ConveRT: Efficient and Accurate Conversational Representations from Transformers

Matt Henderson
PolyAI

Polish Natural Language Processing Meetup

ConveRT: Efficient and accurate conversational representations from transformers
M Henderson, I Casanueva, N Mrkšić, PH Su, I Vulić

A Repository of Conversational Datasets
M Henderson, P Budzianowski, I Casanueva, S Coope, D Gerz, G Kumar, N Mrkšić, G Spithourakis, PH Su, I Vulić, TH Wen

Efficient Intent Detection with Dual Sentence Encoders
I Casanueva, T Temčinas, D Gerz, M Henderson, I Vulić

Span-ConveRT: Few-shot Span Extraction for Dialog with Pretrained Conversational Representations
S Coope, T Farghly, D Gerz, I Vulić, M Henderson

PolyAI

- Dialogue Systems group, Cambridge
- Automated voice agents for customer services, call centres
- Restaurant booking, internet troubleshooting, banking, ...

Creating Task-based Dialogue Systems

Convincing Application

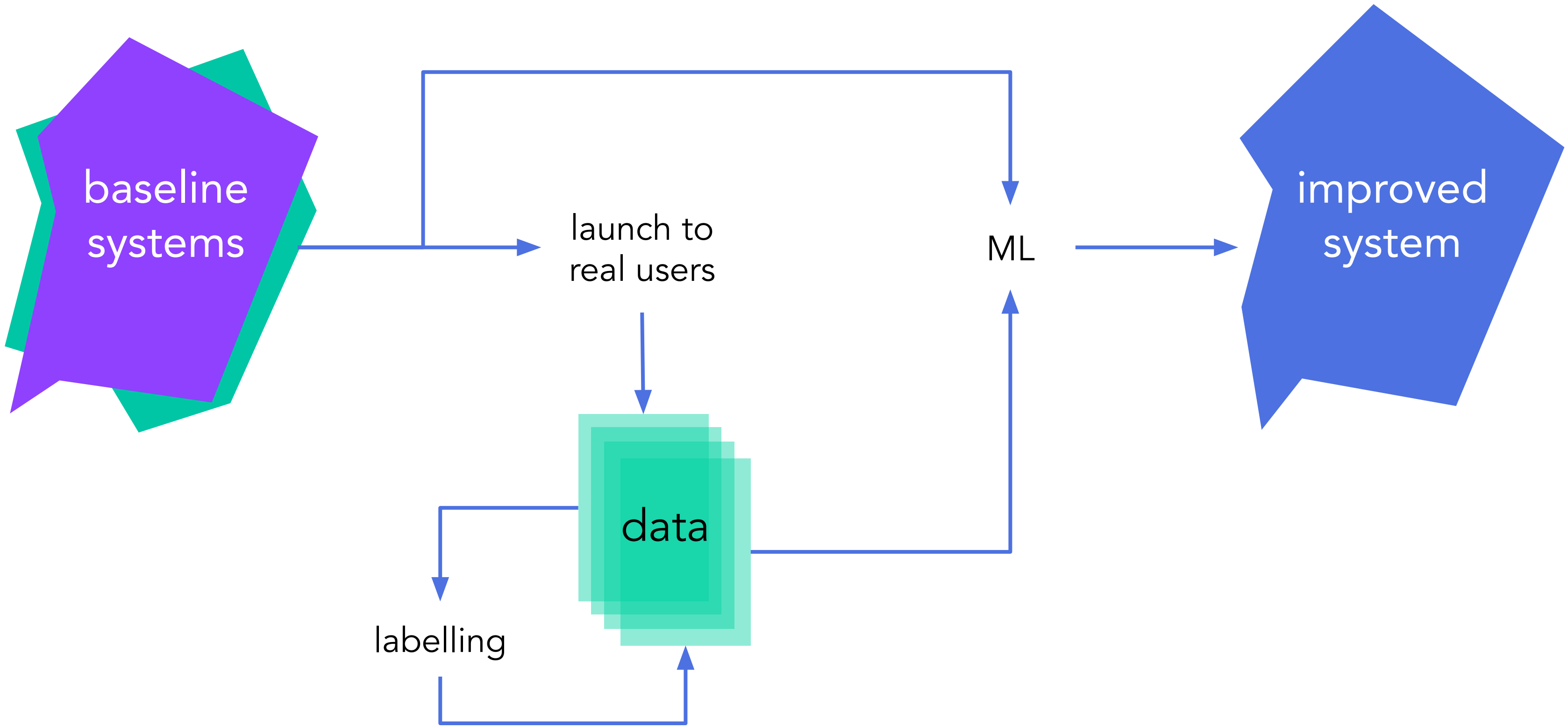
solves a real problem

Meaningful Evaluation

can measure progress

Annotated Data

is machine-learnable


baseline systems
launch to real users
ML
improved system
data
labelling

how do we get a
baseline system?

intent classifiers
slot-value recogniser
response selection/generation

xskills
xdomains
xlanguages

how can we minimise reliance on annotated data?

how can we scale better?
(skills, domains, languages…)

by using large pre-trained models that encapsulate knowledge of conversational response

Pre-training in NLP

- recent trend to pre-train large models of language, then fine-tune

  BERT, ELMo, GPT etc.

- uses unlabelled text + unsupervised objective

  same idea as cbow, skip gram, skip thought etc.

- learns general representations of text, useful for downstream tasks

PolyAI Conversational Datasets

Reddit	OpenSubtitles	AmazonQA
3.7 billion comments from online discussions on many topics	over 400 million lines of subtitles from movies and TV	over 3.6 million product question-answer pairs
↓	↓	↓
727 million examples	316 million examples	3.6 million examples

github.com/PolyAI-LDN/conversational-datasets

Public Conversational Datasets

	~ Turns	Annotations
DSTC 2&3	10^4	response, ASR, SLU
MultiWoz	10^5	response, NLU
DSTC7 Reddit	10^6	response, entities
DSTC7 Ubuntu	10^6	response
PolyAI AmazonQA	10^6	product, response
PolyAI OpenSubtitles	10^8	'response'
PolyAI Reddit	10^9	response

Next word prediction

The launch of India's second lunar mission has been → halted

apple
called
halted
celebrate
passport
...

