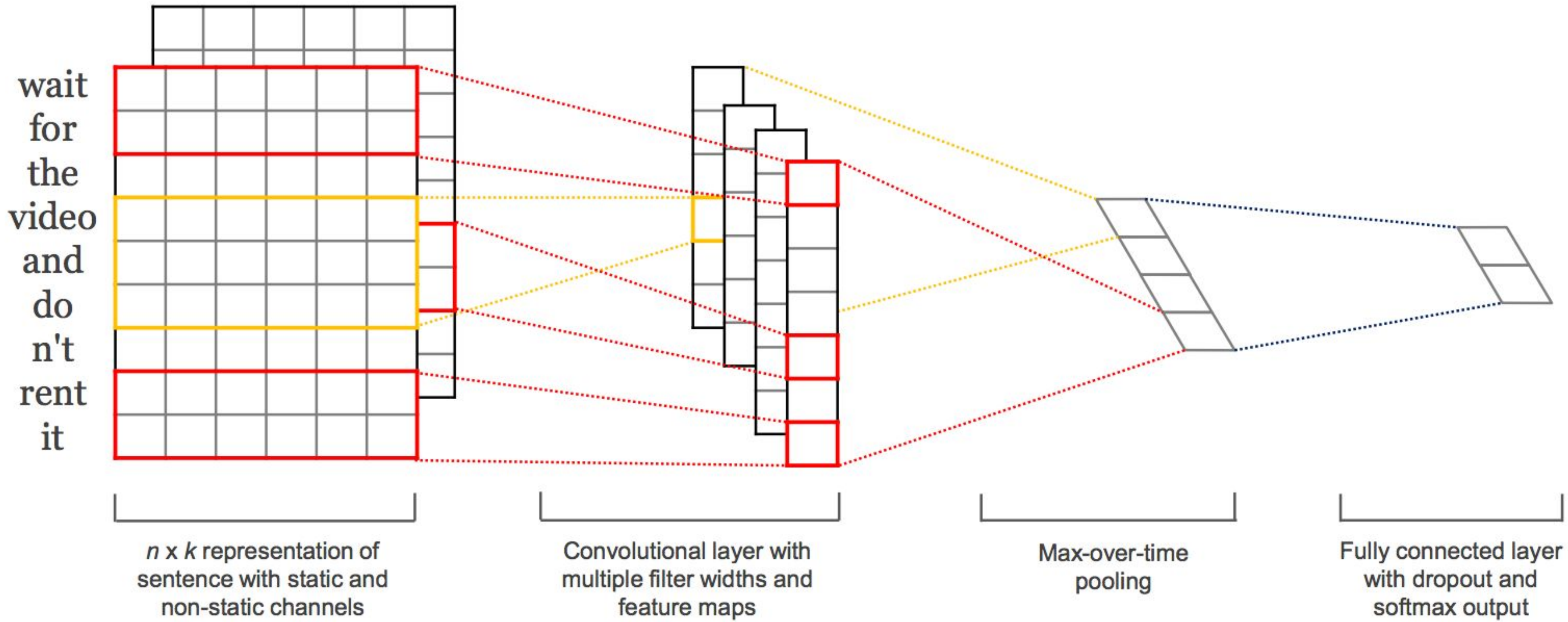
```
"How old are you?"
"What is your age?"
"My phone is good."
...
```

```
[0.3, 0.2, …]
[0.2, 0.1, …]
[0.9, 0.6, …]
...
```

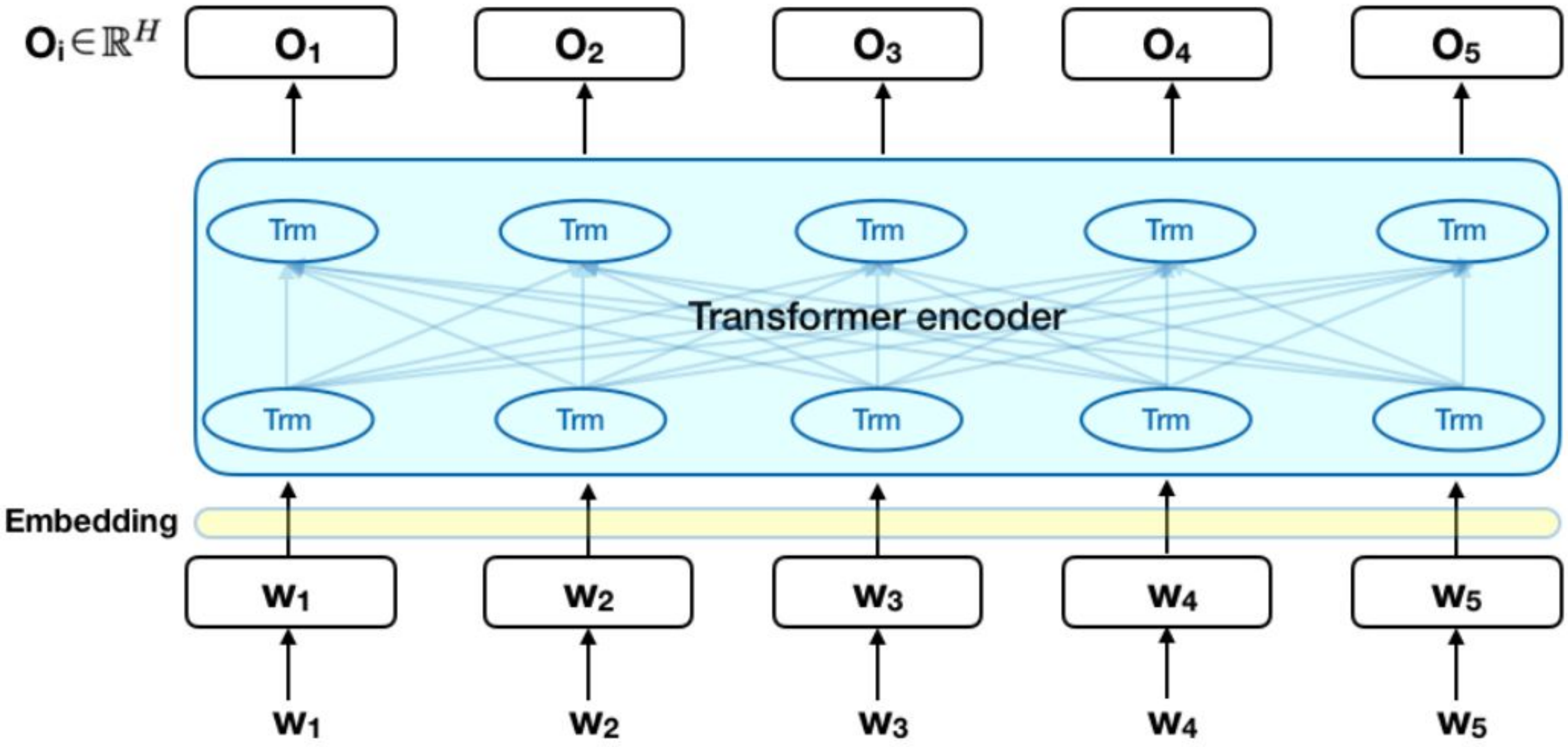
# FastText


Document classes
Softmax
Document vector
Averaging
Word vector $x_1$
Word vector $x_2$
Word vector $x_N$
Document

