Out-of-domain Text Classification

Faculty of Electrical Engineering, CTU in Prague

Tommaso Gargiani
Supervisor: Ing. Petr Lorenc

Introduction to the problem

- Intent classification is crucial to conversational AI
- Terminology
  - in-domain (ID), in-scope – sentence that belongs to a defined intent
  - out-of-domain (OOD), out-of-scope – sentence that does not fall into any supported intent
- Classifiers generally perform very well on in-domain intents
  X they tend to struggle on out-of-domain sentences
- A chatbot that fails to recognise an out-of-domain query will give an unrelated answer, instead of using a fallback response
- Correcting this behaviour will improve the usability of chatbots

Project Goals

- Replicate and verify the results achieved by Larson et al. on their CLINC150 dataset using the same 3 baseline approaches (oos-train, oos-threshold and oos-binary)
- Compute results on two modified versions of the dataset:

  1. Version with a reduced number of in-domain intents

  2. Version with a reduced number of sentences and in-domain intents

Approaches to the problem

- Three baseline approaches to the problem – oos-train, oos-threshold and oos-binary
- oos-train
  - an additional (151st) intent is trained
- oos-threshold
  - Intents with prediction confidence lower than the threshold are classified as out-of-domain
  - Threshold found as the value that yields the highest accuracy on validation set across all intents (including out-of-domain)
- oos-binary
  - First step – binary classification (in- or out-of-domain)
  - Second step – in-domain intent classification if predicted as in-domain, classify as out-of-domain otherwise
  - In-domain classifier always trained on the Full dataset

CLINC150 dataset

- Total of 23,700 queries
  - 22,500 in-domain
  - 1,200 out-of-domain
- 150 in-domain intents + one out-of-domain intent
- Several variants that differ by the number of train queries
- 4 variants of the dataset – Full, Small, Imbalanced and OOS+
  + 2 variants for binary (in- / out-of-domain classification) – Undersample and Wikipedia Augmentation
- Number of queries in the validation/test split
  - Validation – 20 queries per in-domain intent, 100 out-of-domain queries
  - Test – 30 queries per in-domain intent, 1,000 out-of-domain queries

CLINC150 dataset

- Number of train queries
  - Full – 100 queries per in-domain intent, 100 out-of-domain queries
  - Small – 50 queries per in-domain intent, 100 out-of-domain queries
  - Imbalanced – 25/50/75/100 queries per in-domain intent, 100 out-of-domain queries
  - OOS+ – 100 queries per in-domain intent, 250 out-of-domain queries
  - Under – 1,000 in-domain queries, 250 out-of-domain queries
  - Wiki Aug – 15,000 in-domain queries, 15,000 out-of-domain queries

Dataset Modifications [1/2]

Reduction of in-domain intents

- Reduction of the 150 in-domain intents down to 3, 6, 9 and 12
- The random intent selection and measurement is repeated 30 times – the final result is the mean of all results
- To represent scenarios where only a few in-domain intents are needed
- Real-world applications – few in-domain intents are more common than many

Dataset Modifications [2/2]

Reduction of sentences and in-domain intents

- Dataset reduced by the previous modification + imposed limit on the number of sentences of every intent (considering also out-of-domain as one)
- The random intent selection and measurement is repeated 30 times – the final result is the mean of all results
- Sentence limit per in-domain intent
  - Train/validation – 18 sentences
  - Test – 30 sentences
- Sentence limit on out-of-domain
  - Train/validation – 20 sentences
  - Test – 60 sentences
- To represent scenarios where only a few in-domain intents are needed and the dataset size is small
- oos-binary – the sentences of the binary dataset are not limited, only the in-domain

Metrics

- Accuracy over the in-domain intents and recall on out-of-domain queries – same as Larson et al.
  + false acceptance rate (FAR) and false recognition rate (FRR)
- FAR
  - Error of considering out-of-domain queries as in-domain
  - High FAR – the chatbot tends to give an unrelated answer, instead of using a fallback response
- FRR
  - Error of considering in-domain queries as out-of-domain
  - High FRR – the user might think that the system does not work generically, the system fails to guide the user through its dialogue design
- Both FAR and FRR should be as low as possible
- FRR minimisation is preferred in conversational AIs

Classifiers

- Same classifiers and hyperparameters (where possible) as Larson et al.
- FastText – 100 and 300-dimensional pre-trained word vectors in .vec format
- SVM – Scikit-learn implementation
- MLP – Scikit-learn implementation
- BERT – 🤗 Tranformers implementation
- Rasa – 30 train epochs, spaCy language model

Results on original dataset

oos-train

Results on all 150 intents

		In-Domain Accuracy				Out-of-Domain Recall				FAR				FRR			
		Full	Small	Imbalanced	OOS+	Full	Small	Imbalanced	OOS+	Full	Small	Imbalanced	OOS+	Full	Small	Imbalanced	OOS+
FastText	cc.en.100.vec	89.2	86.2	86.7	88.9	23.2	23.1	25.2	43.1	76.8	76.9	74.8	56.9	0.4	0.7	0.5	0.9
FastText	cc.en.300.vec	89.7	87.1	87.1	89.1	24.9	24.4	26.4	43.2	75.1	75.6	73.6	56.8	0.4	0.6	0.5	1.1
FastText	PAPER	89.0	84.5	87.2	89.2	9.7	23.2	12.2	32.2	90.3	76.8	87.8	67.8				
SVM		90.7	88.5	89.0	90.4	37.6	47.3	42.8	58.2	62.4	52.7	57.2	41.8	1.4	2.4	1.8	2.4
SVM	PAPER	91.0	89.6	89.9	90.1	14.5	18.6	16.0	29.8	85.5	81.4	84.0	70.2				
MLP		90.9	88.3	89.8	90.6	10.8	14.9	12.5	23.9	89.2	85.1	87.5	76.1	0.1	0.4	0.3	0.4
MLP	PAPER	93.5	91.5	92.5	94.1	47.4	52.2	35.6	53.9	52.6	47.8	64.4	46.1				
BERT		95.9	95.2	94.7	95.9	35.6	41.8	39.9	53.0	64.4	58.2	60.1	47.0	0.1	0.2	0.2	0.3
BERT	PAPER	96.9	96.4	96.3	96.7	40.3	40.9	43.8	59.2	59.7	59.1	56.2	40.8				
Rasa		91.2	86.4	88.9	90.9	19.1	20.8	28.6	44.6	80.9	79.2	71.4	55.4	0.2	0.3	0.6	0.7
Rasa	PAPER	91.5	88.9	89.2	90.9	45.3	55.0	49.6	66.0	54.7	45.0	50.4	34.0				

oos-threshold

Results on all 150 intents

		In-Domain Accuracy			Out-of-Domain Recall			FAR			FRR		
		Full	Small	Imbalanced	Full	Small	Imbalanced	Full	Small	Imbalanced	Full	Small	Imbalanced
FastText	cc.en.100.vec	88.8	86.5	87.1	11.9	8.0	9.1	88.1	92.0	90.9	0.4	0.3	0.2
FastText	cc.en.300.vec	90.0	87.4	86.8	6.8	2.4	28.6	93.2	97.6	71.4	0.1	0.0	1.7
FastText	PAPER	88.6	84.8	86.6	28.3	6.0	33.2	71.7	94.0	66.8			
SVM		90.6	88.5	88.0	22.7	36.2	35.9	77.3	63.8	64.1	1.2	2.9	3.6
SVM	PAPER	88.2	85.6	86.0	18.0	13.0	0.0	82.0	87.0	100.0			
MLP		90.8	88.5	89.6	14.2	14.9	23.3	85.8	85.1	76.7	0.5	0.7	1.3
MLP	PAPER	93.4	91.3	92.5	49.1	32.4	13.3	50.9	67.6	86.7			
BERT		95.7	94.8	94.6	40.5	26.7	16.7	59.5	73.3	83.3	0.7	0.4	0.5
BERT	PAPER	96.2	96.2	95.9	52.3	58.9	52.8	47.7	41.1	47.2			
Rasa		91.5	86.3	88.0	17.7	8.2	28.5	82.3	91.8	71.5	0.4	0.5	2.0
Rasa	PAPER	90.9	89.6	89.4	31.2	1.0	0.0	68.8	99.0	100.0			

oos-binary

Results on all 150 intents

		In-Domain Accuracy		Out-of-Domain Recall		FAR		FRR	
		Under	Wiki Aug	Under	Wiki Aug	Under	Wiki Aug	Under	Wiki Aug
FastText	cc.en.100.vec	86.6	82.6	29.7	58.8	70.3	41.2	3.4	8.7
FastText	cc.en.300.vec	87.0	87.9	30.7	29.9	69.3	70.1	3.5	2.6
FastText	PAPER	88.1	87.0	22.7	31.4	77.3	68.6		
SVM		89.6	88.7	24.7	31.3	75.3	68.7	1.8	3.5
SVM	PAPER	88.4	89.3	32.2	37.7	67.8	62.3		
MLP		85.6	85.9	43.5	54.9	56.5	45.1	6.5	6.9
MLP	PAPER	90.1	92.9	52.8	32.4	47.2	67.6		
BERT		91.7	94.9	59.7	44.8	40.3	55.2	4.7	1.3
BERT	PAPER	94.4	96.0	46.5	40.4	53.5	59.6		
Rasa		87.0	90.6	41.4	28.2	58.6	71.8	4.1	1.0
Rasa	PAPER	87.5	--	37.7	--	62.3	--		

Results Comparison

All 150 intents

- Good overall accuracy, best in oos-train
- Recall and FAR
  - No approach is significantly better than the others, no simple comparison
  - oos-train (on OOS+) is comparable with oos-binary + it has lower FRRs
  - Results far from the desired recall
- FRR is always relatively low, oos-binary has the highest FRR