Masked word prediction

The launch of ■ 's second lunar mission has been
???
less than an hour before the scheduled blast- ■ ,
due to a ■ problem.

apple
called
halted
celebrate
passport
...

Response Selection

Any recommendations for short trips from Singapore?

→

It doesn't feel like July.
That type of music isn't really my cup of tea.
Bintan is just a quick ferry trip away.
You have to try the vegetarian Haggis!
I'd do a short trip to Paris.
...

Response Selection

- large conversational datasets

Language Modelling

- large text datasets

Response Selection

- large conversational datasets
- representations encode conversational cues

Language Modelling

- large text datasets
- representations encode word/phrase/sentence cues

Response Selection

- large conversational datasets
- representations encode conversational cues
- encodes full sentences

Language Modelling

- large text datasets
- representations encode word/phrase/sentence cues
- encodes words contextually

Response Selection

- large conversational datasets
- representations encode conversational cues
- encodes full sentences
- directly applicable to retrieval-based dialogue

Language Modelling

- large text datasets
- representations encode word/phrase/sentence cues
- encodes words contextually
- maybe applicable to generation/scoring

a lot of the power of neural techniques is finding good embeddings / encodings

- so learn encoder model on large conversational data

- then use various tricks and small models on the learned vector space for domain specific tasks

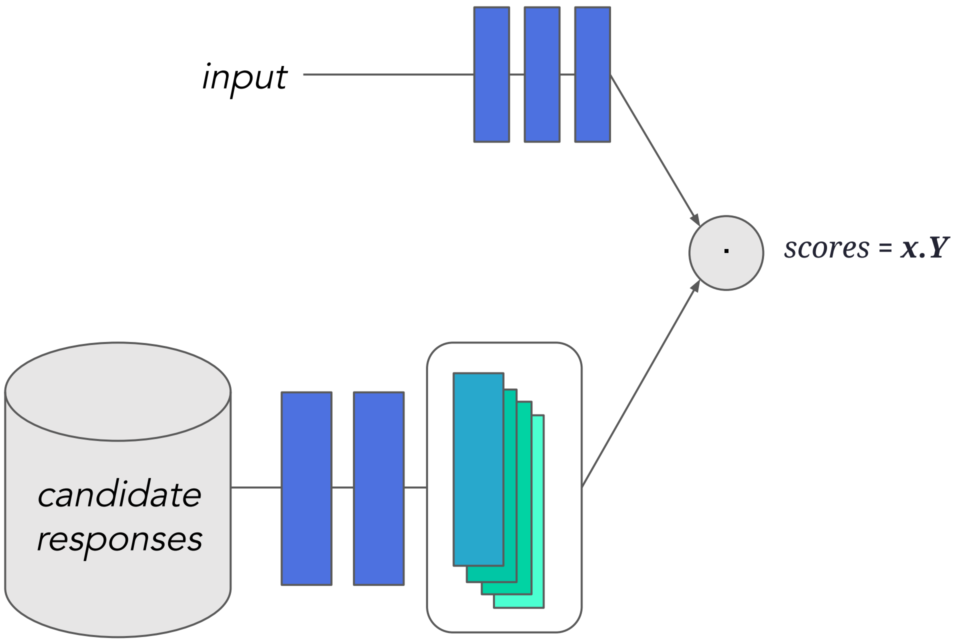
Dual Encoders for Response Selection

dual encoder dot product model

- gmail smart reply
- universal sentence encoder

trained to give a high score for the response found in the data, low score for random responses

final score of an input and response is a dot-product of two vectors



network encodes a batch of inputs to vectors:

$$x_1 \quad x_2 \quad \dots \quad x_N$$

and responses to vectors:

$$y_1 \quad y_2 \quad \dots \quad y_N$$

$x_1.y_1$	$x_1.y_2$	$x_1.y_3$	$x_1.y_4$	$x_1.y_5$
$x_2.y_1$	$x_2.y_2$	$x_2.y_3$	$x_2.y_4$	$x_2.y_5$
$x_3.y_1$	$x_3.y_2$	$x_3.y_3$	$x_3.y_4$	$x_3.y_5$
$x_4.y_1$	$x_4.y_2$	$x_4.y_3$	$x_4.y_4$	$x_4.y_5$
$x_5.y_1$	$x_5.y_2$	$x_5.y_3$	$x_5.y_4$	$x_5.y_5$

the N x N matrix of all scores is a fast matrix product.

large improvement in 1 of 100 ranking accuracy over binary classification.

$x_1 . y_1$	$x_1 . y_2$	$x_1 . y_3$	$x_1 . y_4$	$x_1 . y_5$
$x_2 . y_1$	$x_2 . y_2$	$x_2 . y_3$	$x_2 . y_4$	$x_2 . y_5$
$x_3 . y_1$	$x_3 . y_2$	$x_3 . y_3$	$x_3 . y_4$	$x_3 . y_5$
$x_4 . y_1$	$x_4 . y_2$	$x_4 . y_3$	$x_4 . y_4$	$x_4 . y_5$
$x_5 . y_1$	$x_5 . y_2$	$x_5 . y_3$	$x_5 . y_4$	$x_5 . y_5$