# Convolutional Neural Network


wait
for
the
video
and
do
n't
rent
it
*n* x *k* representation of sentence with static and non-static channels
Convolutional layer with multiple filter widths and feature maps
Max-over-time pooling
Fully connected layer with dropout and softmax output

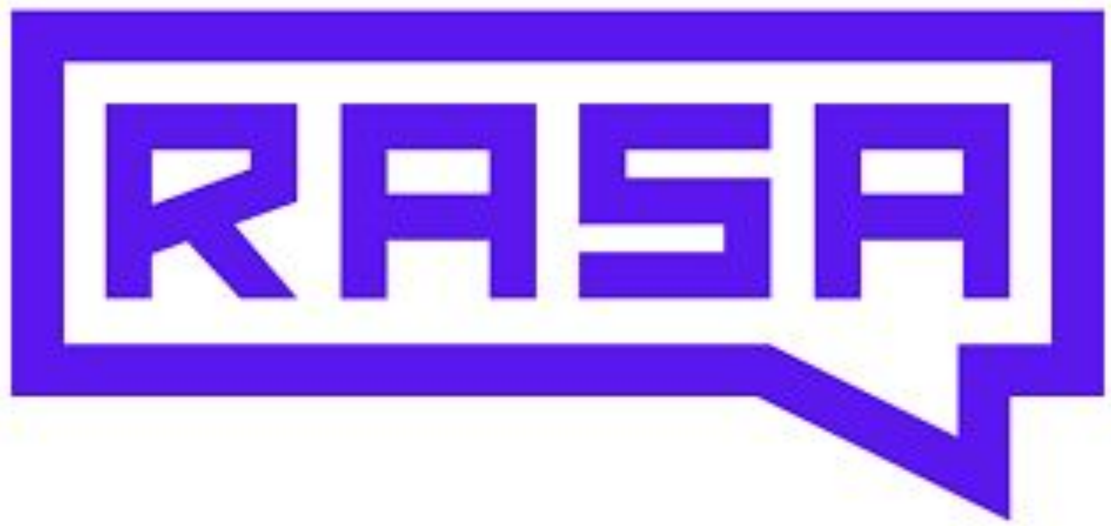
# BERT



$O_i \in \mathbb{R}^H$
$O_1$
$O_2$
$O_3$
$O_4$
$O_5$
Trm
Trm
Trm
Trm
Trm
Transformer encoder
Trm
Trm
Trm
Trm
Trm
Embedding
$w_1$
$w_2$
$w_3$
$w_4$
$w_5$
$w_1$
$w_2$
$w_3$
$w_4$
$w_5$

