Deep Learning in Domain Scaling for Conversational Agents

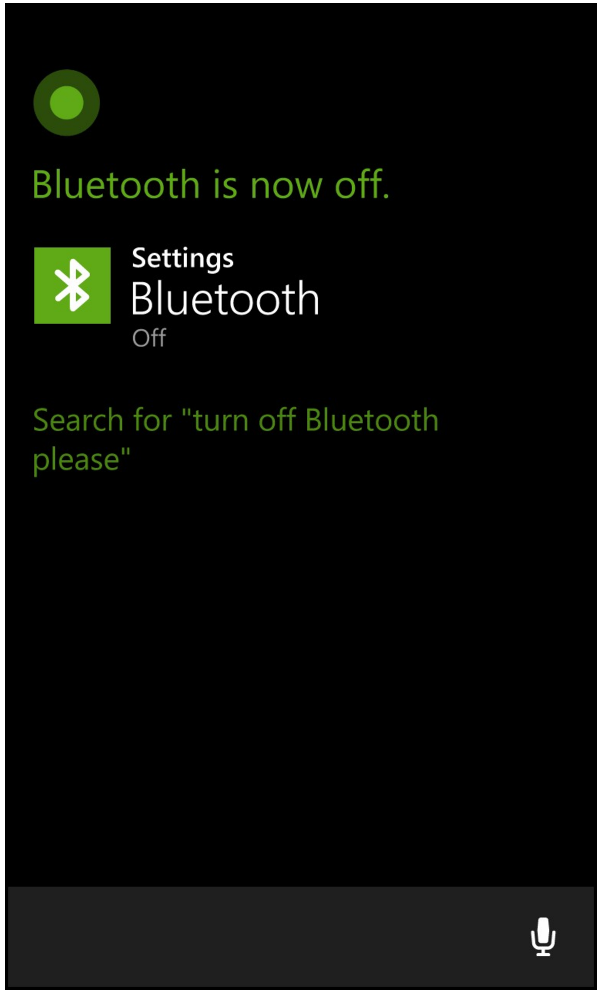
Ye-Yi Wang

In collaboration with Bin Cao, Jianfeng Gao, Ruhi Sarikaya, Gokhan Tur, Asli Celikyilmaz, Bing Intent Science Team and MSR Deep Learning Center

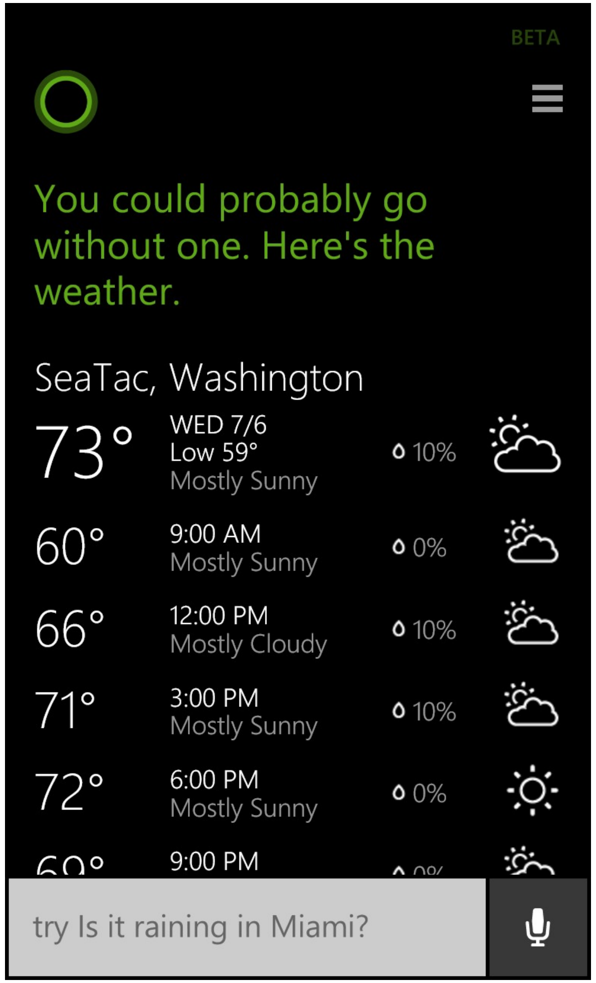
Growing with interACT

Thanks for leading the community to shape the reality
Looking forward to continued leadership in shaping the future