$$\boldsymbol{x}_i = f(\text{input } i)$$

$$\boldsymbol{y}_j = g(\text{response } j)$$

$$S_{ij} = \boldsymbol{x}_i \cdot \boldsymbol{y}_j$$

$$P(\text{ response } j \mid \text{ input } i) \propto e^{Sij}$$

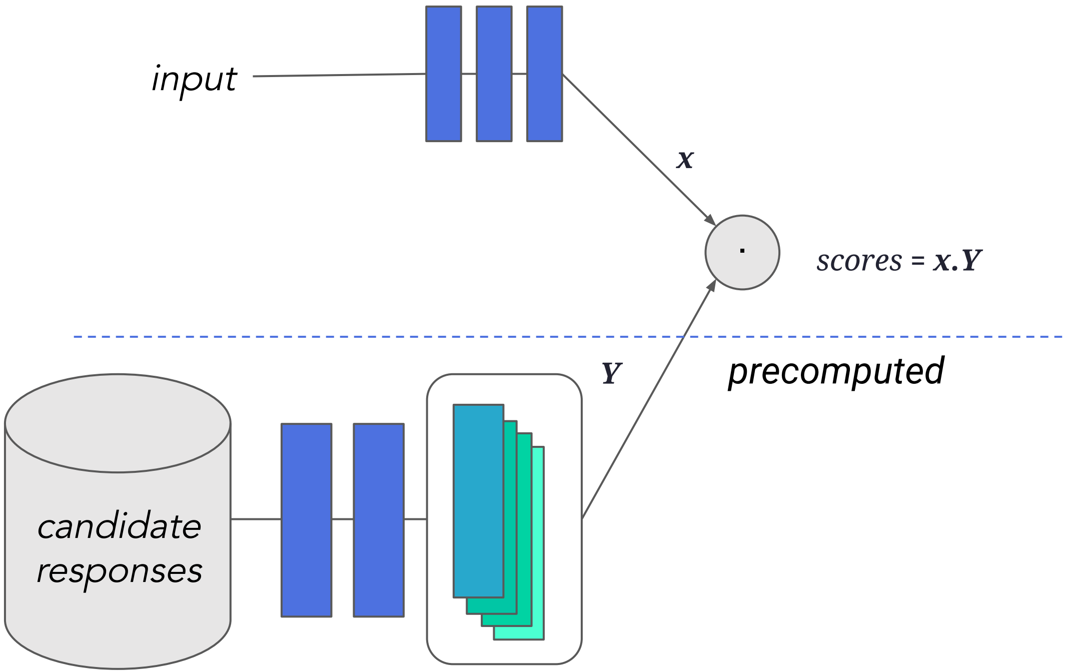
$$-\log P(\text{ example } i) = -S_{ii} + \log \Sigma_j e^{Sij}$$

"dot product loss"

Precomputation for dot product model

the representations of the candidates $\boldsymbol{Y}$ can be precomputed

approximate nearest neighbor search can speed up the top N search



at inference, a user query has N words, there are M responses with N_R words each

- dot product model
    - $O(N)$ to encode input to vector space
    - $O(\log M)$ to find top scoring response with approximate search

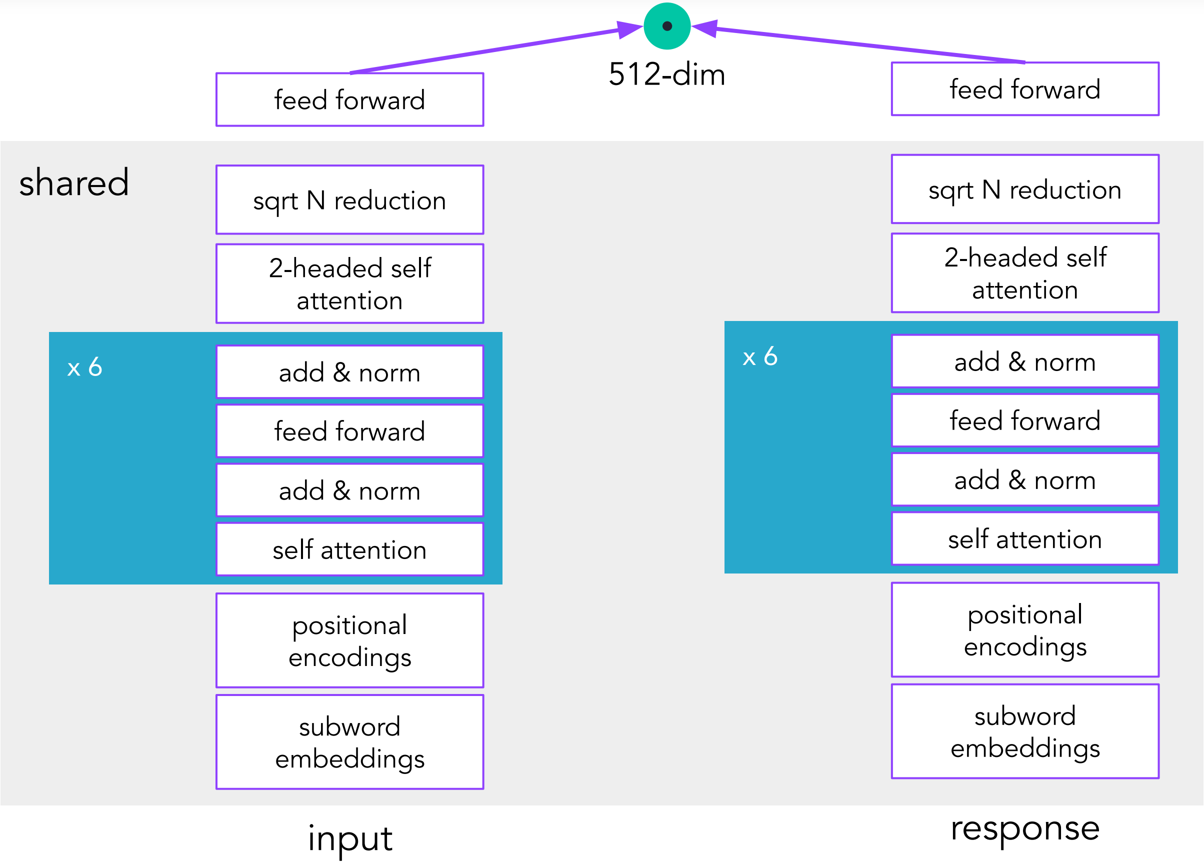
at inference, a user query has N words, there are M responses with N_R words each