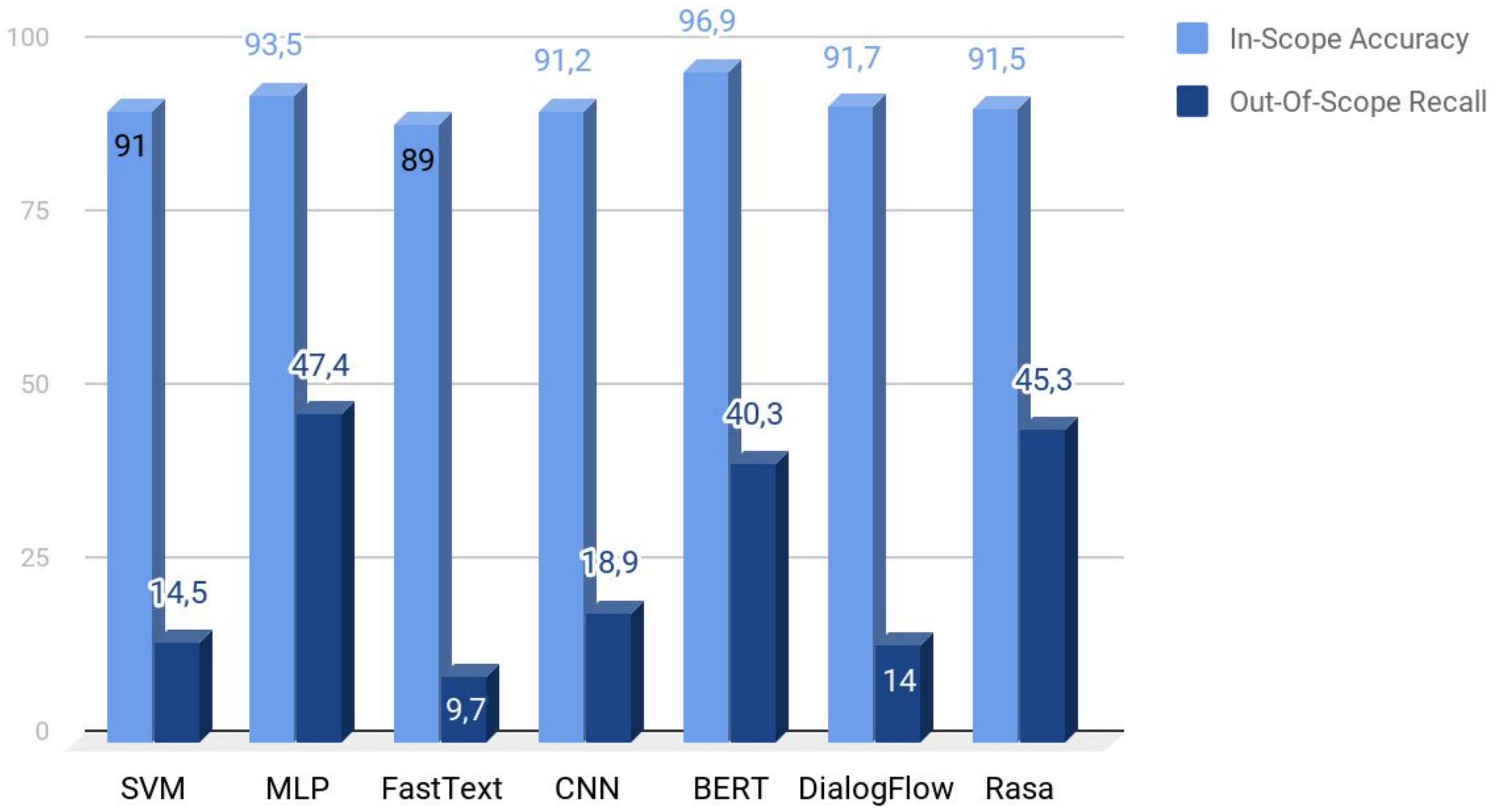
# Google’s DialogFlow

# Rasa NLU

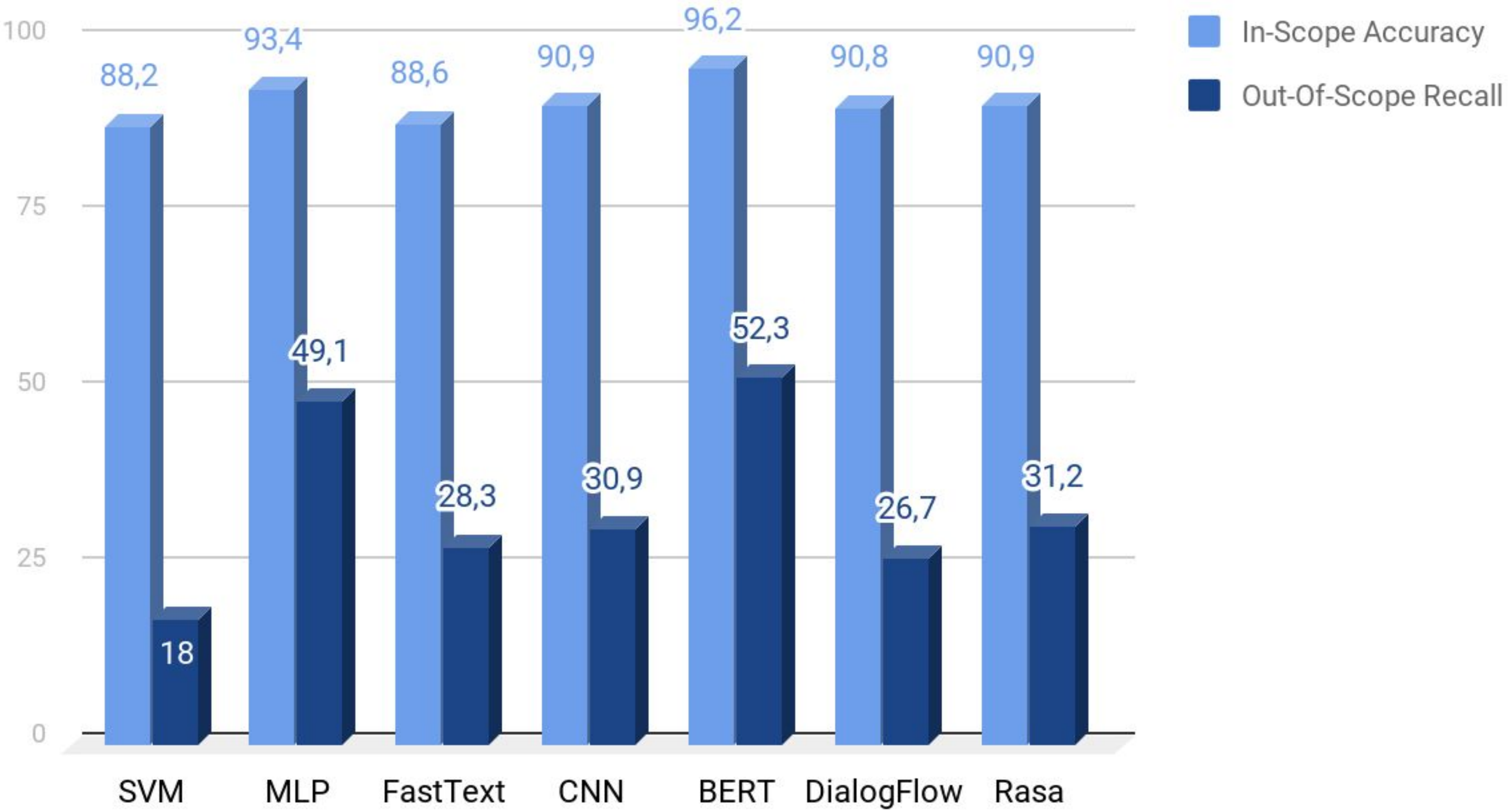


# Baseline Results

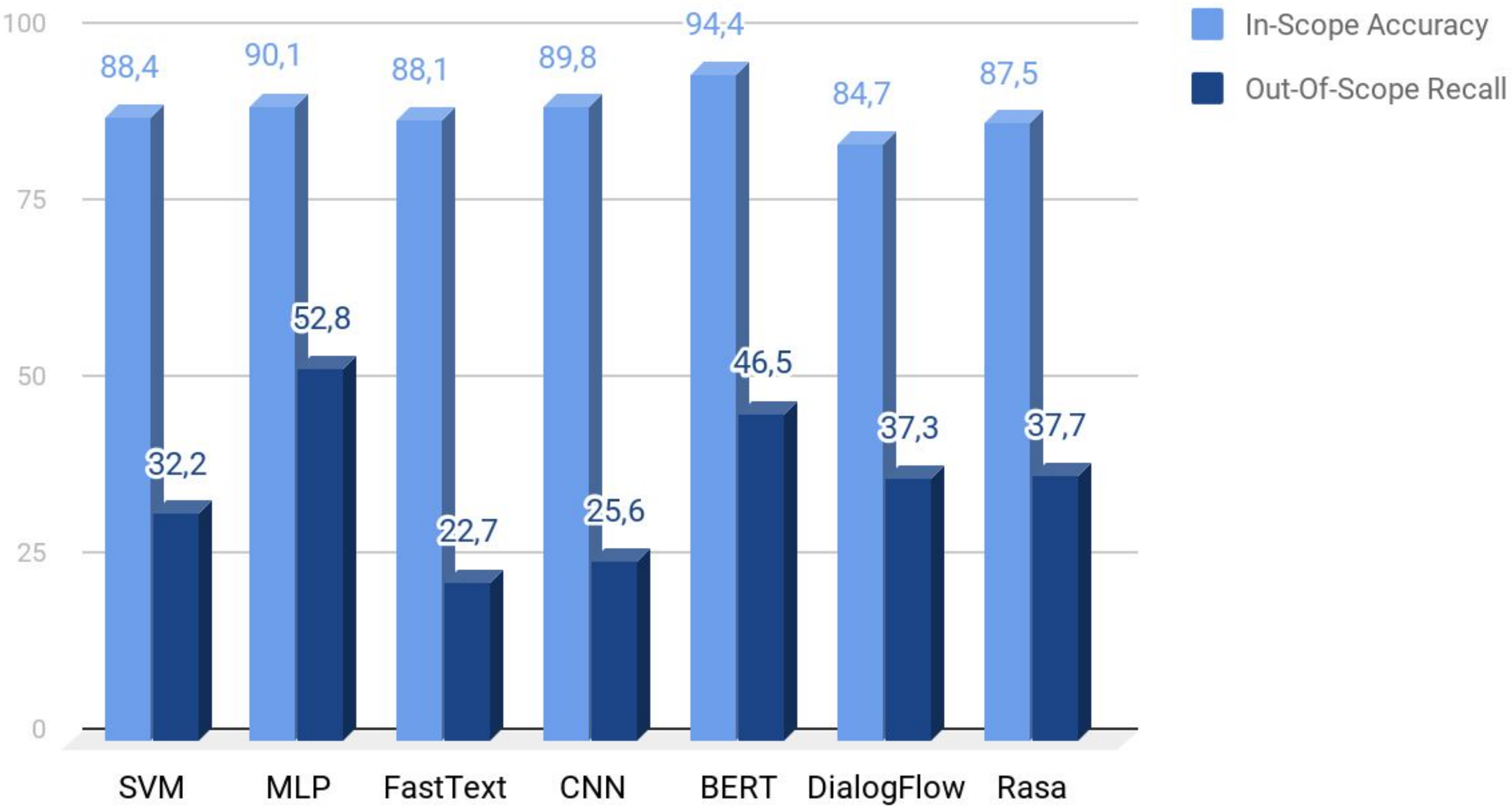
# OOS-Train


100
75
50
25
0
In-Scope Accuracy
Out-Of-Scope Recall
SVM
MLP
FastText
CNN
BERT
DialogFlow
Rasa
91
14,5
93,5
47,4
89
9,7
91,2
18,9
96,9
40,3
91,7
14
91,5
45,3

