Personal Information Leakage Detection in Conversations

Qiongkai Xu[1,2], **Lizhen Qu**[3]*, **Zeyu Gao**[1], **Gholamreza Haffari**[3]
[1]The Australian National University, Canberra, ACT, Australia
[2]Data61 CSIRO, Canberra, ACT, Australia
[3]Monash University, Clayton, VIC, Australia
{Qiongkai.Xu,Zeyu.Gao}@anu.edu.au
{Lizhen.Qu,Gholamreza.Haffari}@monash.edu

Abstract

The global market size of conversational assistants (chatbots) is expected to grow to USD 9.4 billion by 2024, according to MarketsandMarkets. Despite the wide use of chatbots, leakage of personal information through chatbots poses serious privacy concerns for their users. In this work, we propose to protect personal information by warning users of detected suspicious sentences generated by conversational assistants. The detection task is formulated as an alignment optimization problem and a new dataset PERSONA-LEAKAGE is collected for evaluation. In this paper, we propose two novel constrained alignment models, which consistently outperform baseline methods on PERSONA-LEAKAGE[1]. Moreover, we conduct analysis on the behavior of recently proposed personalized chit-chat dialogue systems. The empirical results show that those systems suffer more from personal information disclosure than the widely used Seq2Seq model and the language model. In those cases, a significant number of information leaking utterances can be detected by our models with high precision.

1 Introduction

According to Opus Research[2], 4.5 billion dollars will be invested in conversational assistants (chatbots) by 2021. Among diverse types of chatbots, Google Duplex, first introduced at Google I/O 2018, represents the kind of AI personal assistants (PAs) that *act on behalf of* people to perform simple tasks, such as making reservations at restaurants and hair salons. In order to successfully complete those tasks, PAs are granted the access to personal information (PI) of their owners, such as number of children, working hours, home address, and vacation plans. Thus, these PAs pose privacy concerns when they communicate with real-life people, or other bots in natural language.

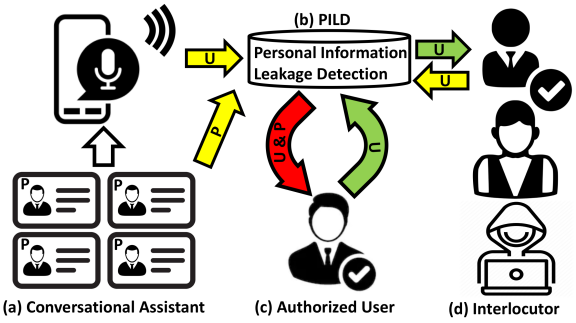


Figure 1: Given utterances (U) and personal information descriptions (P) from a conversational assistant (a), PILD module (b) detects risky utterances with corresponding personal information and sends a warning (red arrow) to an authorized user (c). The authorized user manually approve or reject the utterances. Then, only the approved utterances (green arrow) are sent to interlocutors (d) who could be authorized or malicious.

Another major source of personal information leakage is online social networks, which store a huge amount of possibly sensitive information on users and their interactions (Zhang et al., 2010). However, a recent study shows that *none* of the popular social network platforms (Facebook, Wechat, Google+, etc) have developed a perfectly non-leaky privacy protection mechanism (Yu et al., 2018). In addition, internet users (including a vast number of children and teenagers) often show a phenomenon called *privacy paradox*, which states that even users with high level of privacy concerns do not always take appropriate actions although those measures are fairly easy to perform (Norberg et al., 2007). As an unfortunate example, children's privacy is often unconsciously compromised by their parents' online behaviour, such as online posting and mes-

*Corresponding author

[1]The dataset and our model implementation is available at https://github.com/xuqiongkai/PILD.

[2]https://www.opus.global/media/44137/opus-q3-2018-report-eng.pdf

Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, pages 6567–6580,
November 16–20, 2020.

saging (Minkus et al., 2015).

An ideal privacy protection solution is *not* to stop using PAs or discourage online socialization, but to have the ability to *control the dissemination of personal information* (Yu et al., 2018). Personal information can be dispersed through various types of media. In this work, we focus on natural language utterances in conversations articulated by PAs or humans. The ways of controlling such textual information vary significantly w.r.t. platforms, PAs, user preferences, and social circles. Since there is no universally applicable control strategy, we take the first step towards privacy protection by designing a Personal Information Leakage Detection module (PILD) that *warns* users or *alerts* PAs whenever an utterance is associated with personal information, as illustrated in Figure 1. The warning module gives authorized users the capability to control information leakage from the start. Then, it is up to users and the design of PAs to decide how they deal with utterances leaking personal information. PAs will communicate with other interlocutors using secure or approved utterances.