- dot product model
  - $O(N)$ to encode input to vector space
  - $O(\log M)$ to find top scoring response with approximate search
- general sequence model (e.g. BERT next sentence scoring)
  - $O(M(N + N_R))$ to score all responses
  - $O(M)$ to find top response

1-of-100 accuracy

how often the correct response is ranked top vs 99 random

ConveRT - Conversational Representations from Transformers



ConveRT - Conversational Representations from Transformers

		reddit 1-of-100 accuracy
keyword-based	TF-IDF	26.7%
	BM25	27.6%
MAP dot product models	ELMo	19.3%
	BERT	24.5%
	USE	40.8%
	USE_QA	46.3%
	BERT dot-product model	55.0%
PolyAI	ConveRT	**68.2%**

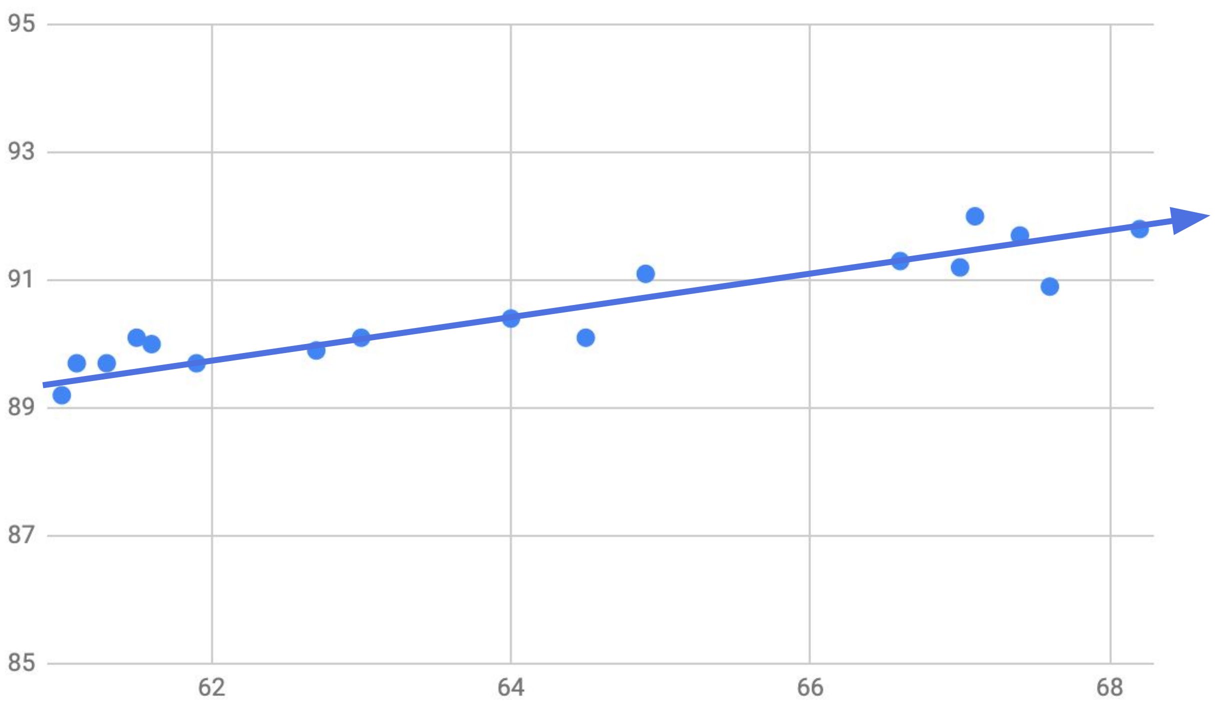
ConveRT

resource-constrained optimization:
pick the best model after training 18 hours on 12 GPUs

- fast ML engineering cycle, rapid progress
- we own the whole training pipeline

- training costs under $100
- model runs fine on CPU
- final model is 40MB

task-based
accuracy
(no fine-tuning)



reddit 1-of-100
(progress over 3 weeks)