## OOS-Threshold

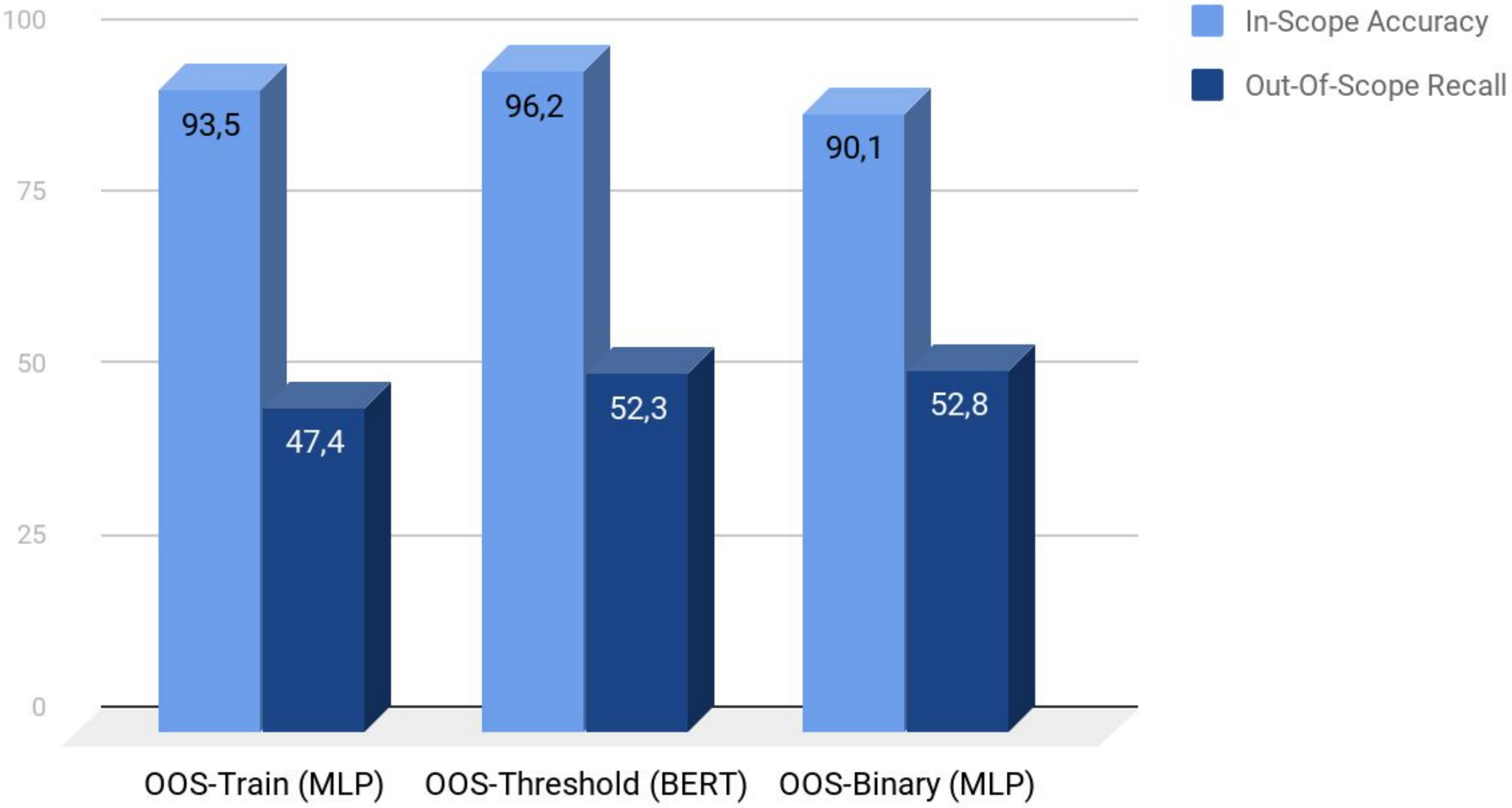


## OOS-Binary

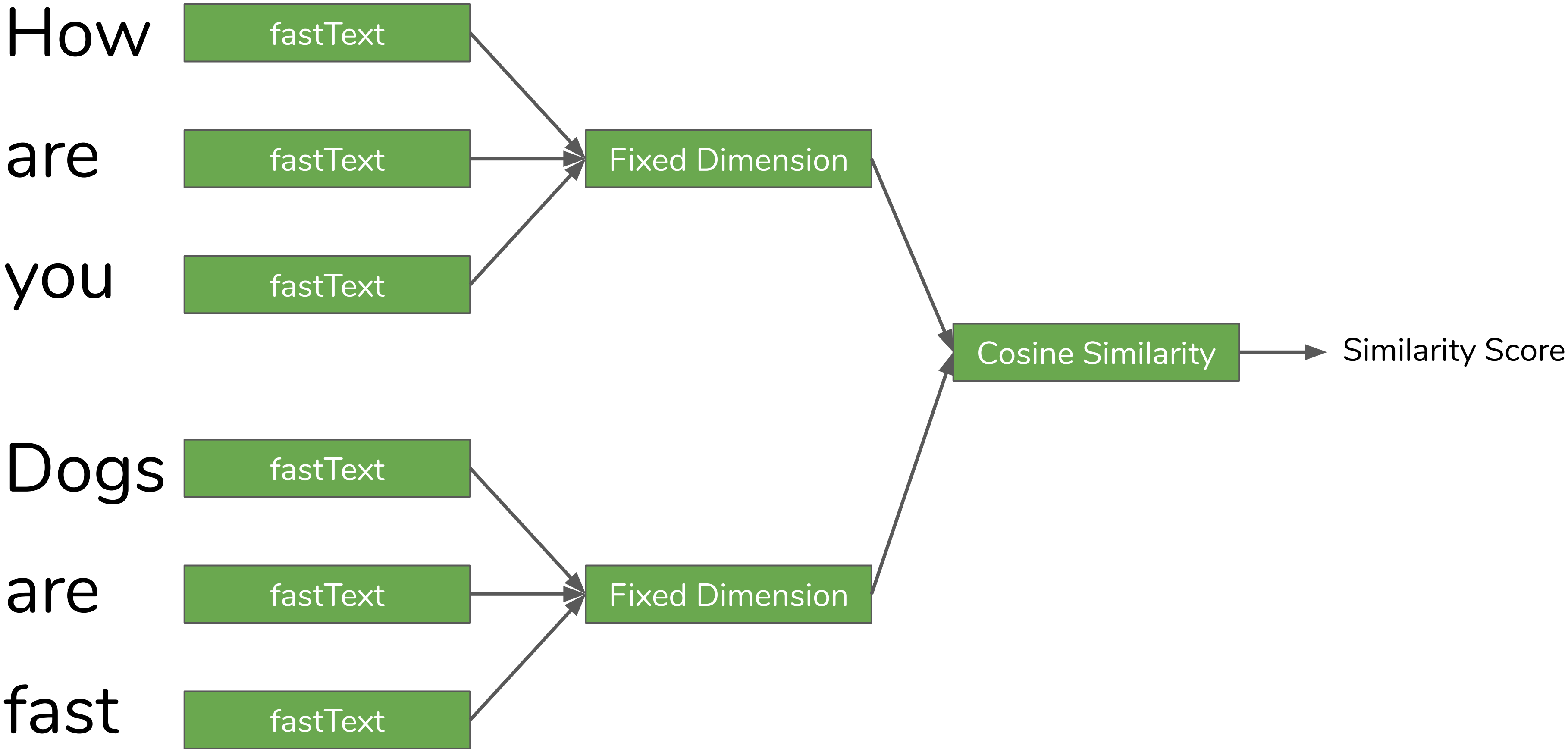


100
75
50
25
0
88,4
32,2
90,1
52,8
88,1
22,7
89,8
25,6
94,4
46,5
84,7
37,3
87,5
37,7
SVM
MLP
FastText
CNN
BERT
DialogFlow
Rasa
In-Scope Accuracy
Out-Of-Scope Recall