Cortana: Task Completion & QA

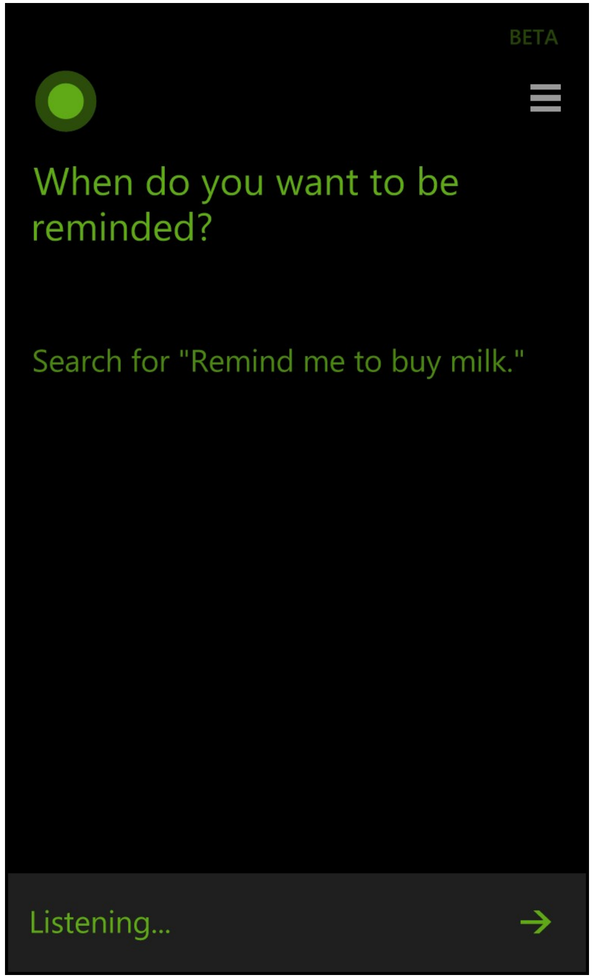


Turn off Bluetooth please

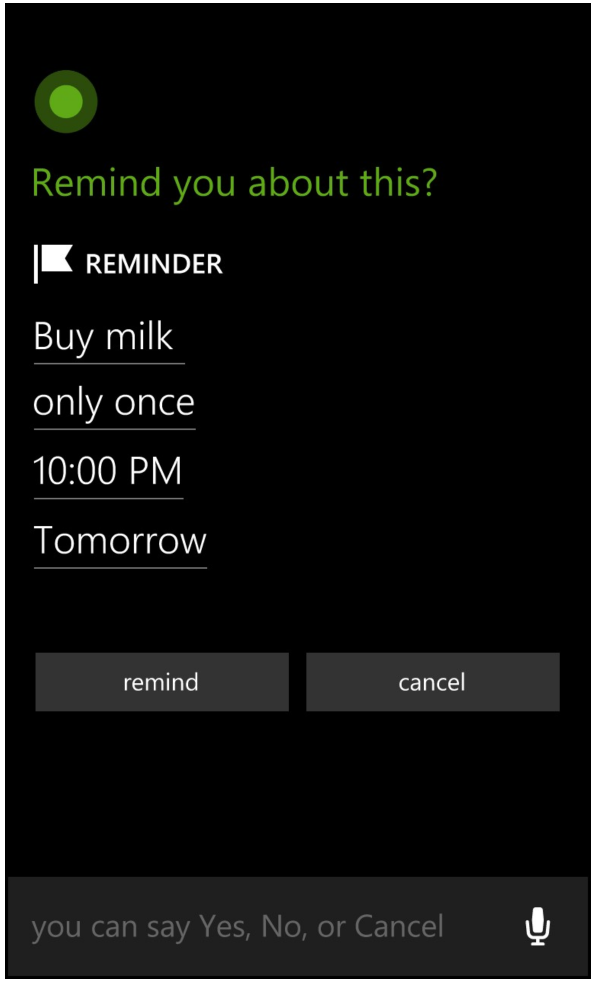


Do I need a jacket today?