intent classification

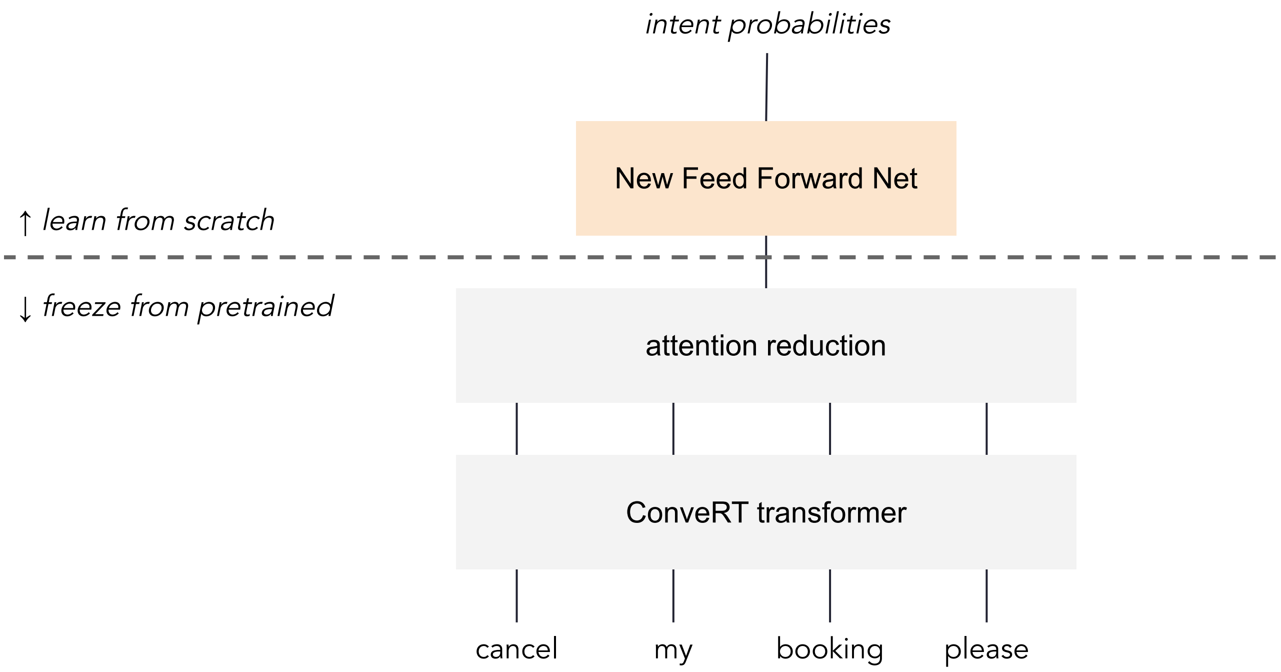
Intent Classification

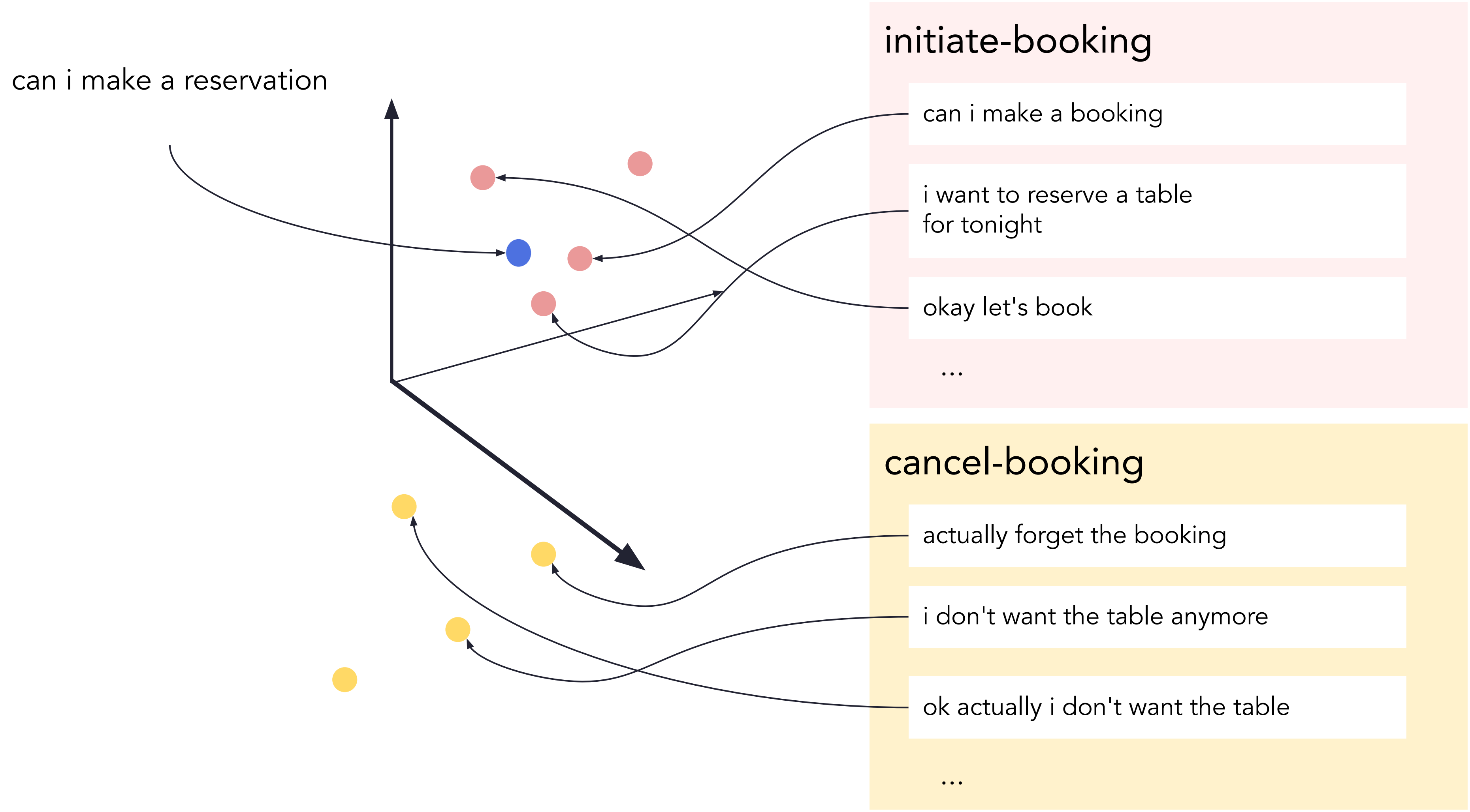
initiate-booking
- can i make a booking
- can i reserve a table
- okay i want to book a table for tonight

cancel-booking
- cancel it
- i don't want the table anymore

restart
- let's start over
- forget this

Recap - Intent Classification




can i make a reservation
initiate-booking
can i make a booking
i want to reserve a table for tonight
okay let's book
...
cancel-booking
actually forget the booking
i don't want the table anymore
ok actually i don't want the table
...