Possible Reasons of Measured Differences

All 150 intents

- Classifier implementation – the paper did not specify the used implementations
- Hyperparameters – not all could be set / were specified
- FastText – the paper did not specify the format
- Rasa – the paper did not specify the number of epochs
- oos-binary – the paper did not specify the dataset for in-domain training

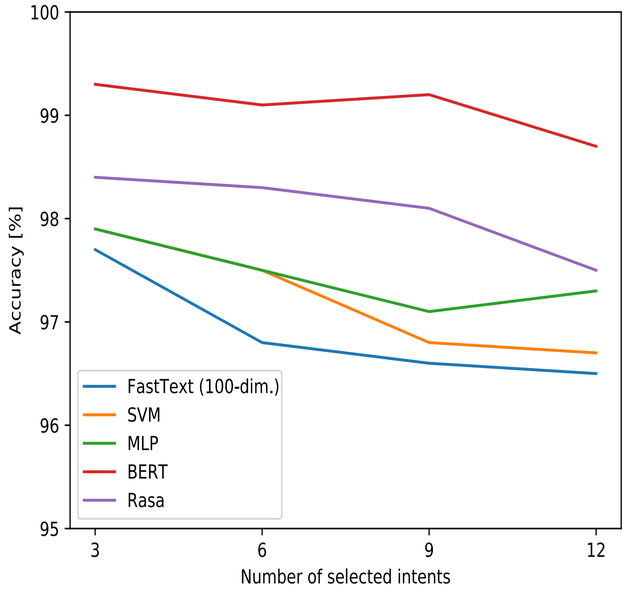
Results on modified datasets

oos-train

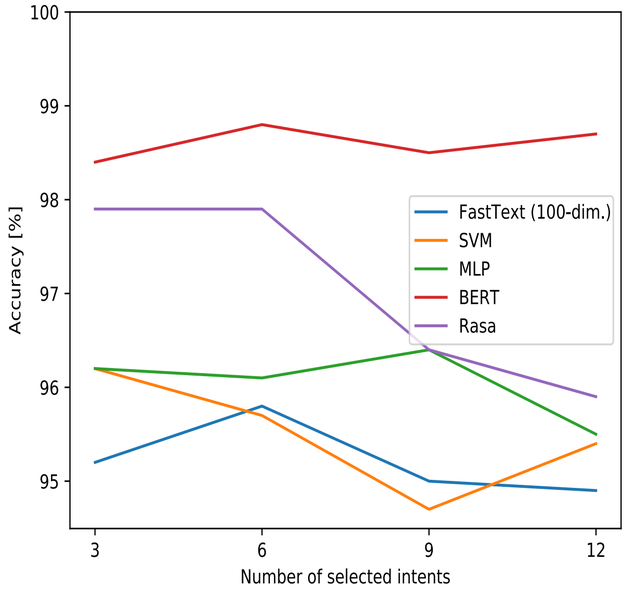
oos-train

Accuracy on reduced in-domain intents

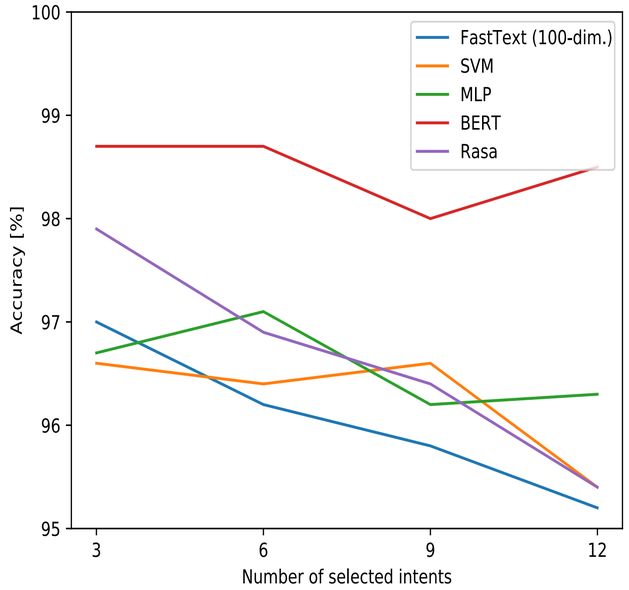
Full



