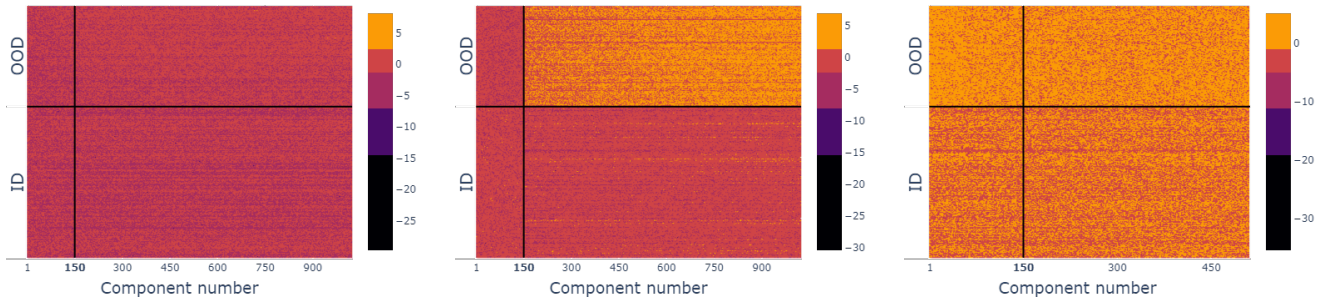

(a) RoBERTa, no fine-tuning (b) RoBERTa, fine-tuned (c) CNN, trained

Figure 3: These heatmaps represent each utterance from the CLINC150 test set as the vector of Mahalanobis distance terms, computed according to Eq. 5 and sorted in the decreasing order of explained variance. Each row stands for an utterance. The horizontal solid line separates the OOD utterances (above the line) from the ID ones (below the line). The vertical solid line splits each heatmap into two parts: to the left are components numbered lower than 150, to the right are components numbered above 150. 150 is the number of classes in the CLINC150 dataset. Only fine-tuned RoBERTa-based vectors clearly distinguish ID and OOD utterances. The difference between ID and OOD is less evident in (c) and almost indistinguishable in (a). However, in (b), the values of the components, starting from the 150^{th} one (in yellow), are lower than those of ID ones (in red).

are close to class centroids, and the number of classes is significantly lower than the dimension of the embeddings ($N \ll d$).

For further analysis, we consider several Mahalanobis distance variants. Following **?**, we introduce the equivalent Mahalanobis distance form, based on Principal Component Analysis of the class-wise centered ID data:

$$d(\psi(x)) = \min_{c} \sum_{i=1}^{d} \frac{y_i^2(\psi(x) - \mu_c)}{\lambda_i}, \tag{5}$$

where $y_i(\psi(x))$ is the i-th component of the PCA transform of $\psi(x)$, λ_i are explained variances of the corresponding principal components, μ_c are class centroids.

? introduced two modifications of Eq. 5, namely, **marginal Mahalanobis distance**, which ignores class information and uses instead a single mean vector for all ID classes, (see Eq. 6) and **partial Mahalanobis distances**: it is the version of the equations (5) and (6) with the summation starting from N-th component. Eq. 7 corresponds to the **partial marginal** variant. Marginal Mahalanobis distance aims at using more compact data representation in the form of a single ID centroid, helping to reduce the amount of data needed for OOD detection. Partial variant utilizes the most important terms only.

$$d(\psi(x)) = \sum_{i=1}^{d} \frac{y_i^2(\psi(x) - \mu)}{\lambda_i} \tag{6}$$

$$d(\psi(x), N) = \sum_{i=N}^{d} \frac{y_i^2(\psi(x) - \mu)}{\lambda_i}, \tag{7}$$

where

$$\mu = \frac{1}{N} \sum_{x \in \mathcal{D}_{in}} \psi(x),$$

stands for the ID data centroid.

Mahalanobis distance can efficiently utilize low-dimensional nature of ID data. Following the properties of PCA (**?**), if the data is approximately N-dimensional, it is explained by the first N principal components. That means that for ID data, all the terms in the Eq. 6, 7 are little, while OOD data can be detected by important loadings of the terms $\frac{y_i^2}{\lambda_i}$ with $i > N$. To check this, we plot the terms of the Eq. 7 for ID and OOD data, Fig. 3. Fig. 3 shows that when decomposed with the Mahalanobis distance embeddings of fine-tuned RoBERTa fall into two parts. The last components of OOD embeddings have a higher variance when compared to the first ones. This phenomena is observed neither for ID embeddings nor for RoBERTa without fine-tuning nor for the trained CNN.

Comparison of other distances.

We compare Mahalanobis distance variants to explore this matter: original, marginal Mahalanobis, and their partial versions. Additionally, we exploit Euclidean distance to complete our evaluation.

All Mahalanobis distance variants outperform Euclidean distance by far; see Fig. 4. Euclidean distance does not take the correlation between features into account. Although there is little difference between Mahalanobis distance variants, partial and marginal variants are more stable when varying training data size. Marginal Mahalanobis distance is less affected by the reduction of training data; see 5.