Intent Evaluation

Model	BANKING77			CLINC150			HWU64		
	10	30	Full	10	30	Full	10	30	Full
BERT-FIXED	67.55	80.07	87.19	80.16	87.99	91.79	72.61	79.78	85.77
BERT-TUNED	83.42	90.03	**93.66**	91.93	95.49	96.93	84.86	88.27	92.10
USE	84.23	89.74	92.81	90.85	93.98	95.06	83.75	89.03	91.25
CONVERT	83.32	89.37	93.01	92.62	95.78	**97.16**	82.65	87.88	91.24
USE+CONVERT	**85.19**	**90.57**	93.36	**93.26**	**96.13**	**97.16**	**85.83**	**90.16**	**92.62**

Efficient Intent Detection with Dual Sentence Encoders

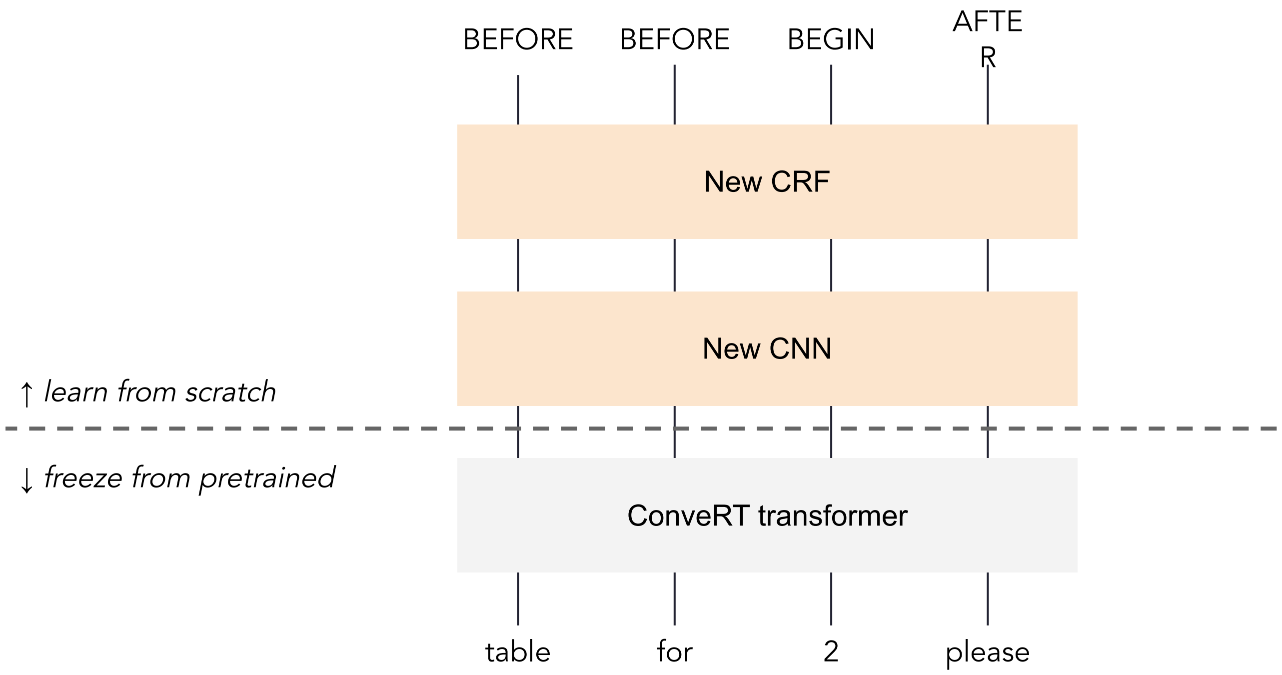
I Casanueva, T Temčinas, D Gerz, M Henderson, I Vulić

value extraction

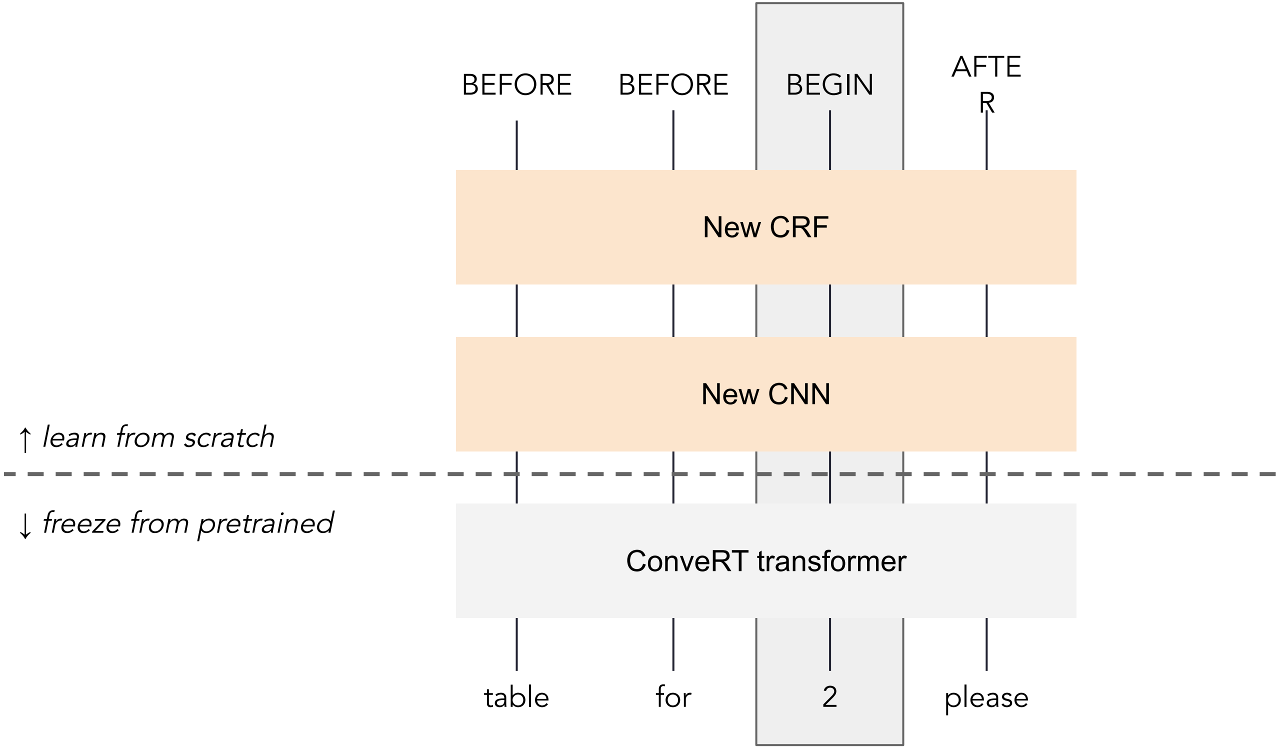
Value Extraction / Slot Filling

please book Harry's bar at 9 pm	INITIATE_BOOKING
2 people under the name Henderson	INFORM
do you do christmas bookings yet	CHRISTMAS_BOOKINGS