# Comparison of Approaches



# My Initial Work



How fastText

are fastText

you fastText

Fixed Dimension

Dogs fastText

are fastText

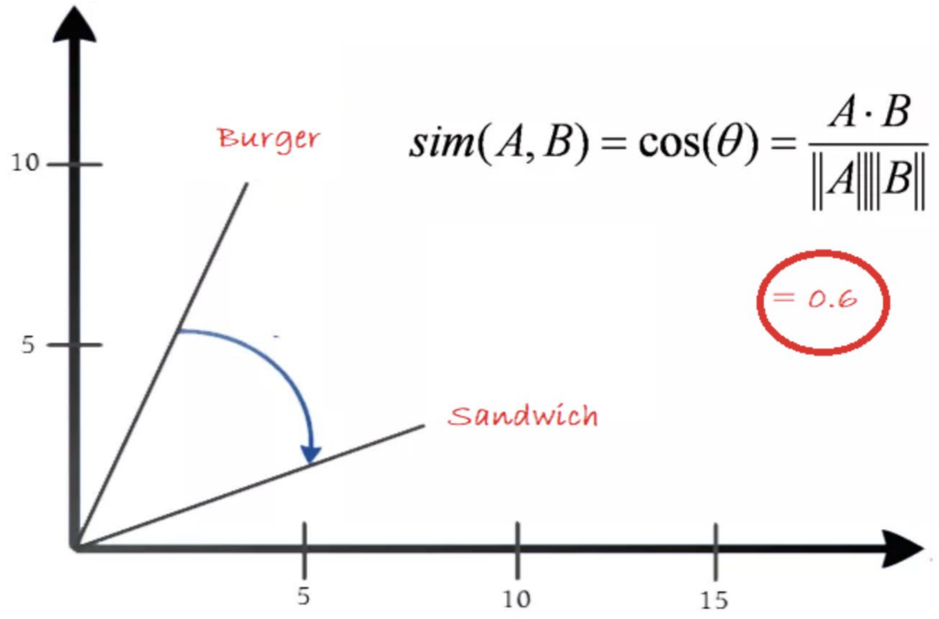
fast fastText

Fixed Dimension

Cosine Similarity

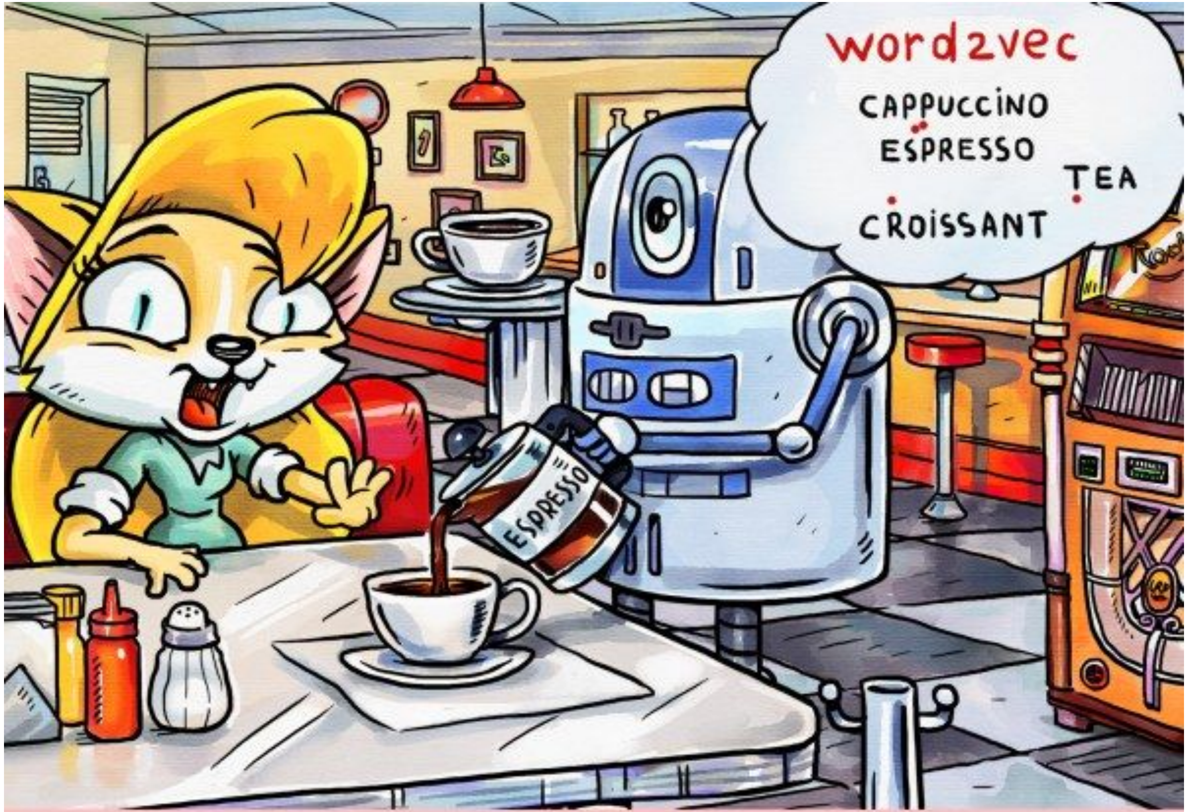
Similarity Score

$$similarity(A,B) = \frac{A \cdot B}{\|A\| \times \|B\|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \times \sqrt{\sum_{i=1}^{n} B_i^2}}$$