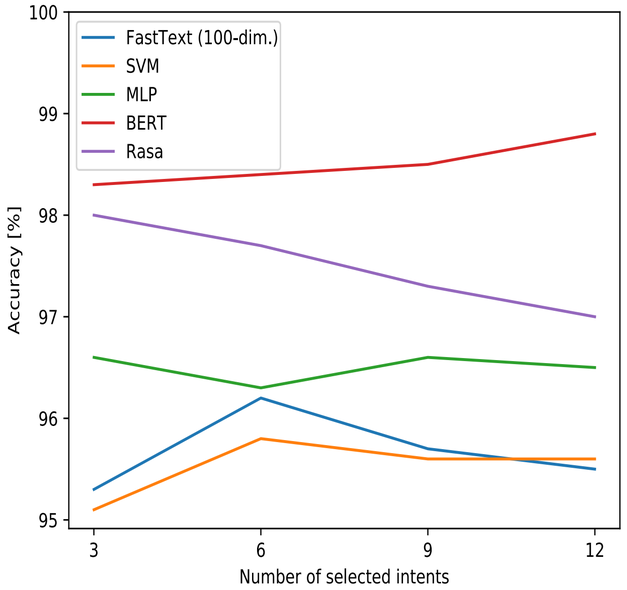
Small



Imbalanced



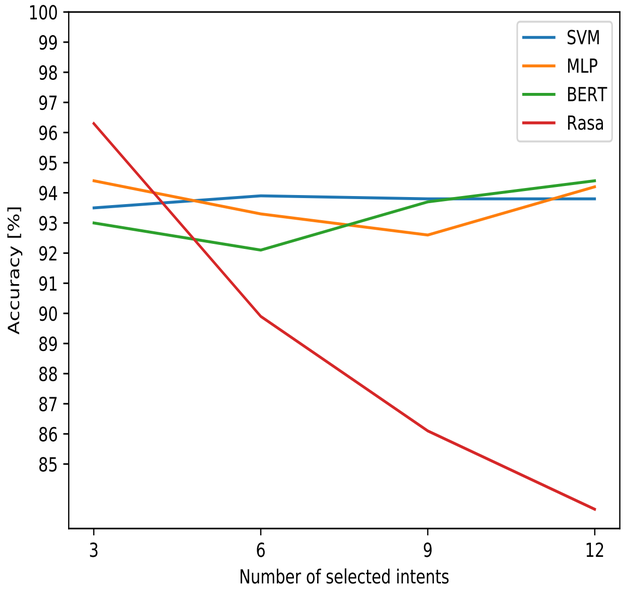
OOS+



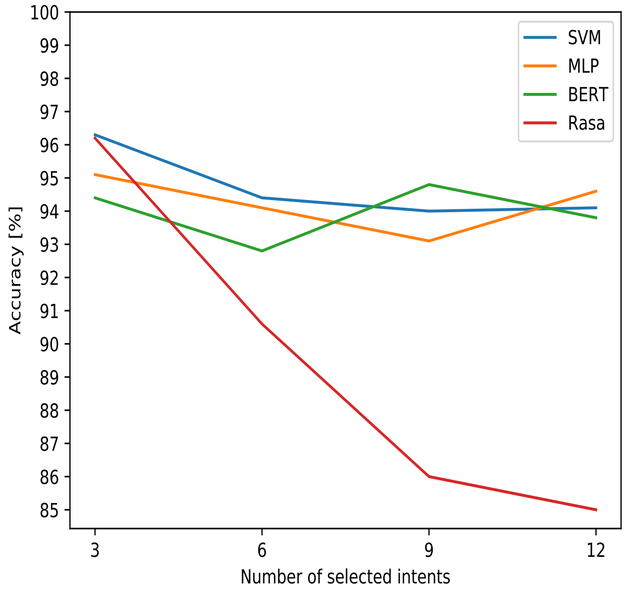
oos-train

Accuracy on reduced in-domain intents and limited sentences

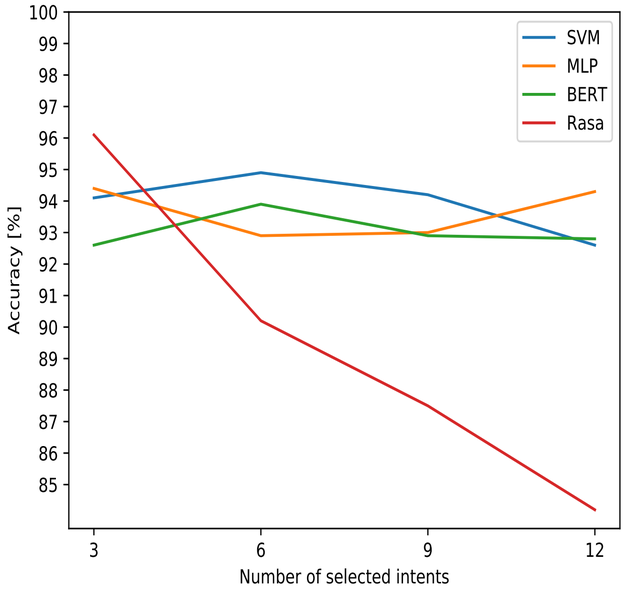
Full



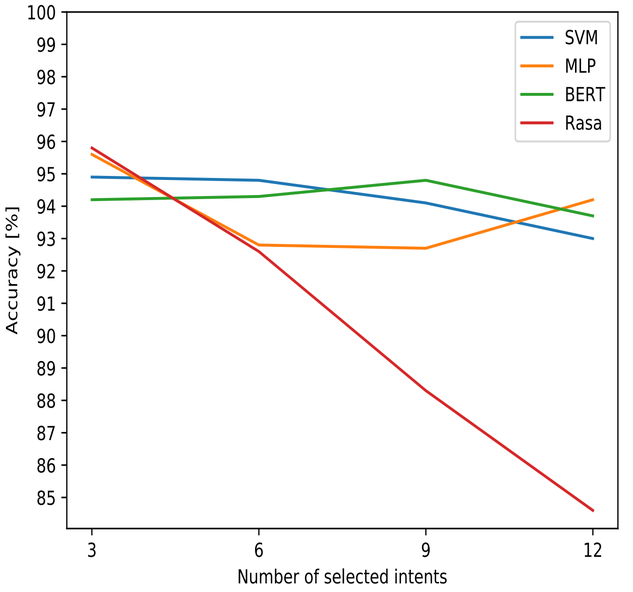
Small



Imbalanced



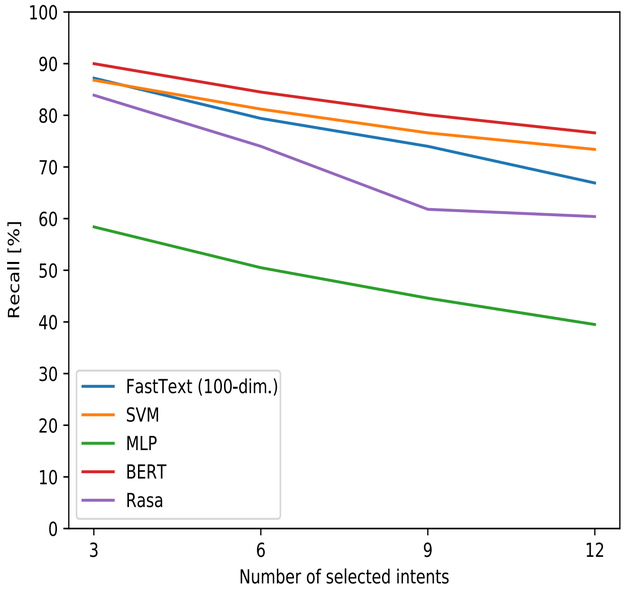
OOS+



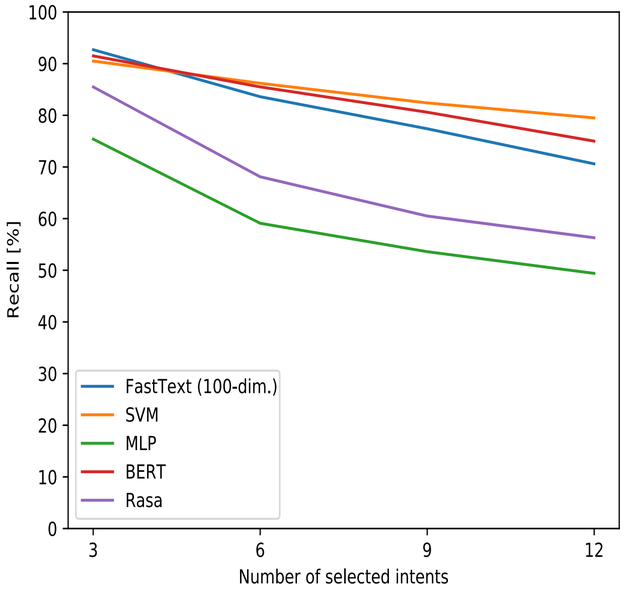
oos-train

Out-of-domain recall on reduced in-domain intents

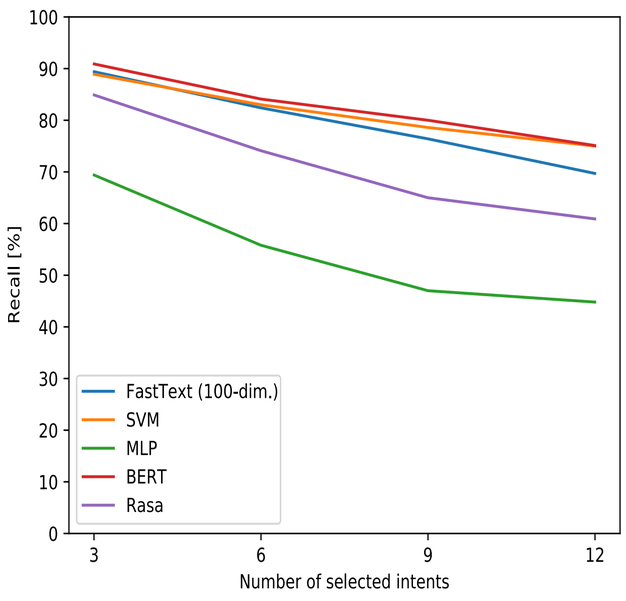
Full



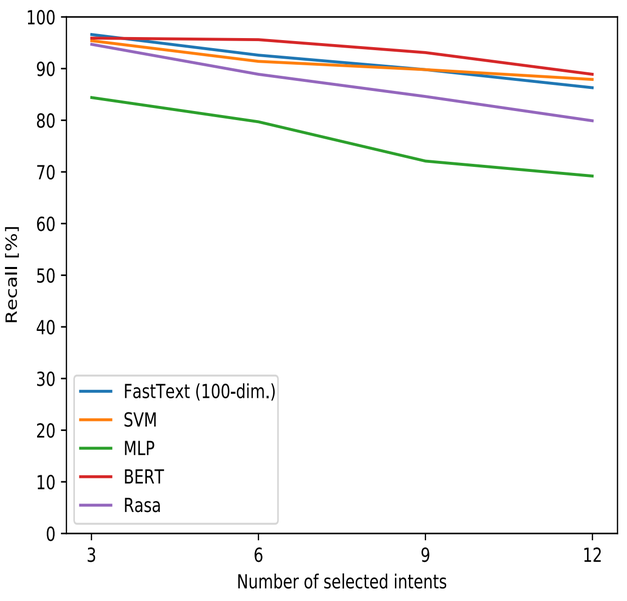
Small



Imbalanced



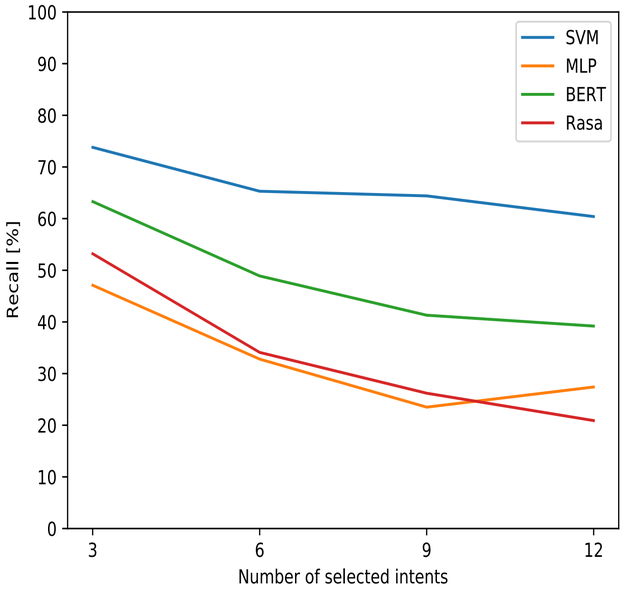