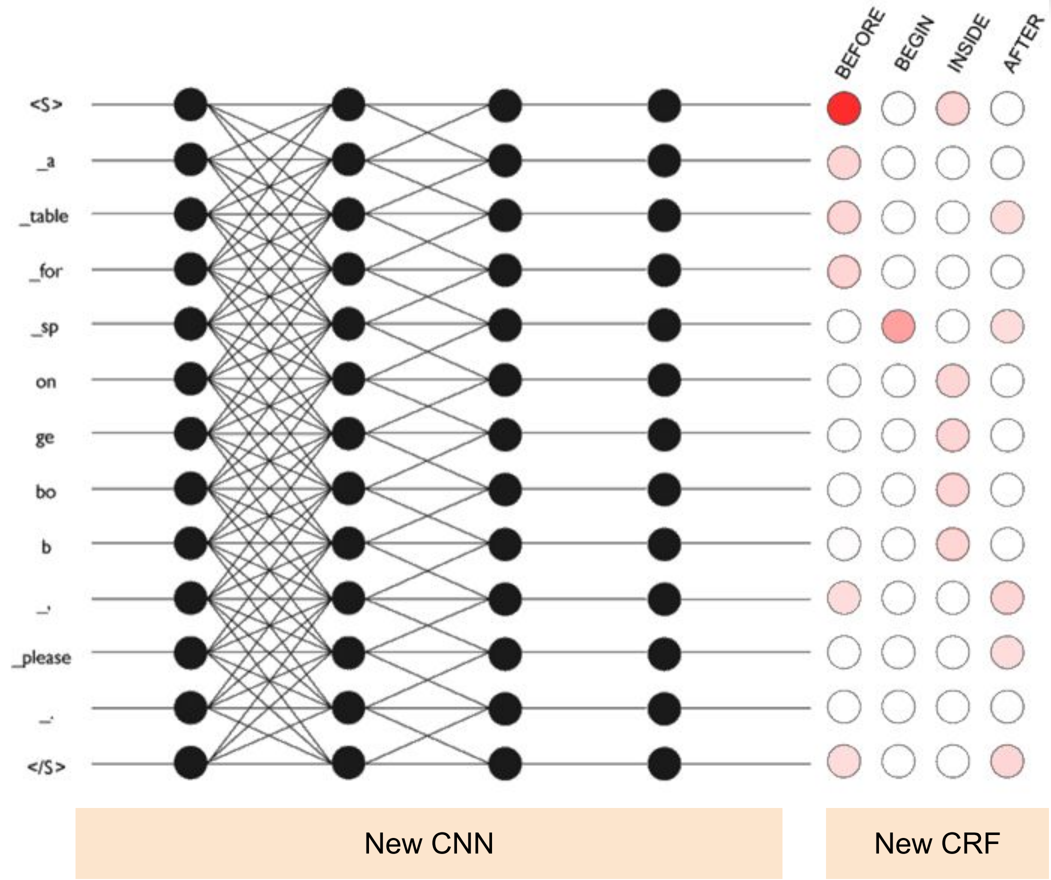
SpanConveRT



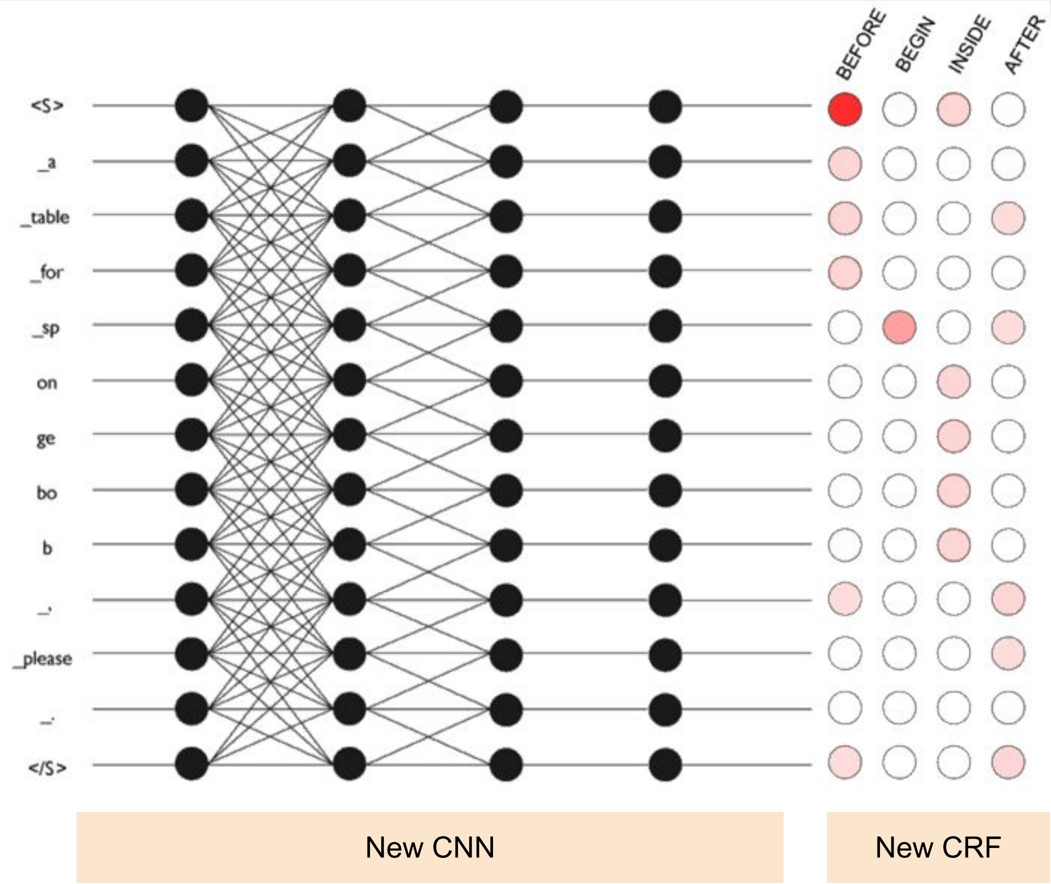
SpanConveRT



SpanConveRT



SpanConveRT



Value Extraction Evaluation

Fraction	Span-ConveRT	V-CNN-CRF	Span-BERT
1 (8198)	**0.96**	0.94	0.93
1/2 (4099)	**0.94**	0.92	0.91
1/4 (2049)	**0.91**	0.89	0.88
1/8 (1024)	**0.89**	0.85	0.85
1/16 (512)	**0.81**	0.74	0.77
1/32 (256)	**0.64**	0.57	0.54
1/64 (128)	**0.58**	0.37	0.42
1/128 (64)	**0.41**	0.26	0.30

Span-ConveRT: Few-shot Span Extraction for Dialog with Pretrained Conversational Representations
S Coope, T Farghly, D Gerz, I Vulić, M Henderson