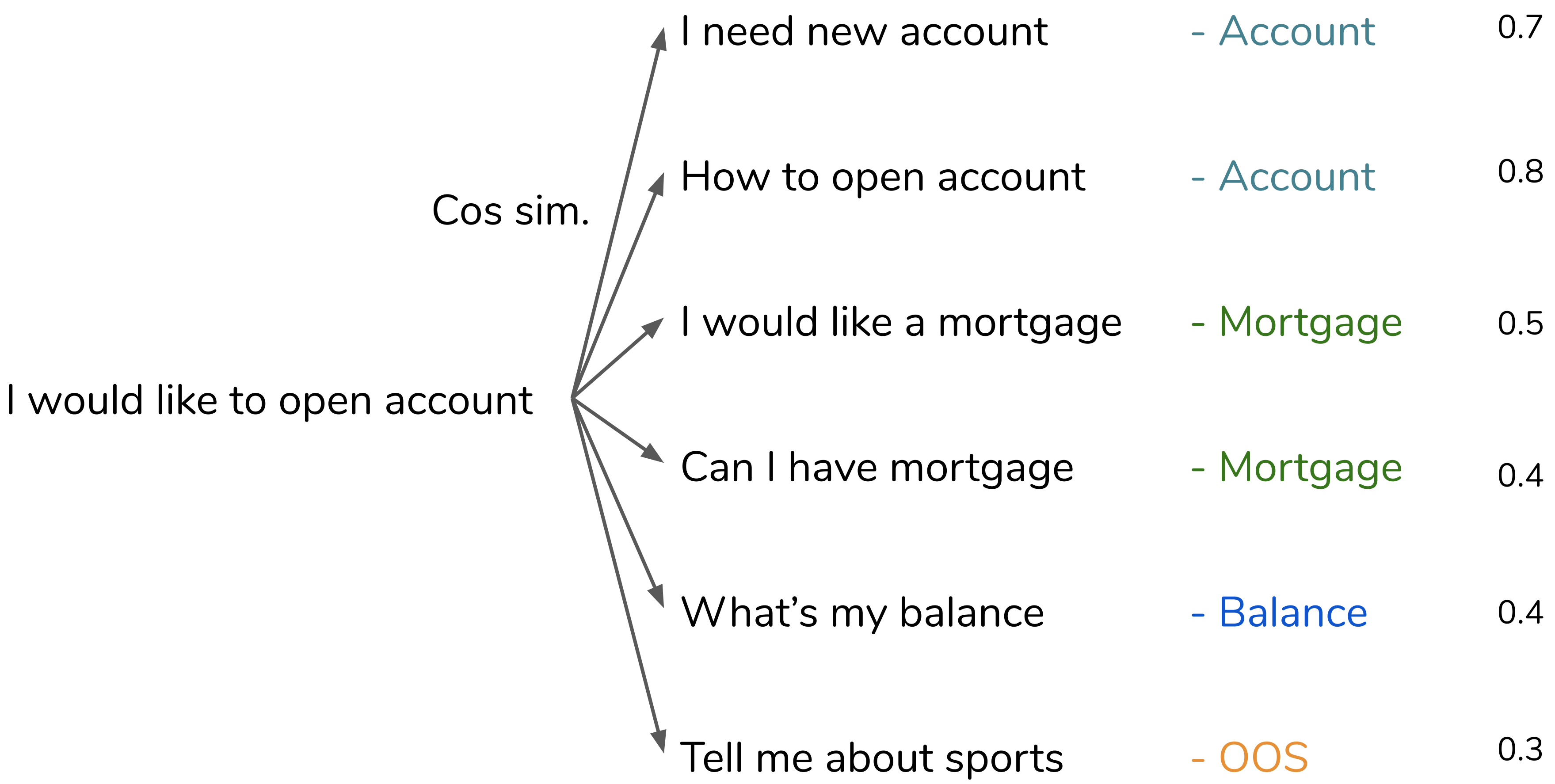


$$sim(A,B) = \cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|}$$

= 0.6



- Espresso? But I ordered a cappuccino!
- Don't worry, the cosine distance between them is so small that they are almost the same thing.



I would like to open account

Cos sim.