OOS+



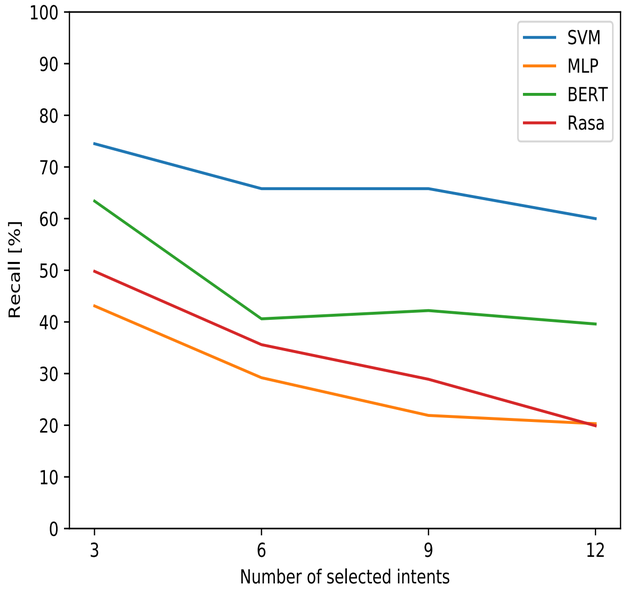
oos-train

Out-of-domain recall on reduced in-domain intents and limited sentences

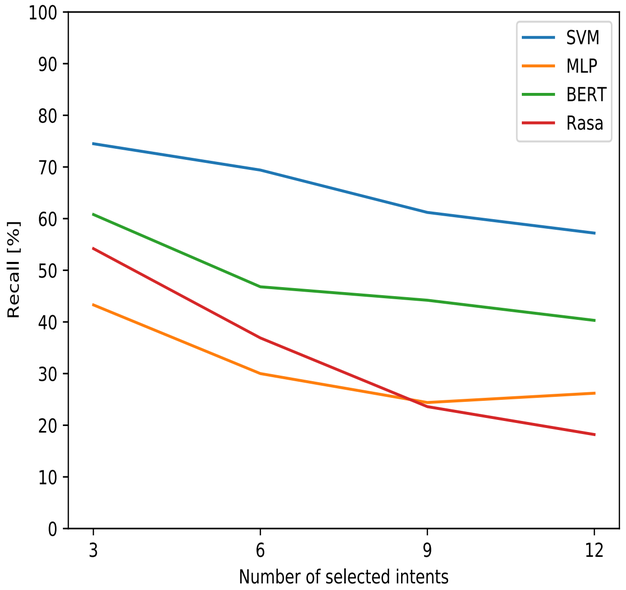
Full



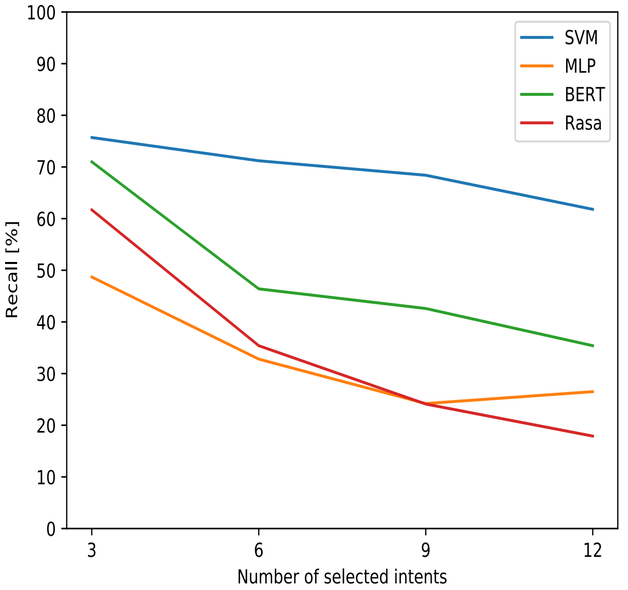
Small



Imbalanced



OOS+

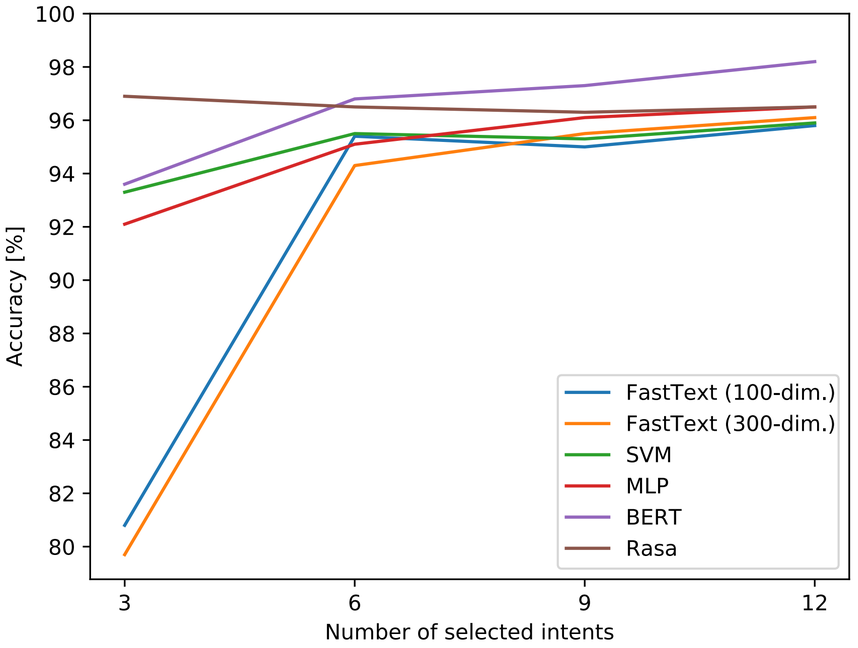


oos-threshold

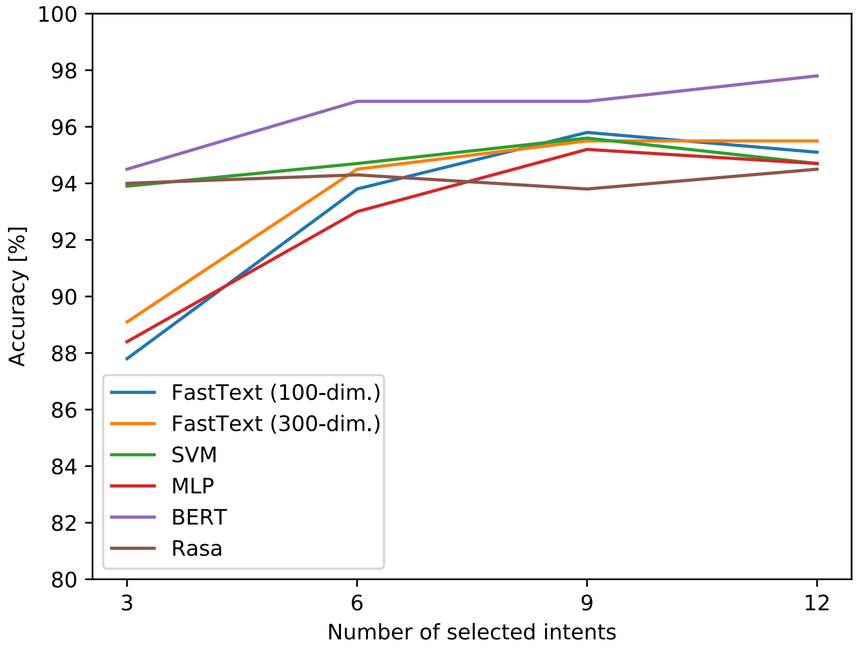
oos-threshold

Accuracy on reduced in-domain intents

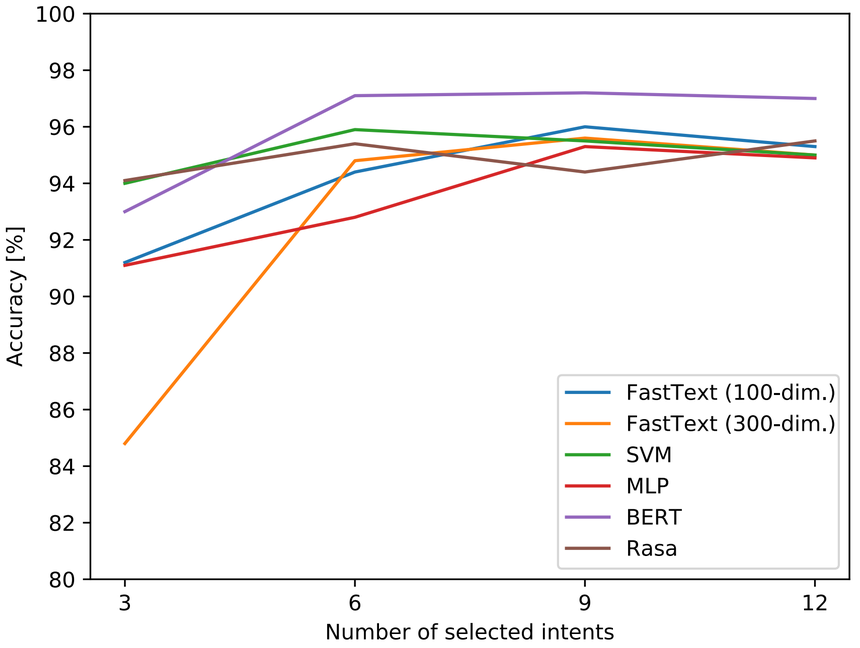
Full



Small



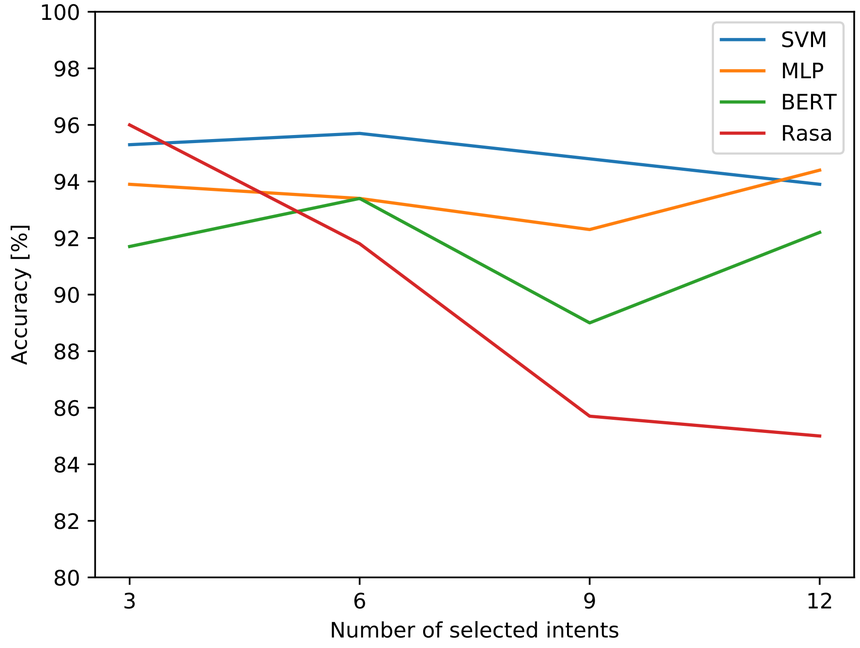
Imbalanced



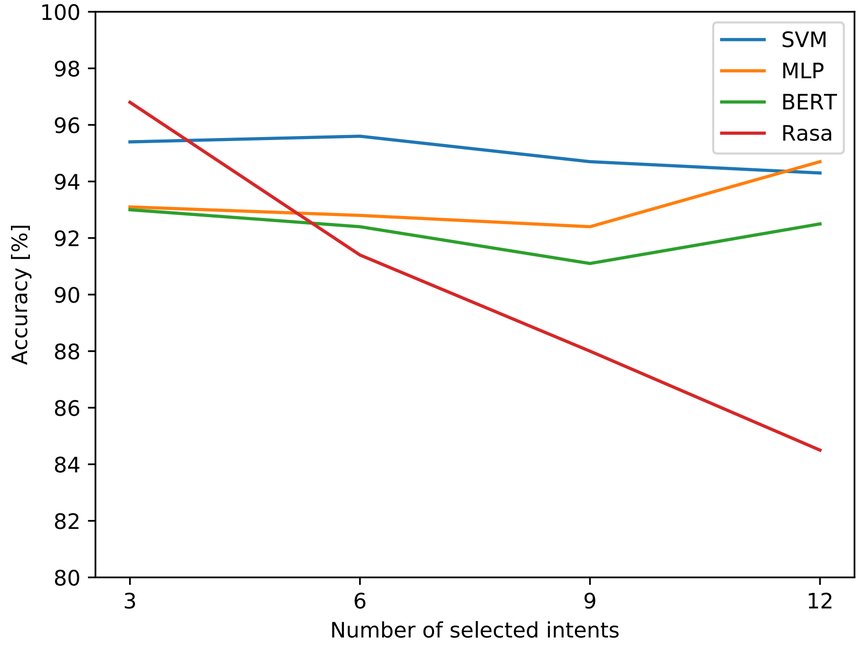
oos-threshold

Accuracy on reduced in-domain intents and limited sentences