Cortana: Multi-turn Conversations



Remind me to buy milk



10 Pm tomorrow

Cortana: Language Understanding

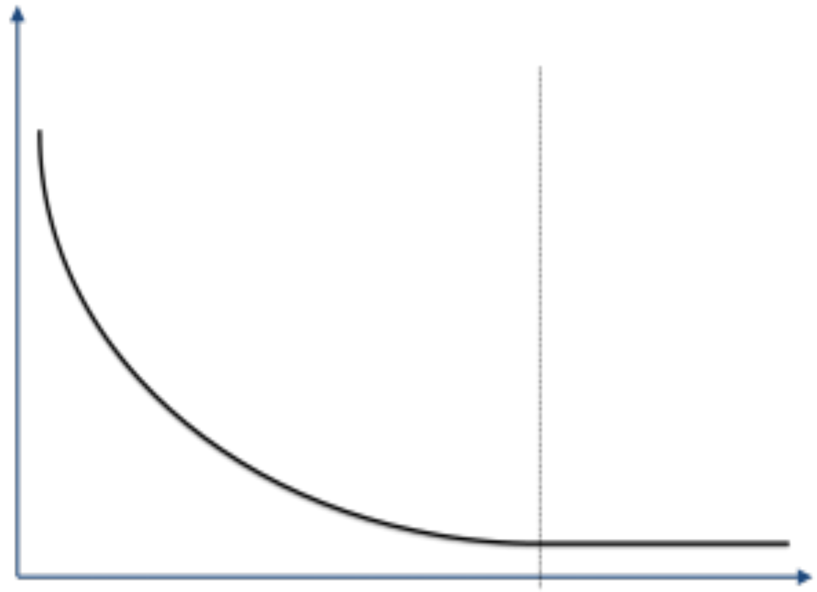
- What is "Understanding"?
  - Explicit or implicit? Generic or domain specific
  - Practical solution: Query → Semantic Frame
- Semantic Frame: structured meaning representation
  - Domain (Weather, Device Control, Play Music, …) SVMs
  - Intent (5 day forecast, Get temperature, …) SVMs
  - Slots (e.g., weather in <loc>**Boston**</loc>) CRFs
- Model Training
  - Domain by domain, locale by locale
  - Annotators provide labeled data for initial coldstart model training
  - Annotators label the feedback data after deployment for continuous improvement
  - Hard to scale

Cortana: Dialog Modeling

- 1st generation (past): manually designed finite state dialog flow/policy
- 2nd generation (now): a platform that hides the complexity of flow design, fixed dialog policy
- 3rd generation (future): deep reinforcement learning for dialog policy learning/tuning.

Why Language Understanding is hard

- Ambiguity
- Power Law



“there is no data like more data”
“data is the new oil, intelligence is the new power”

The Language Understanding Scaling Problem

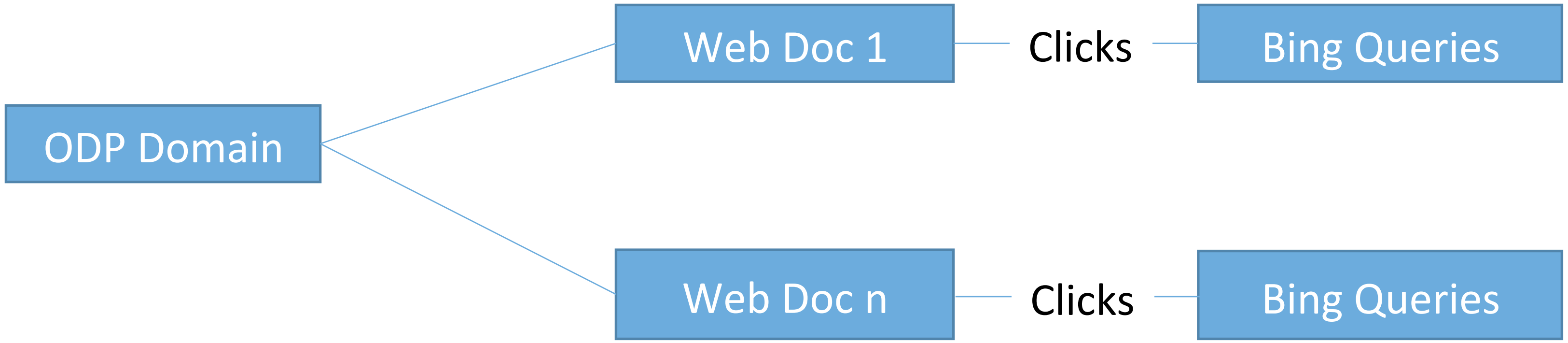
- Domain scaling: a demand/supply problem of supervision data
- Increase the supply: Automatic offline data labeling & feedback loop
  - Multi-task deep learning for domain classification against an existing taxonomy (ODP)
  - HITS and EM algorithm for entity tagging
  - Feedback loop
- Reduce the demand
  - Features with better generalization capability (Multi-task embedding learning)
  - Models that generalize better (LSTM, Seq2Seq)