I need new account - Account 0.7

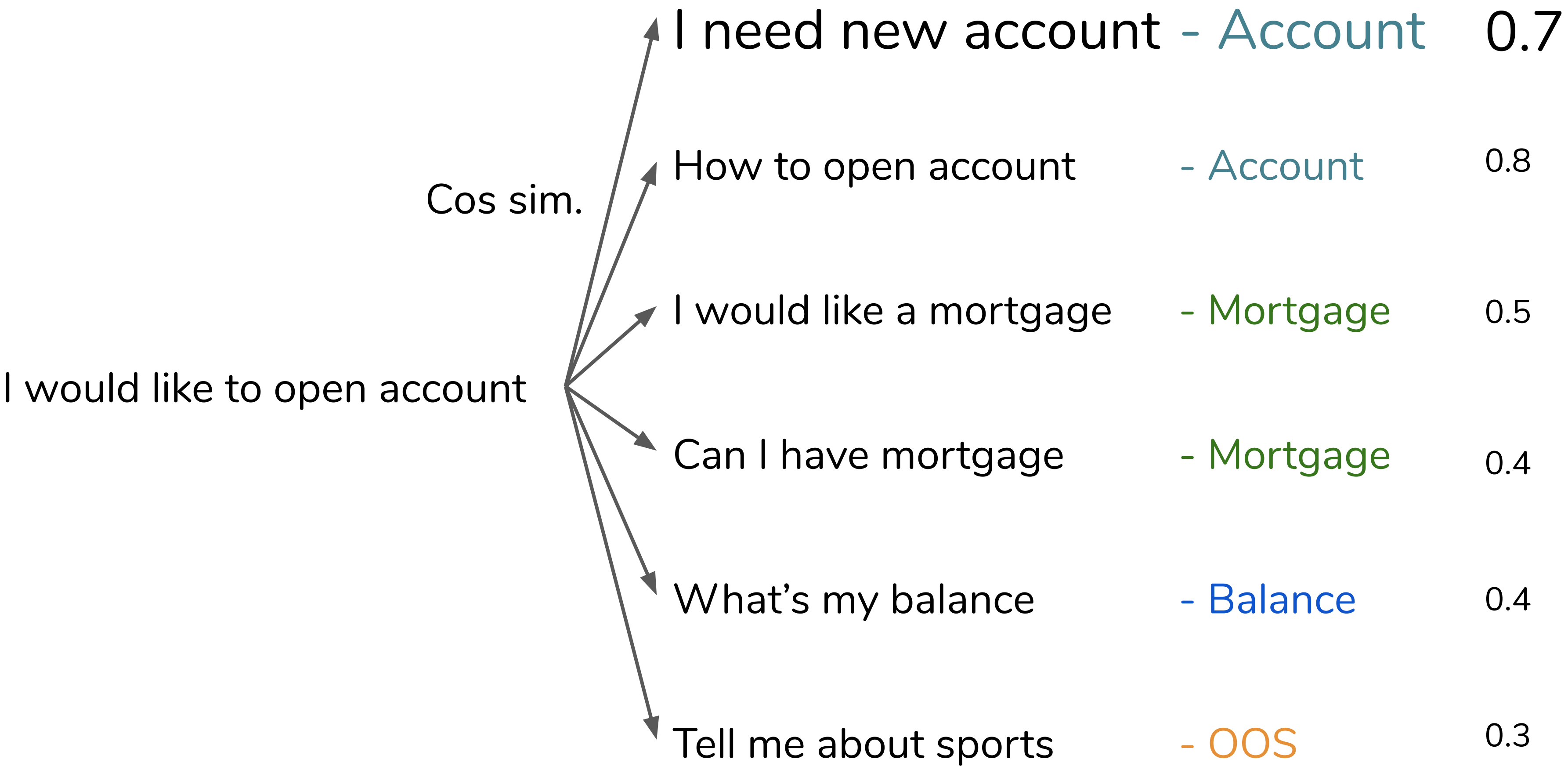
How to open account - Account 0.8

I would like a mortgage - Mortgage 0.5

Can I have mortgage - Mortgage 0.4

What’s my balance - Balance 0.4

Tell me about sports - OOS 0.3

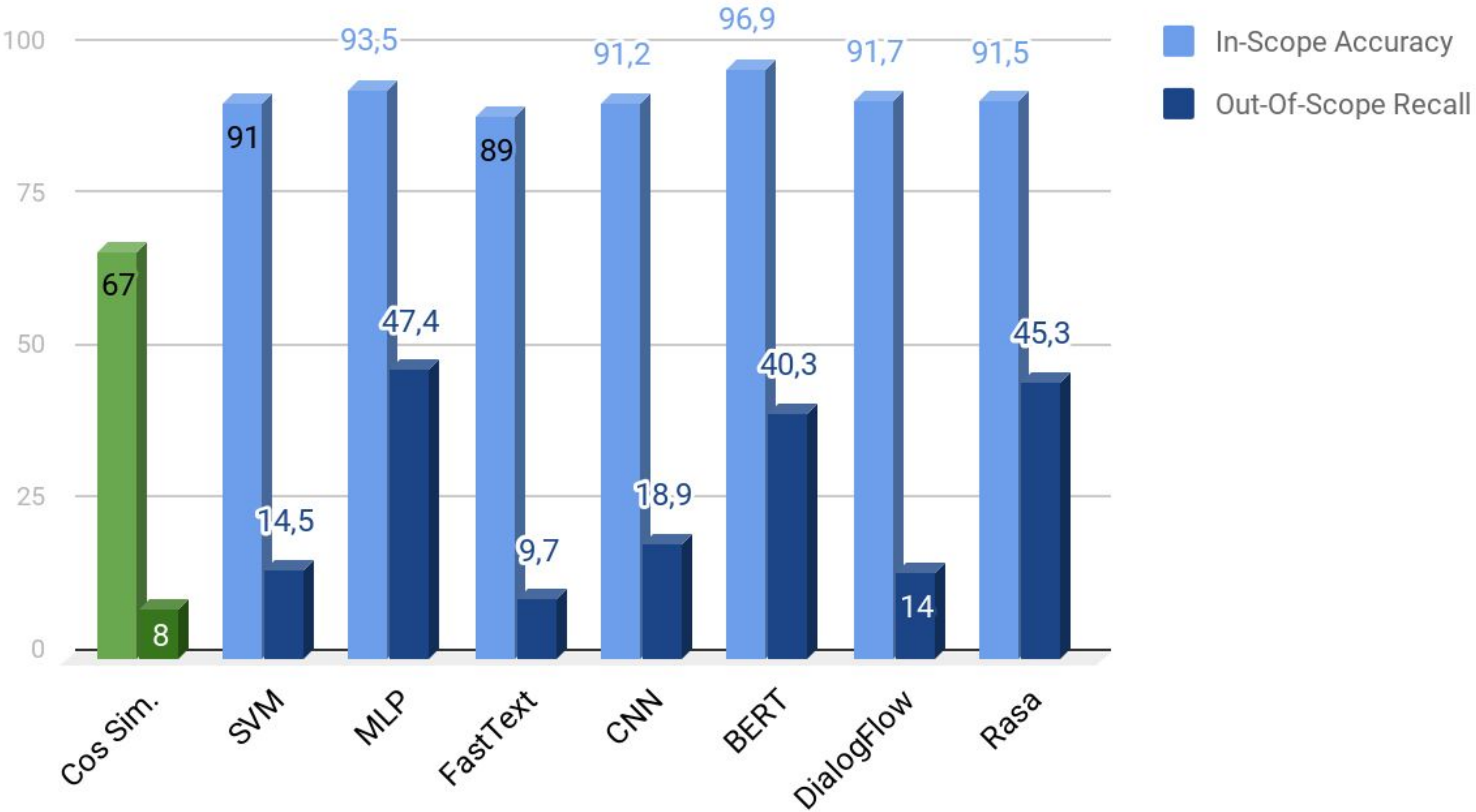


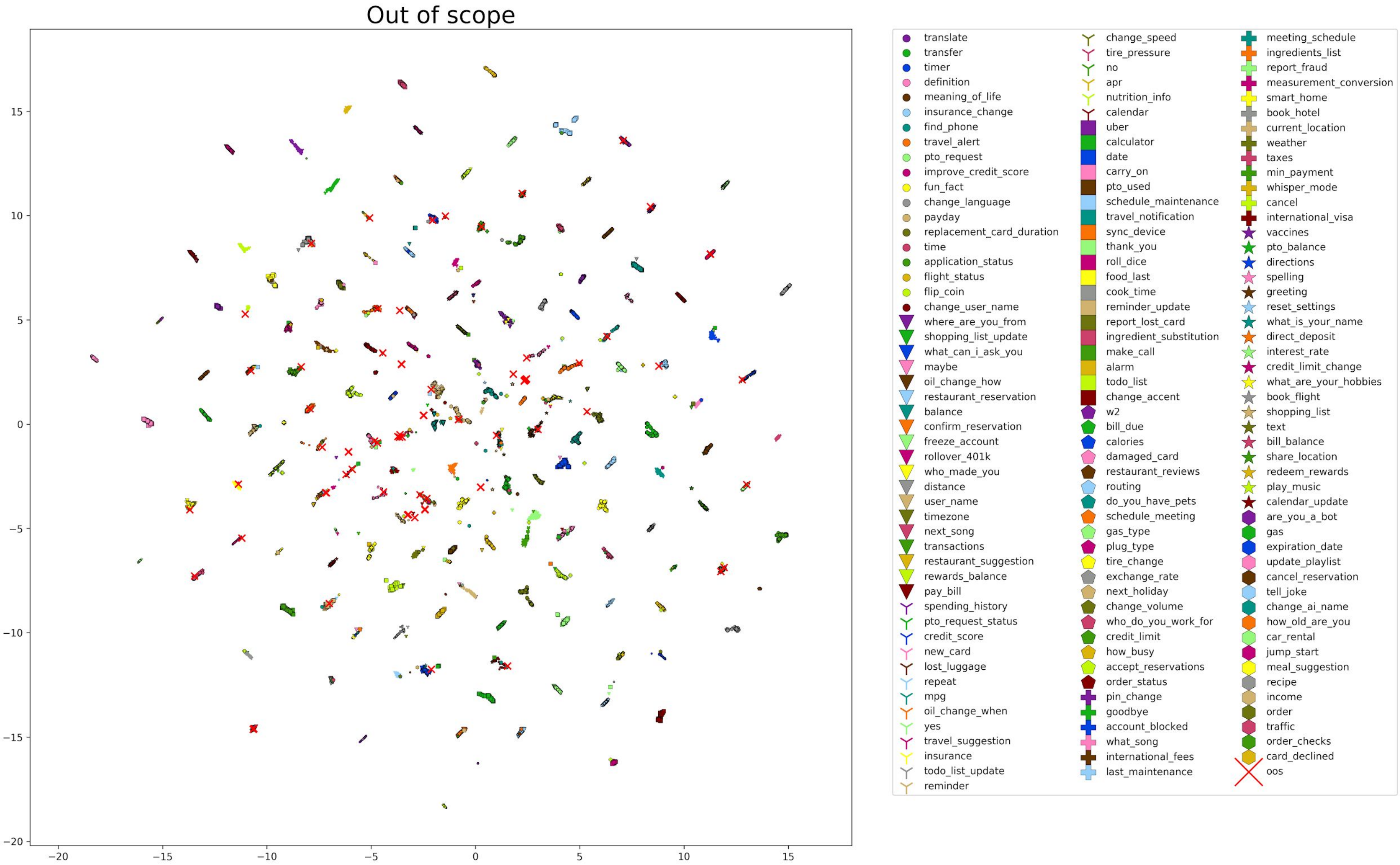
I would like to open account

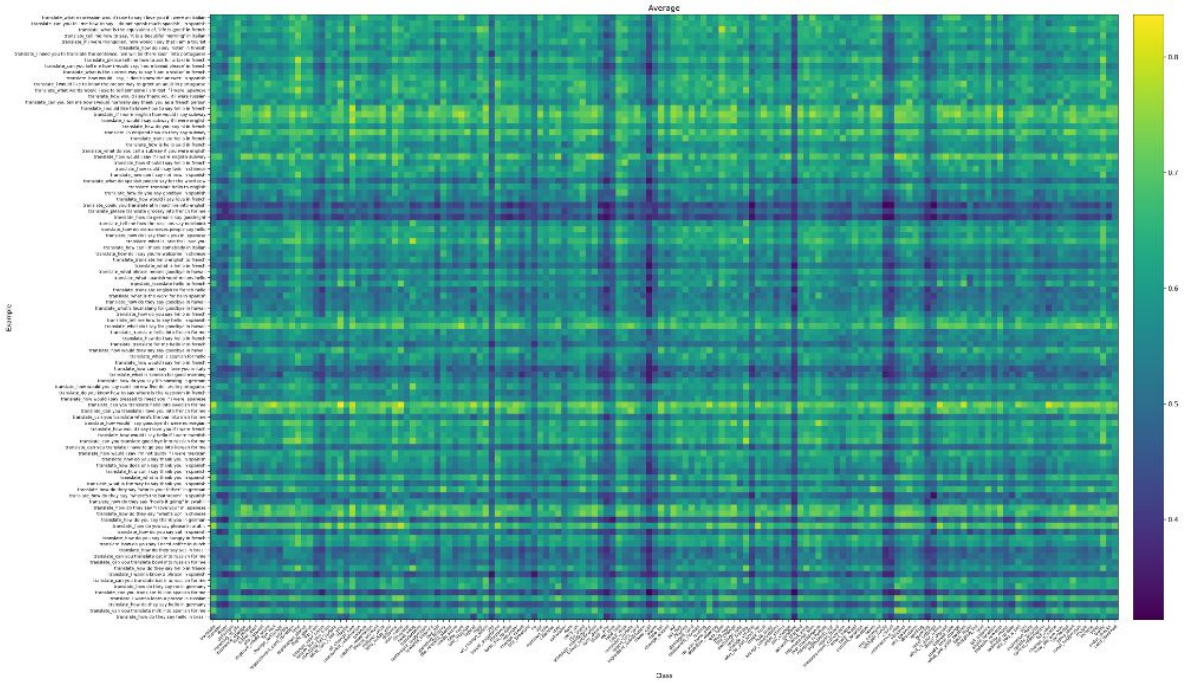
Cos sim.

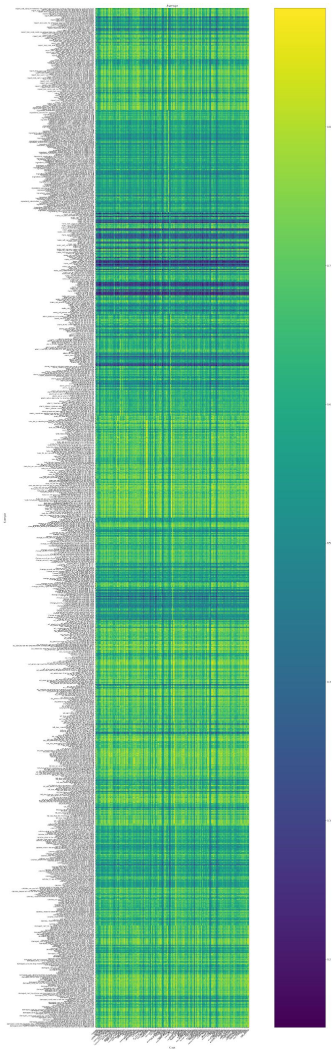
| | | |
|---|---|---|
| I need new account | - Account | 0.7 |
| How to open account | - Account | 0.8 |
| I would like a mortgage | - Mortgage | 0.5 |
| Can I have mortgage | - Mortgage | 0.4 |
| What’s my balance | - Balance | 0.4 |
| Tell me about sports | - OOS | 0.3 |

# OOS-Train


100
75
50
25
0
In-Scope Accuracy
Out-Of-Scope Recall
67
8
91
14,5
93,5
47,4
89
9,7
91,2
18,9
96,9
40,3
91,7
14
91,5
45,3
Cos Sim.
SVM
MLP
FastText
CNN
BERT
DialogFlow
Rasa


Out of scope
15
10
5
0
−5
−10
−15
−20
−20
−15
−10
−5
0
5
10
15
translate
transfer
timer
definition
meaning_of_life
insurance_change
find_phone
travel_alert
pto_request
improve_credit_score
fun_fact
change_language
payday
replacement_card_duration
time
application_status
flight_status
flip_coin
change_user_name
where_are_you_from
shopping_list_update
what_can_i_ask_you
maybe
oil_change_how
restaurant_reservation
balance
confirm_reservation
freeze_account
rollover_401k
who_made_you
distance
user_name
timezone
next_song
transactions
restaurant_suggestion