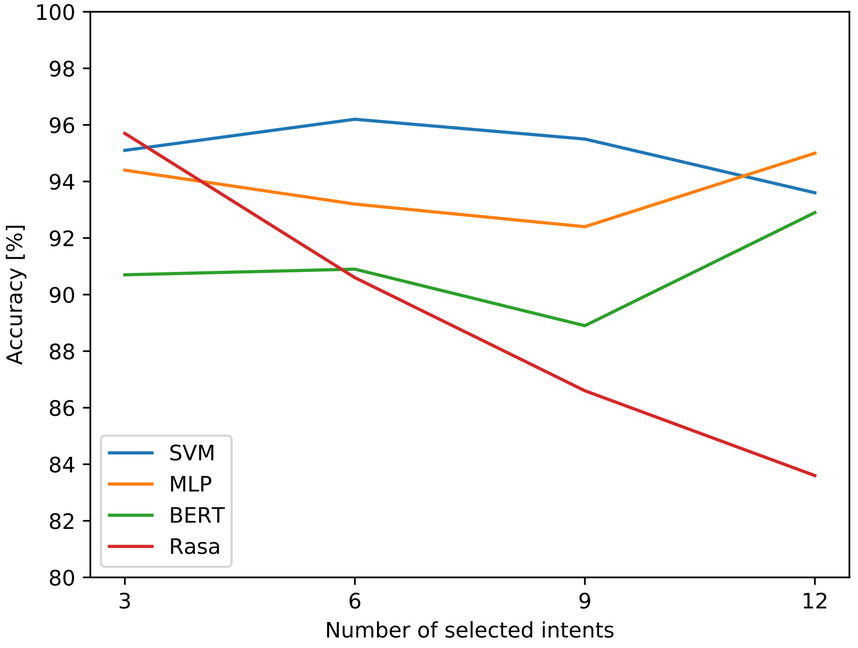
Full



Small



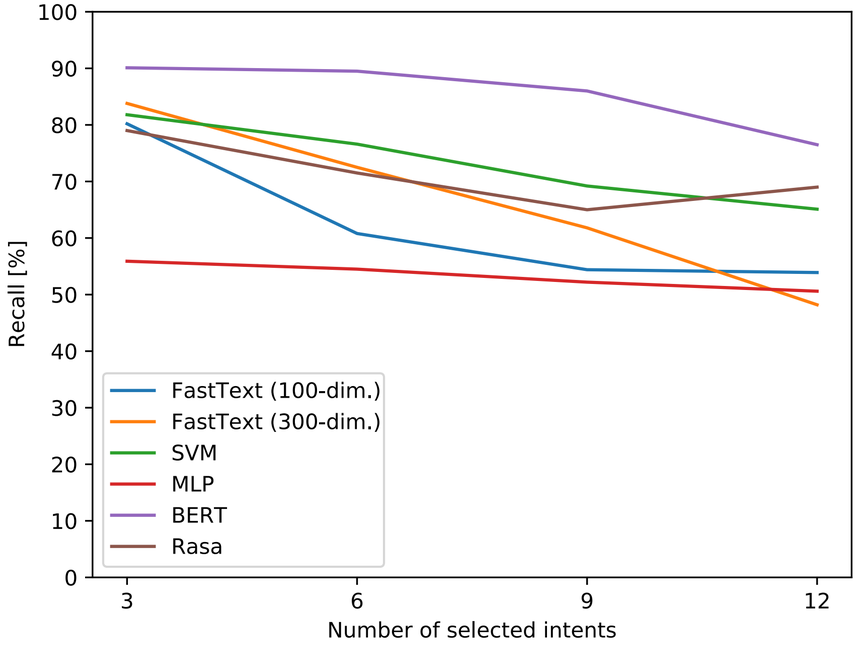
Imbalanced



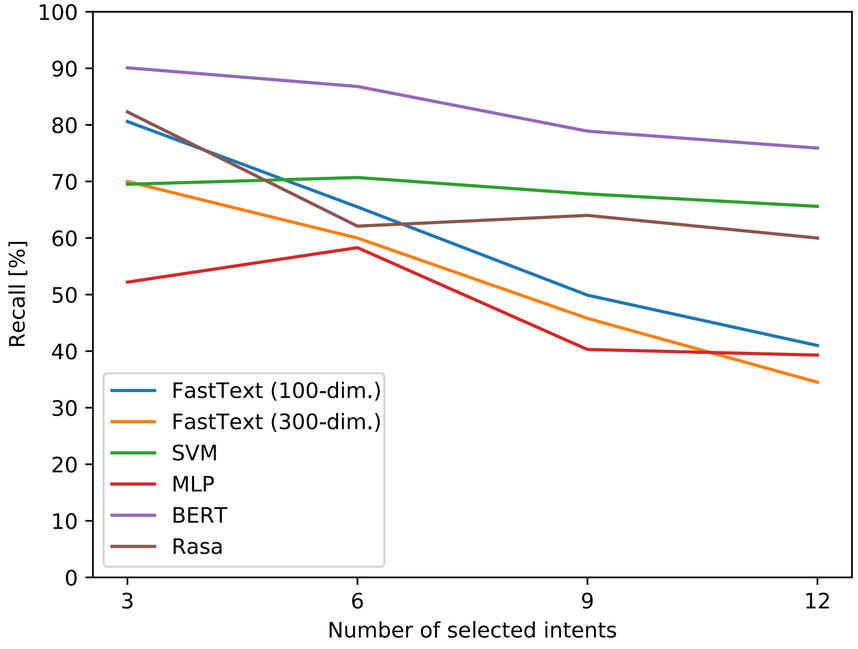
oos-threshold

Out-of-domain recall on reduced in-domain intents

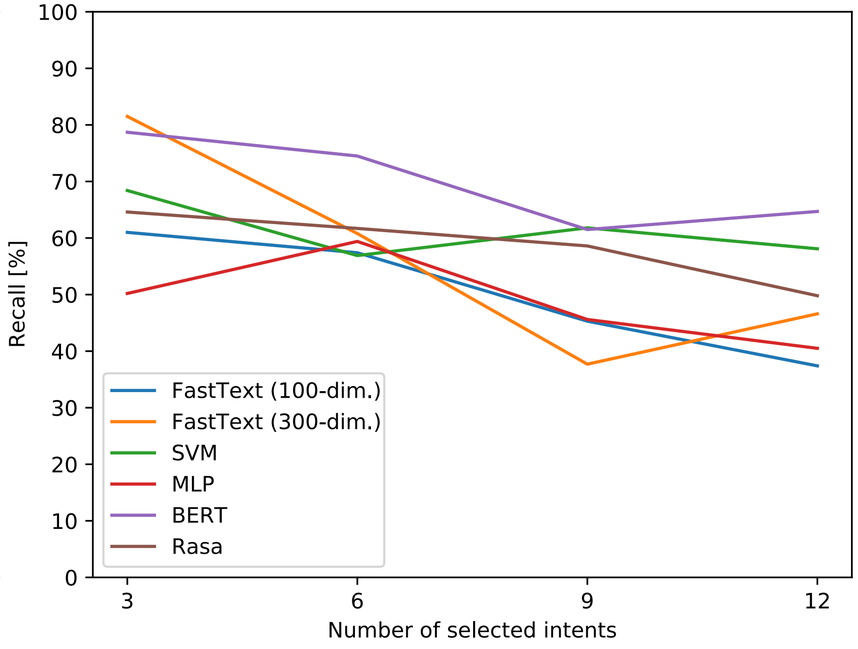
Full



Small



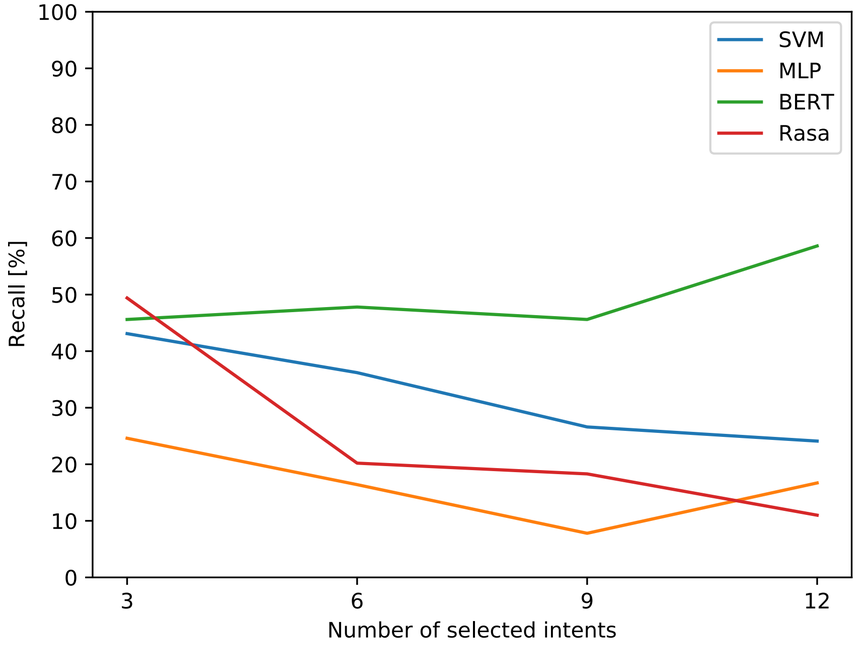
Imbalanced



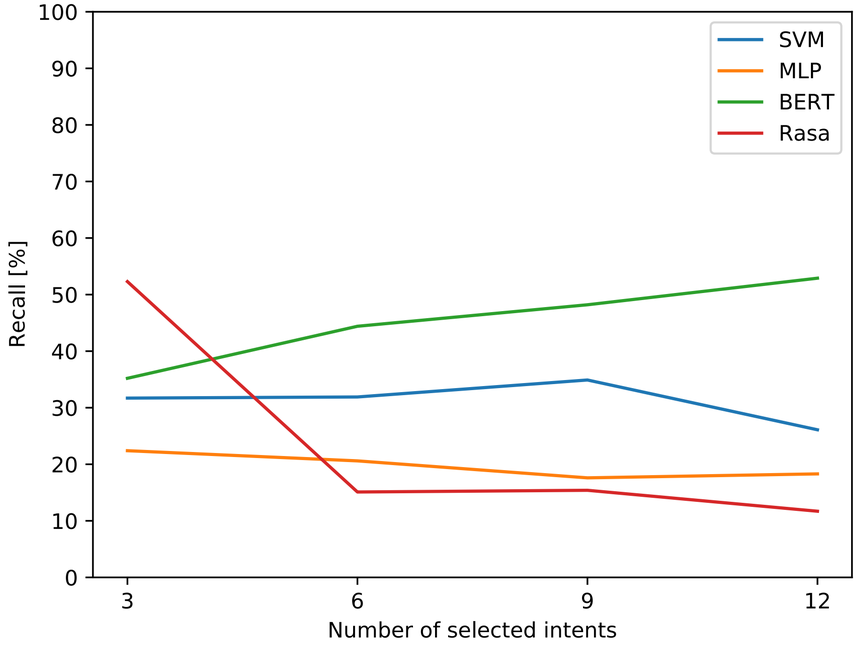
oos-threshold

Out-of-domain recall on reduced in-domain intents and limited sentences

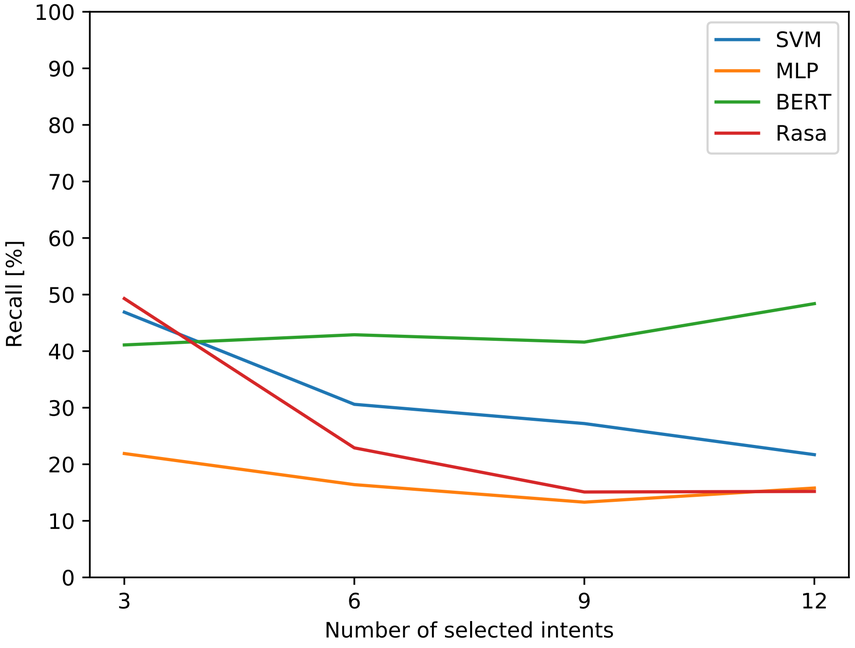
Full



Small



Imbalanced

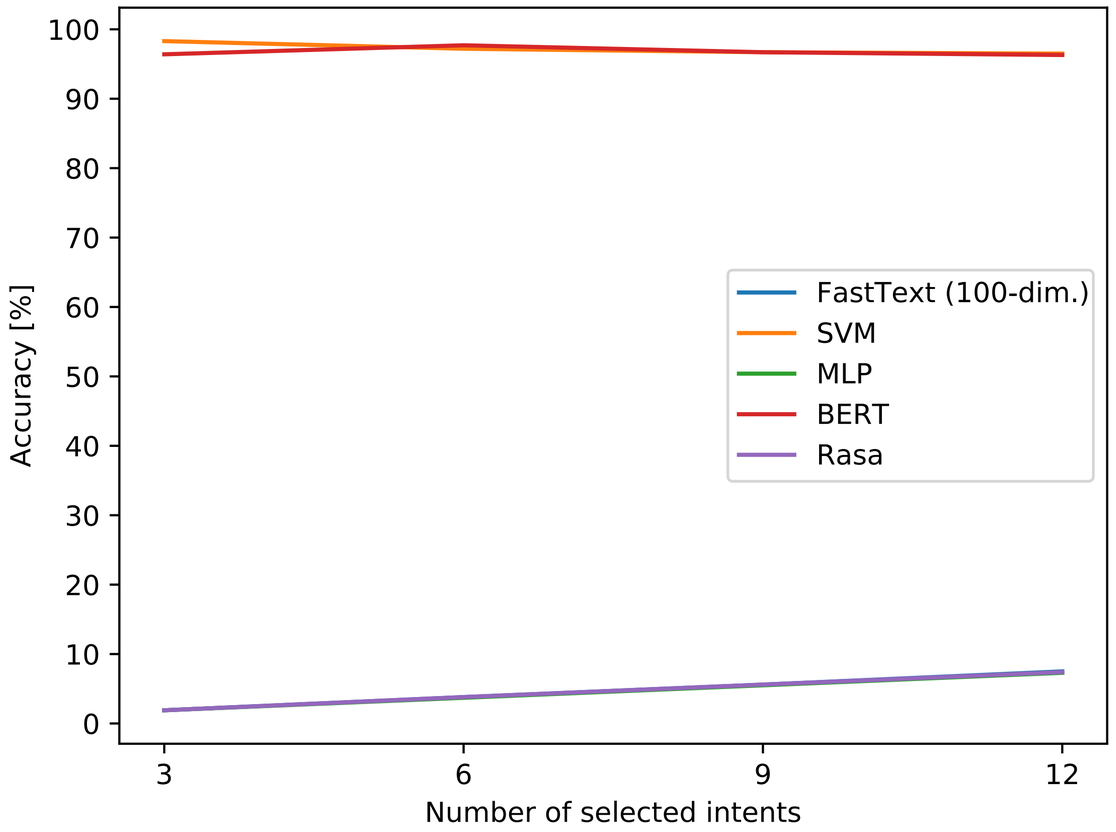


oos-binary