Increase the Supply

Tools for users to select from pre-labeling big data via semi-supervised or unsupervised learning

Semi-supervised/Unsupervised Labeling of Big Data

- Classification with weak supervisions

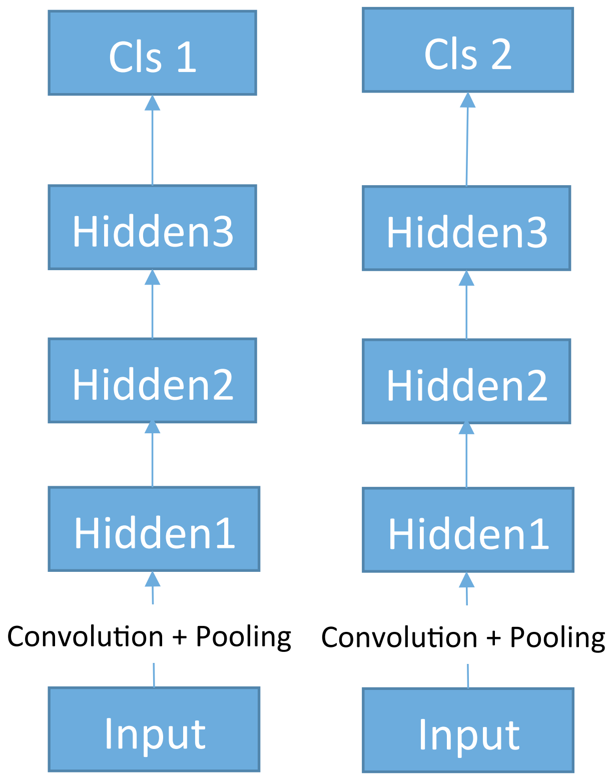


- Slot tagging with EM algorithms + Knowledge base
  - Substring match against entities in the knowledge base
  - Disambiguation via pattern statistics (contextual dependency)
  - Iteratively repeated the process (EM algorithm)
  - Initialize EM with HITS algorithm