ConveRT

efficient task tailored to dialogue

smaller cheaper faster models

robust performance on downstream tasks

competitive NLU accuracy

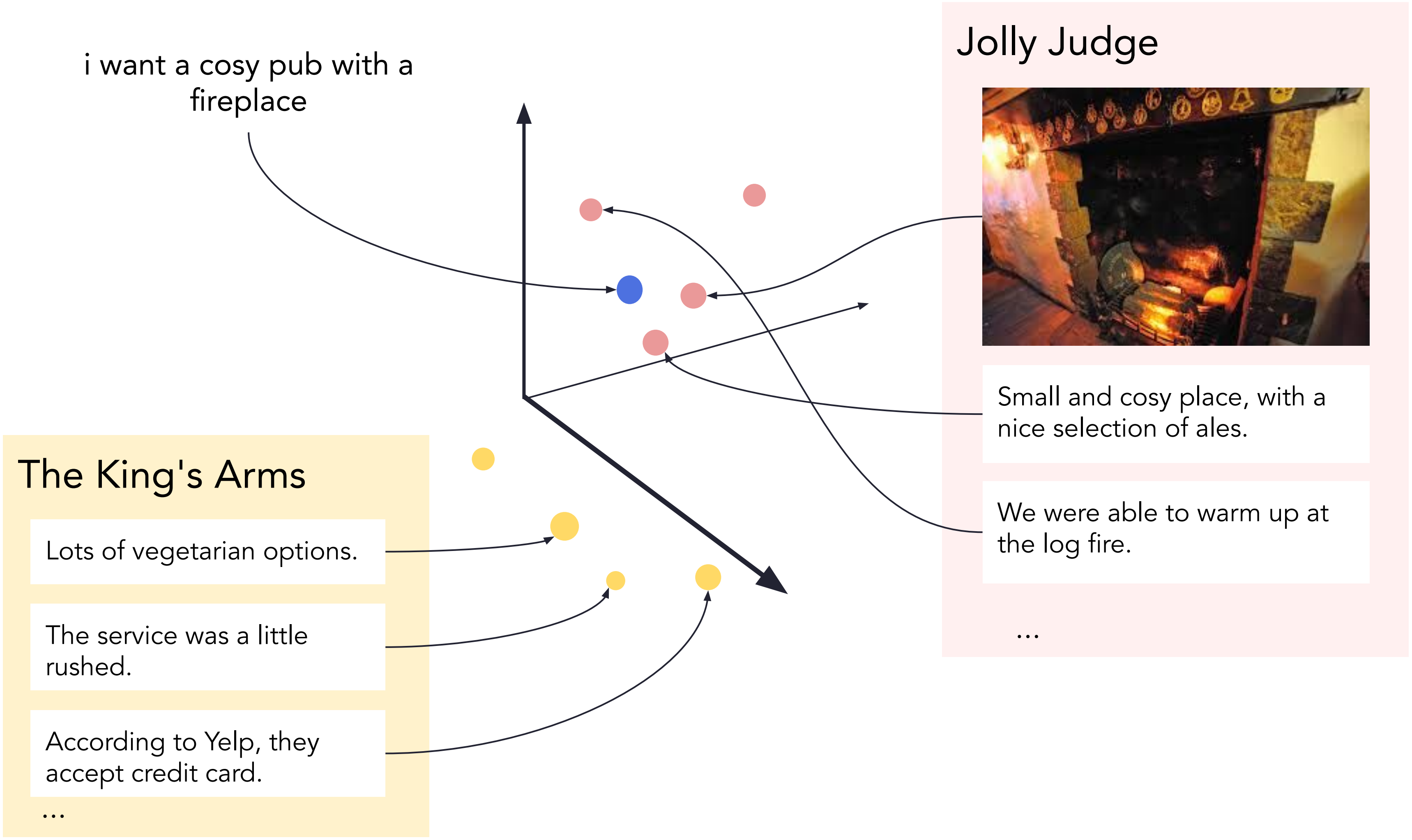
powers conversational search

efficient search

reduced dependency on strict ontology

PolyAI

find your voice


i want a cosy pub with a fireplace
Jolly Judge
Small and cosy place, with a nice selection of ales.
We were able to warm up at the log fire.
...
The King's Arms
Lots of vegetarian options.
The service was a little rushed.
According to Yelp, they accept credit card.
...