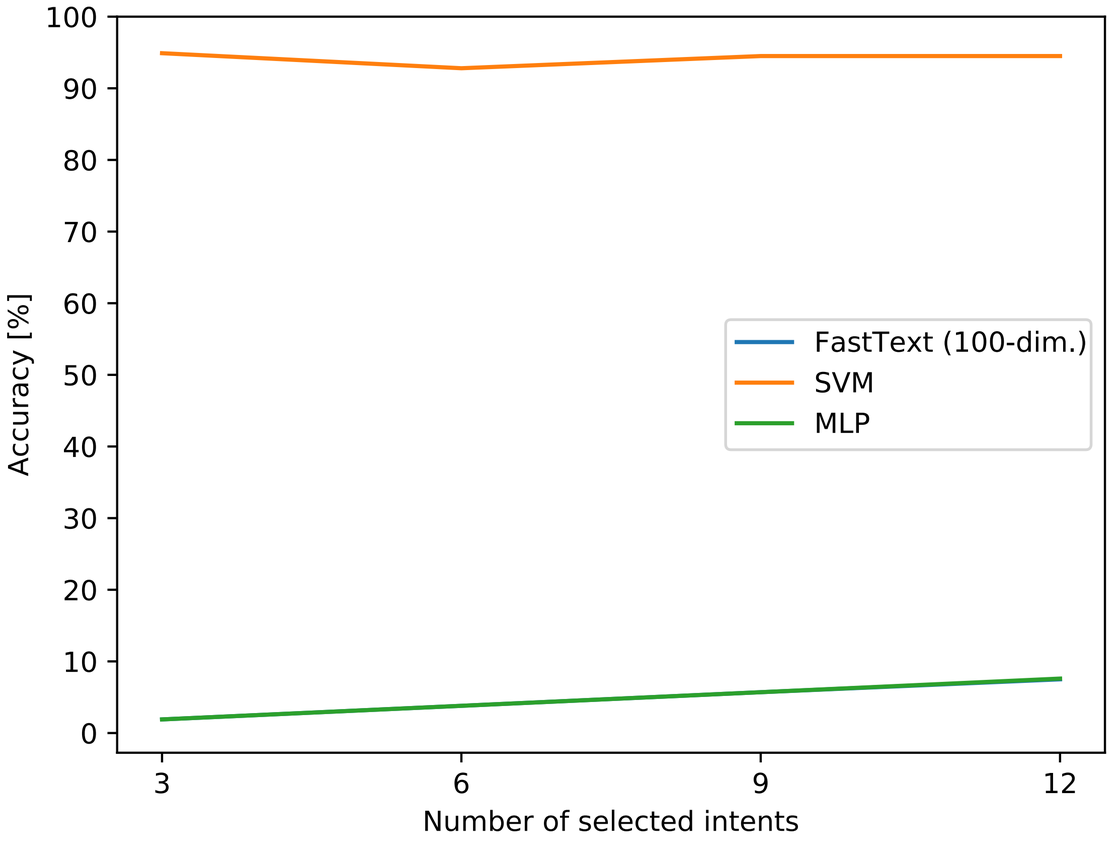
oos-binary

Accuracy on reduced in-domain intents

Under



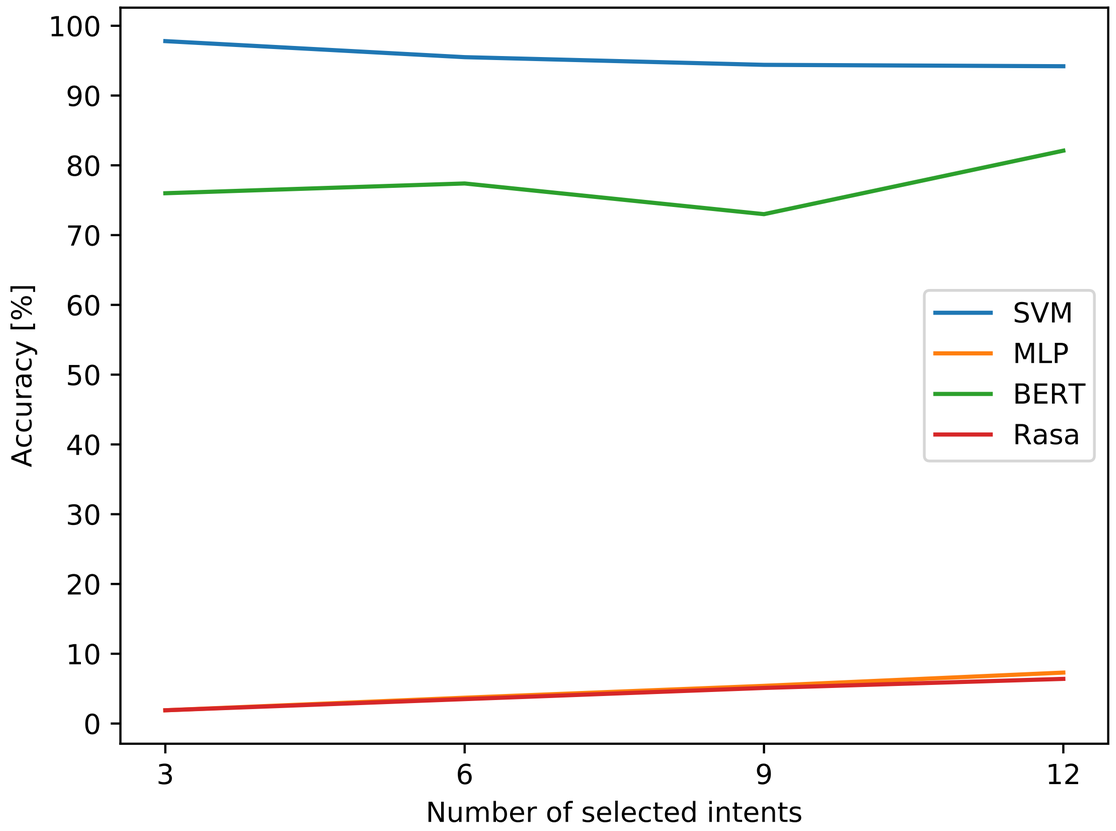
Wiki Aug



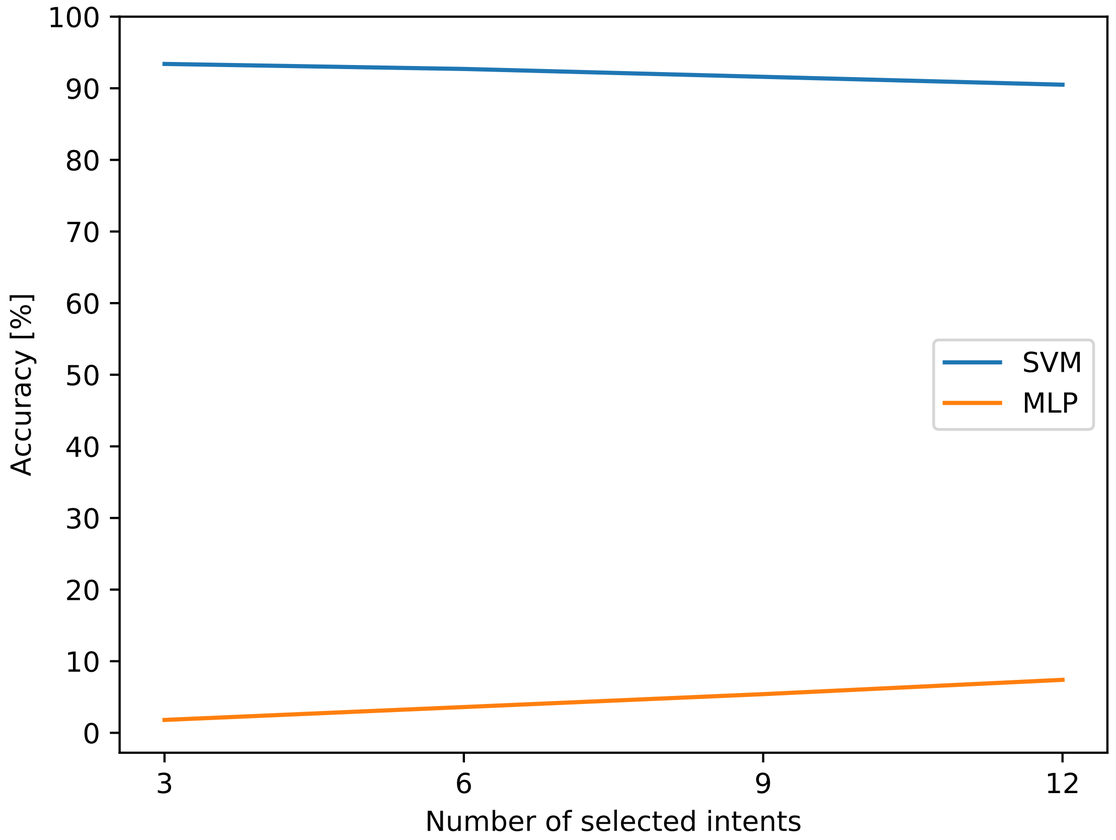
oos-binary

Accuracy on reduced in-domain intents and limited sentences

Under



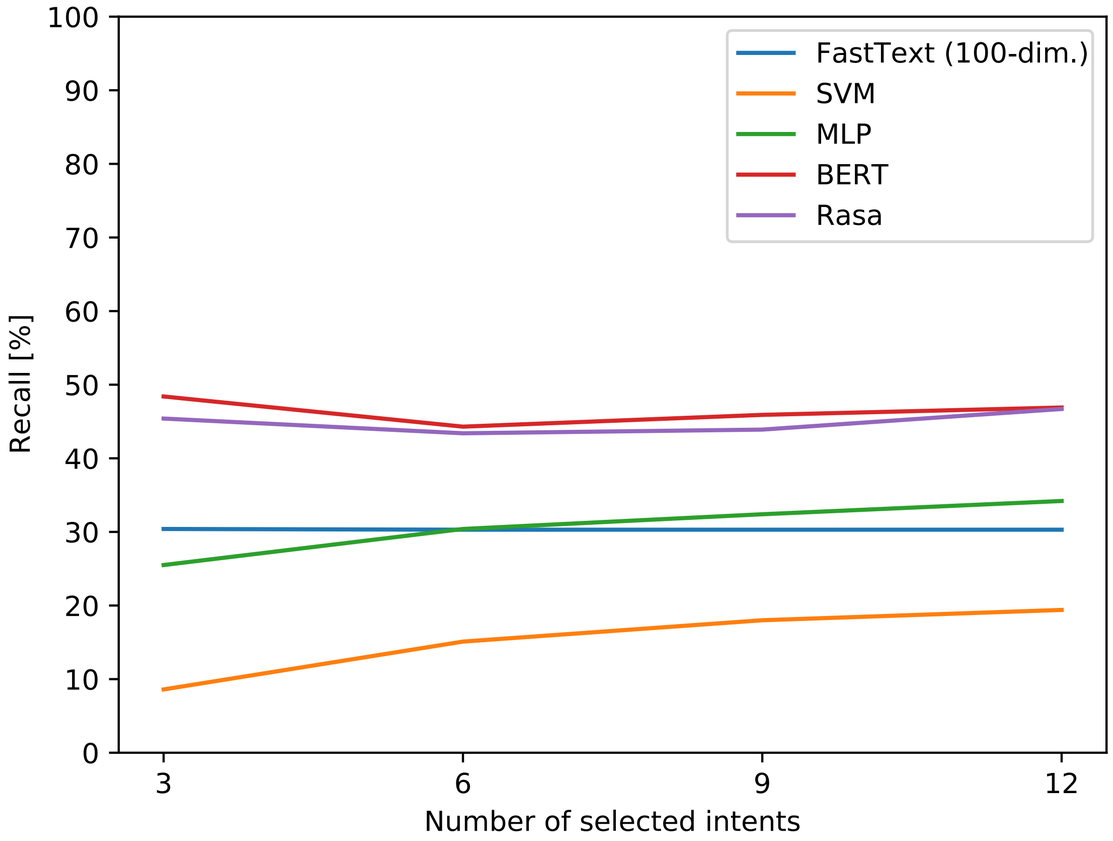
Wiki Aug



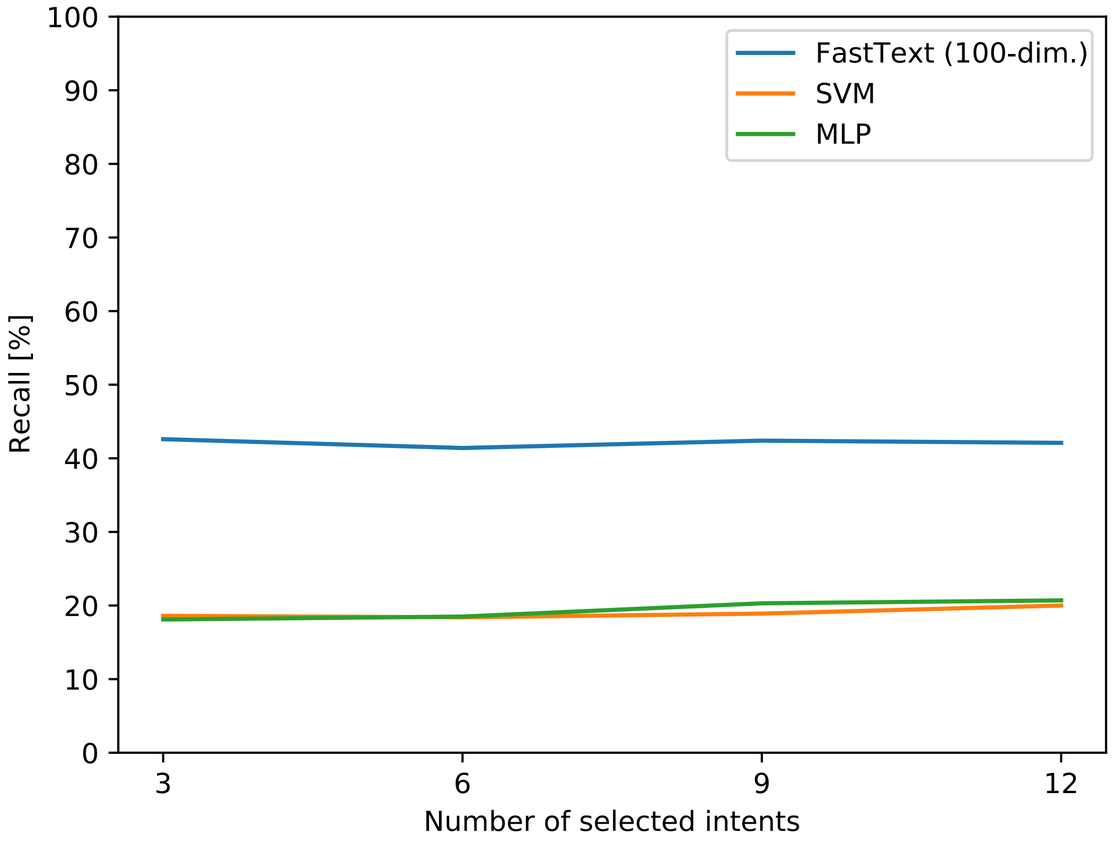
oos-binary

Out-of-domain recall on reduced in-domain intents

Under



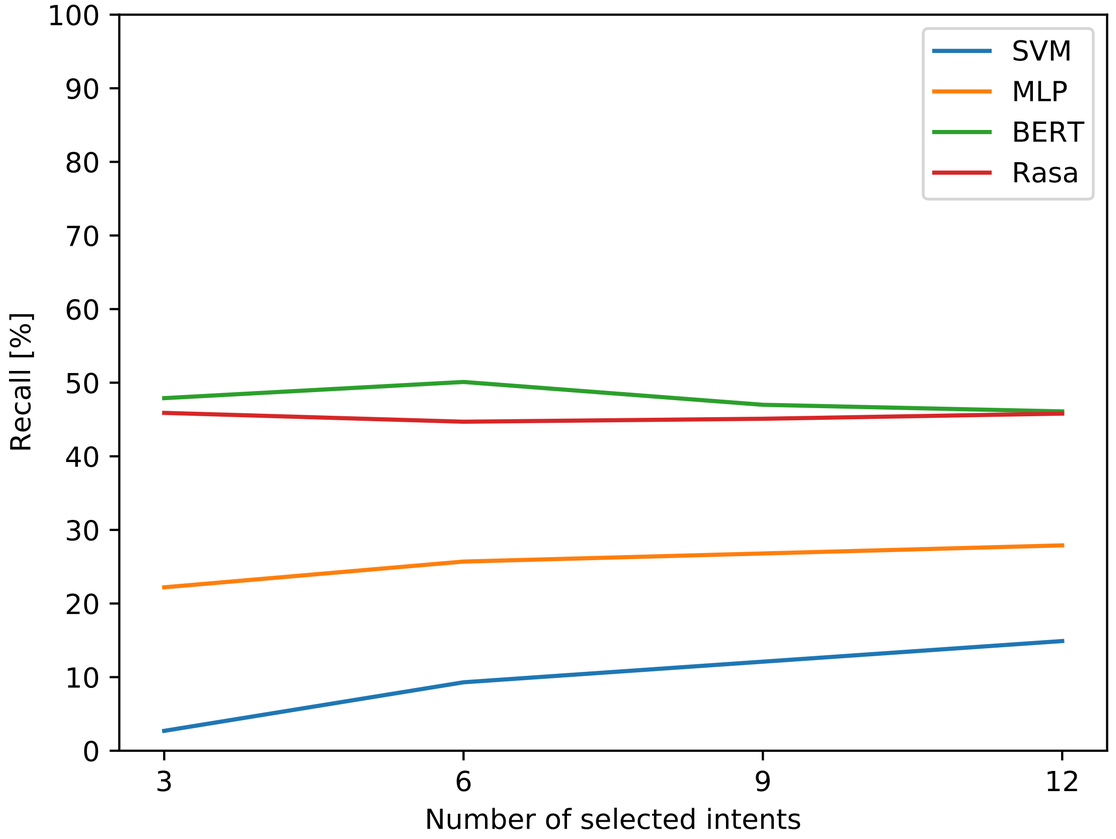
Wiki Aug



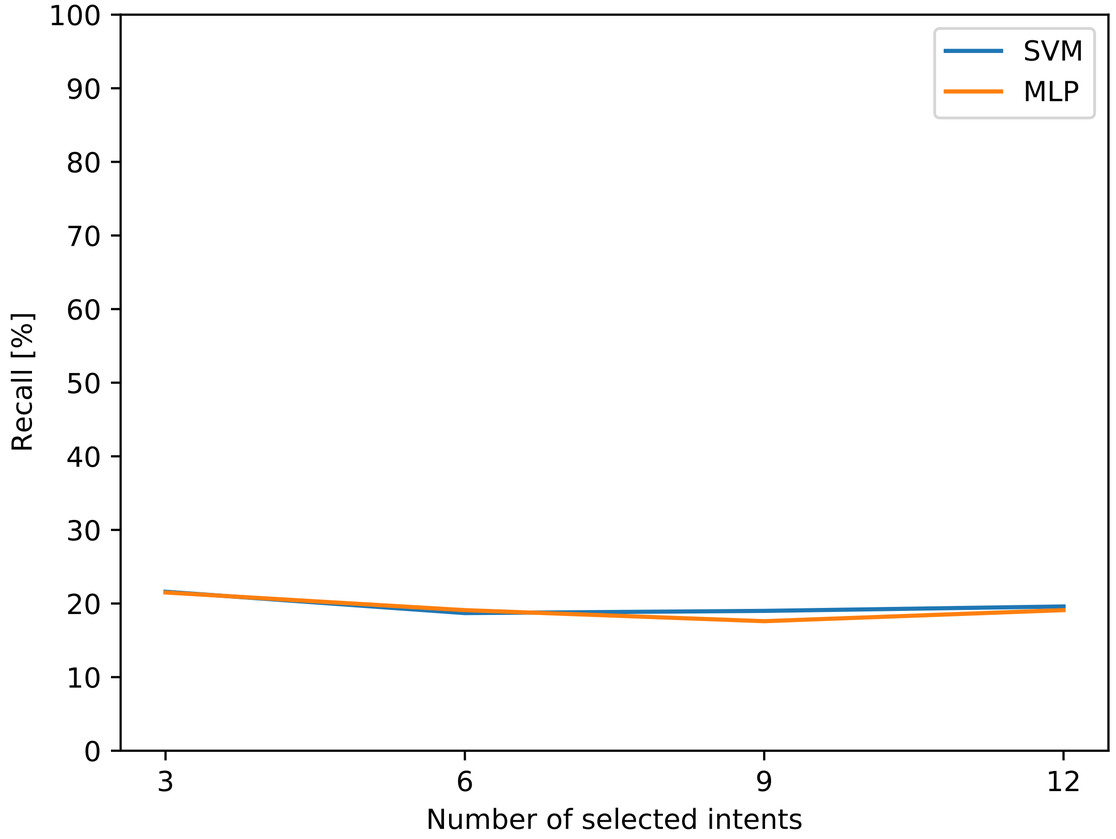
oos-binary

Out-of-domain recall on reduced in-domain intents and limited sentences

Under



Wiki Aug



FAR and FRR - all approaches

- FAR – dependent on recall, using the formula FAR = 100 - recall
- FRR – always only a few percent on all approaches

Results Comparison [1/2]

- Accuracy on reduced in-domain intents
  - Both oos-train and oos-threshold (excl. 3 selections) very high results
  - oos-binary – results extremely depend on used classifier
- Accuracy on reduced in-domain intents and limited sentences
  - oos-train, oos-threshold – similar (high) results, Rasa accuracy plummets on increasing number of intents
  - oos-binary – again, very different results based on classifier

Results Comparison [2/2]

- Recall on reduced in-domain intents
  - Best overall results – oos-train, worst results – oos-binary
  - Large difference between best and worst approach
- Recall on reduced in-domain intents and limited sentences
  - Best overall results – oos-train, worst results – oos-binary

Conclusion

- Out-of-domain performance tends to improve with
  - More out-of-domain data
  - Smaller imbalance between total number of in-domain and out-of-domain in the train set (provided that there is enough out-of-domain data – does not apply to the previous sentences limiting)
- No simple outcomes

Thank you for your attention!