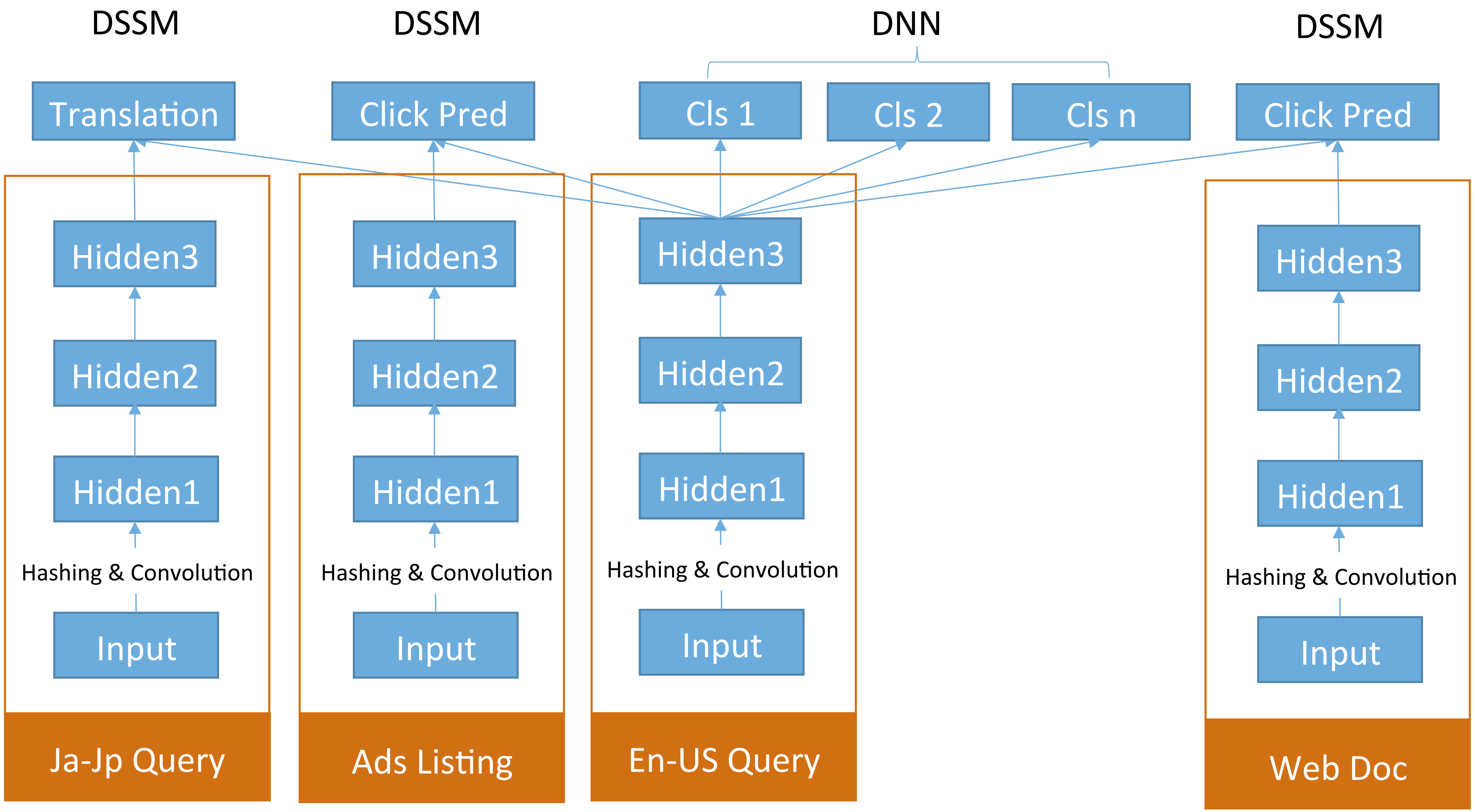
Tools for data selection

- Browsable – organized according to a give domain taxonomy (ODP) and (finer-grained) clusters from topic modeling
- Searchable – semantic similarity ranking based on query embedding

Multi-Task Deep Learning



Multi-Task Deep Learning



Multi-Task Deep Learning: learn generic semantics

- DNN/DSSM based multi-task learning has been applied to domain classification in IntentExplorer
- Significant improvement on Ads team's ODP experiments

	Avg AUC	Top1 Accuracy	Top5 Accuracy	Top10 Accuracy
MT-DNN	95.0%	51.3%	82.4%	89.5%
SVMs	95.6%	41.1%	72.3%	83.6%

- However, query level embedding doesn't help slot tagging

Reduce the Demand

Embedding as features for better generalization

Embedding learning for Cold Start LU
Reducing the Demand on Labeled Data

Domain	Baseline	Baseline	LSTM + Embedding
	(Production Model)	(SVM + Ngram)	
alarm	0.999	0.997	0.9995
calendar	0.997	0.992	0.9976
communication	0.996	0.976	0.9958
mediacontrol	0.999		0.9989
mystuff	0.997	0.997	0.9973
note	0.999	0.999	0.9995
ondevice	0.993		0.9944
places	0.989	0.984	0.9885
reminder	0.999	0.979	0.999
weather	0.999	0.998	0.9989
web	0.969	0.941	0.9734
webnavigation		0.998	0.9967

On par performance can be achieved with a fraction of training data for slot tagging