rewards_balance
pay_bill
spending_history
pto_request_status
credit_score
new_card
lost_luggage
repeat
mpg
oil_change_when
yes
travel_suggestion
insurance
todo_list_update
reminder
change_speed
tire_pressure
no
apr
nutrition_info
calendar
uber
calculator
date
carry_on
pto_used
schedule_maintenance
travel_notification
sync_device
thank_you
roll_dice
food_last
cook_time
reminder_update
report_lost_card
ingredient_substitution
make_call
alarm
todo_list
change_accent
w2
bill_due
calories
damaged_card
restaurant_reviews
routing
do_you_have_pets
schedule_meeting
gas_type
plug_type
tire_change
exchange_rate
next_holiday
change_volume
who_do_you_work_for
credit_limit
how_busy
accept_reservations
order_status
pin_change
goodbye
account_blocked
what_song
international_fees
last_maintenance
meeting_schedule
ingredients_list
report_fraud
measurement_conversion
smart_home
book_hotel
current_location
weather
taxes
min_payment
whisper_mode
cancel
international_visa
vaccines
pto_balance
directions
spelling
greeting
reset_settings
what_is_your_name
direct_deposit
interest_rate
credit_limit_change
what_are_your_hobbies
book_flight
shopping_list
text
bill_balance
share_location
redeem_rewards
play_music
calendar_update
are_you_a_bot
gas
expiration_date
update_playlist
cancel_reservation
tell_joke
change_ai_name
how_old_are_you
car_rental
jump_start
meal_suggestion
recipe
income
order
traffic
order_checks
card_declined
oos


Average
Example
Class
0.8
0.7
0.6
0.5
0.4


Average