Conclusion

Out-of-Domain (OOD) detection task is becoming core to modern dialog systems. Successful detection and rejection of OOD utterances in real-life applications increase the dialog assistant's credibility and improves user experience. This paper compared multiple techniques for unsupervised OOD detection, applied to three commonly used NLU datasets, in particular, CLINC150, ROSTD, and SNIPS. We exploited different text representation models, ranging from the old-fashioned bag-of-word modes to the most recent pre-trained

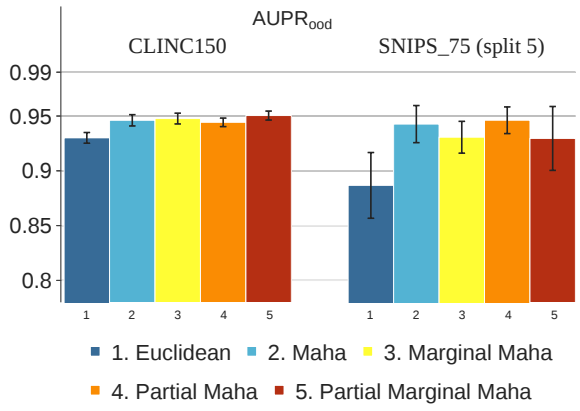

Figure 4: Comparison of different distances. Mahalanobis distance and its variants outperform Euclidean distance by a wide margin.

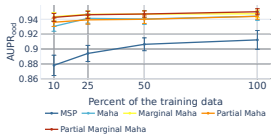

Figure 5: OX: fraction of train data used, CLINC150, OY: performance of OOD detection score. Mahalanobis distance and its variants need less data for OOD detection.

Transformers. We adopted best practices used in the vision domain and previously established state-of-the-art methods within the scope of unsupervised methods, namely, Maximum Softmax Probability, Likelihood Ratio, and Mahalanobis distance, along with its modifications.

With the help of Transformer-based models, equipped with Mahalanobis distance, we establish new state-of-the-art results. To that end, we show that fine-tuning with ID data's supervision plays a crucial role, allowing re-shaping, favorable for the task, of the embedding space. These results are supported in line with (**?**), confirming that fine-tuning Transformers improves the performance of the downstream unsupervised tasks. The proposed pipeline, i.e., fine-tuning a Transformer and using Mahalanobis distance, is robust to distillation. Supporting smaller models is essential for edge devices, where distilled models are usually deployed. Reduced in size, distilled versions of pre-trained Transformers models perform on par with the full-size models. Mahalanobis distance remains stable, even when used with a distilled model. Still, there are some limitations to the Mahalanobis OOD score. In the first place, it depends on the geometrical features of the embedding space, which could be spoilt if, for example, the embedder is used simultaneously as a classification model and overfits. The greatest challenge is then semantically similar utterances, of which one is in ID, and the other is OOD. For example, this can happen if the dialog assistant supports only one of two related actions. Future research directions should consider such cases and the trade-off between the accuracy of intents classification and OOD detection.

Acknowledgments

Ekaterina Artemova is partially supported by the framework of the HSE University Basic Research Program and funded by the Russian Academic Excellence Project "5-100". Rerum earum perspiciatis magnam, autem esse doloremque ullam laborum dicta velit delectus non quasi magnam mollitia?Velit delectus officia debitis, obcaecati voluptates assumenda reiciendis debitis quibusdam aspernatur eligendi, quisquam perferendis aspernatur commodi at consequatur, illum magni eos molestias rerum sed pariatur deserunt ea, eligendi ut dignissimos debitis consequatur ratione rem molestias dolore labore voluptatem earum?Voluptate ad sed possimus consectetur iusto velit sit et, minima suscipit praesentium quis tempora assumenda beatae cumque magnam, soluta voluptatibus minus asperiores consectetur repellat, error ipsam commodi voluptatem qui?Ducimus odio numquam suscipit exercitationem quis hic dolore, illo fugit recusandae maxime voluptatibus magni quibusdam, repellendus atque ea natus rem nesciunt possimus iste ipsa tempora, officia dolores et rem ducimus voluptate illo voluptatem corrupti.Odio ea quidem enim, cum molestias ipsam ratione dicta maxime?Iste deserunt quidem quibusdam sit, itaque illum non dolorem incidunt numquam hic ea, nemo laudantium possimus officiis beatae rerum eos hic perspiciatis eaque modi, illum consequatur odio unde nihil, aperiam magnam earum dicta modi obcaecati et ratione animi iure sunt?Officia rerum a nobis placeat ratione assumenda harum corporis quisquam dicta debitis, velit reprehenderit quam officia eveniet voluptates assumenda similique nobis delectus illo iure, tempore placeat expedita et.Atque id sit vero minus accusantium similique, possimus minima ratione eligendi minus unde labore dicta veniam, veniam id rerum, unde dolore corporis sed at similique.Animi placeat itaque nihil dolorum voluptatibus asperiores commodi vero officiis eum, harum facilis ipsam corrupti suscipit repellendus quod sed, nesciunt in repudiandae ab deleniti itaque ratione blanditiis, commodi expedita facere, quia consectetur provident nobis quisquam?Illo odit aspernatur ipsa molestias nostrum dolorum itaque neque,