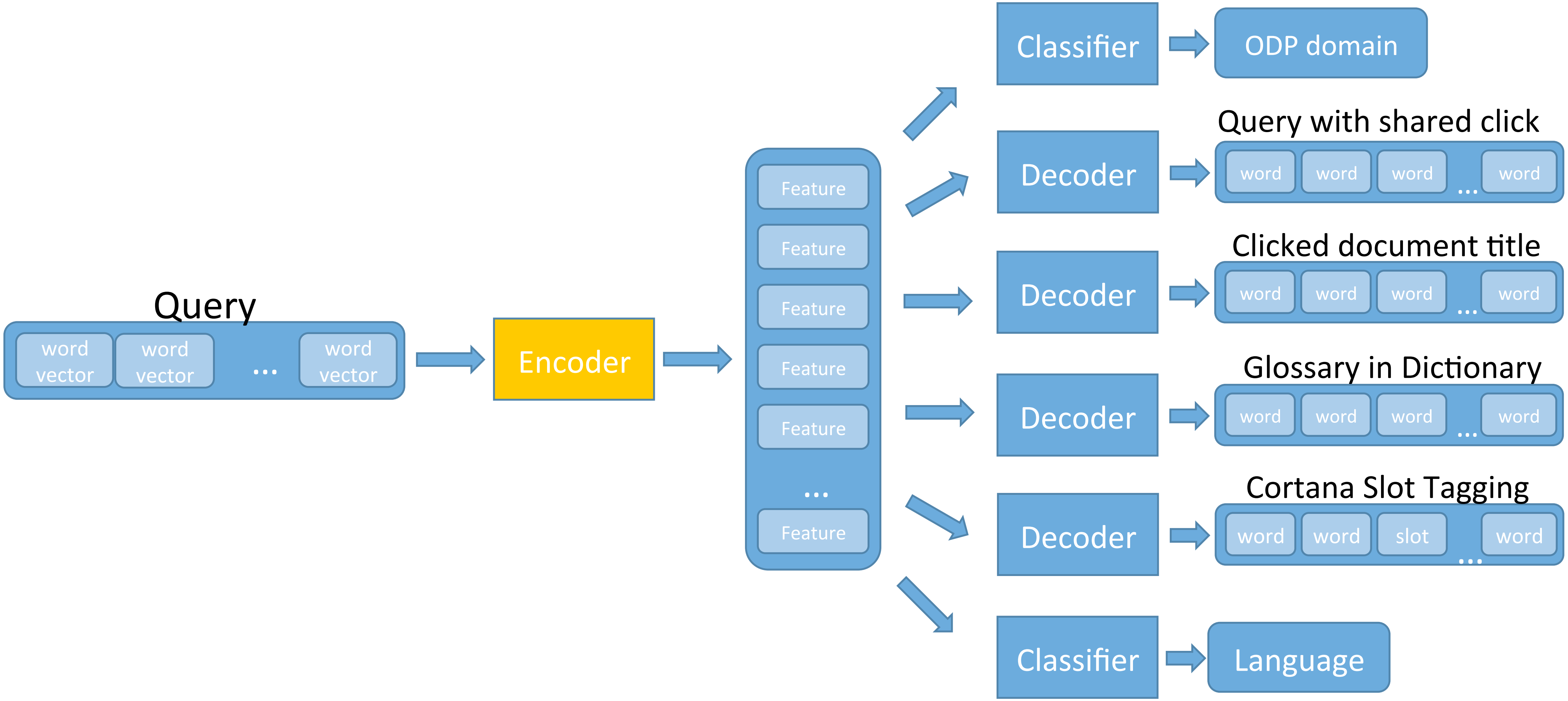
On par performance can be achieved without engineered features for domain classification

Opportunity for Improvement

Using Oracle embedding, the classification results were much better when fraction of training data were used

#Samples	Embedding	Optimal Embedding
494	0.6197	0.8800
1080	0.7581	0.8972
2312	0.8418	0.9139
4974	0.8765	0.9290

Multi-Task Deep Sequence Learning



Summary

- Challenges in scaling up or democratizing the conversational experiences
- The key issue is here is a demand/supply problem
- Increasing demand – auto-labeled data for selection
- Reducing the cost – project into a continuous space via embedding learning for better generalization