We formulate detection of utterances causing personal information leakage as a text alignment problem, which aims to link information leaking utterances to the corresponding textual descriptions of personal information. We consider personal information provided in text, because i) user profiles on popular social network platforms include a significant proportion of textual descriptions, and ii) it is natural for users to share their information with PAs in natural language. Figure 2 demonstrates an example of aligning utterances in a dialogue with a set of personal information descriptions. Those red lines depict the ground-truth alignments between utterances and personal information descriptions. The true alignments are sparse as not all utterances leak personal information, e.g., U1, U3 and U6. Meanwhile, an utterance may be associated with more than one descriptions of personal information, e.g., U2 and U4, and vice versa.

In the absence of direct supervision signals, we explore low annotation-cost solutions to this text alignment problem by considering a weakly supervised setting. In this setting, we only know *who speaks what* and what are the PI descriptions of each interlocutor during training, without knowing true alignments. The additional challenges are imposed by the complex relationships between utterances and descriptions of PI, which could be sparse

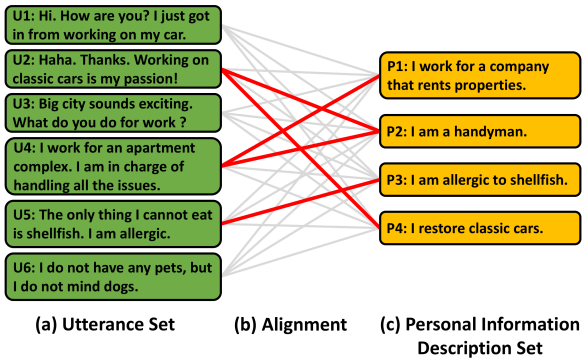


Figure 2: The alignment (b) of an utterance set (a) and a personal information description set (c) by a user. The matched sentence-level utterance-PI pairs are highlighted using red lines.

alignment, and one-to-one, one-to-many, many-to-one, or many-to-many mapping.

To address the aforementioned challenges, we propose two models SHARP-MAX and SPARSE-MAX by formulating the text alignment problem as constrained optimization problems. The training procedure takes the form of contrastive learning (Mnih and Kavukcuoglu, 2013; Dai and Lin, 2017). Herein, we encourage aligning an utterance with the descriptions of its interlocutor subject to sparsity constraints, while penalizing its alignments with those of other speakers. Thus, sentence-level alignments are not employed during training.

The main contributions are the following:

- We propose to protect privacy in conversation using PILD. Due to the lack of datasets for the new task, we construct a testing dataset PERSONA-LEAKAGE by extending the test set of the personalized dialogue corpus PERSONA (Zhang et al., 2018) with alignment annotations through crowdsourcing.

- Under weakly supervised setting, we propose two novel alignment models SHARP-MAX and SPARSE-MAX, which leverage coarse grained alignment signals to deliver sparse solutions. Our experiments on PERSONA-LEAKAGE show that our models achieve superior performance than competitive baselines.

- We empirically evaluated four representative dialogue models as PAs on PERSONA-LEAKAGE by letting them act as one of the interlocutors in a dialogue. We found that more advanced dialogue

models are prone to leak higher proportion of personal information of the interlocutors they represent. Our PILD module works well on recently proposed dialogue agents.

2 Alignment Models

In this section, we formally define the problem of PI leakage detection as text alignment between utterances and descriptions of PI in the weakly supervised setting, followed by presenting the architecture shared by the two proposed alignment models SPARSE-MAX and SHARP-MAX. The two models differ in the sparsity regularization for alignments during training. We then detail the training algorithms as well as how to derive the regularizers.

2.1 Problem Statement

A dialogue between two interlocutors A and B is composed of two sets of utterances U_A and U_B. The corresponding persona profiles P_A and P_B are two sets of PI descriptions. A personalized dialogue dataset $\mathcal{D} = \{\langle U_i, P_i\rangle | i = 1, 2, \cdots, N\}$ consists of $\langle U_i, P_i\rangle$ associated with the same interlocutor i in a conversation, where $U_i = \{u_{i,j} | j = 1, 2, \cdots, n_i\}$ and $P_i = \{p_{i,k} | k = 1, 2, \cdots, m_i\}$. In the weakly supervised setting, a $\langle U_i, P_i\rangle$ from the 'same interlocutor' provides a set-level training signal for learning an alignment between the utterance set and the PI description set. An alignment is a set of links between an utterances set and an description set. This can also be viewed as identifying the edges of a bipartite graph between the two sets of vertices U_i and P_i. In the absence of alignment annotation during training, we relax the problem by learning alignment strength between $u_{i,j}$ and $p_{i,k}$ as an association score $a_{i,j,k}$, which constitute an association matrix $\mathbf{A}_i \in \mathbb{R}^{n_i \times m_i}$ for each $\langle U_i, P_i\rangle$. Then, it is up to the system design of a PA or the preference of an interlocutor to decide if an association score indicates that $p_{i,k}$ is leaked through $u_{i,j}$. For example, one can check if $a_{i,j,k}$ is above a pre-specified threshold.

2.2 Model Architecture

Recent advances in pre-trained language models, such as BERT (Devlin et al., 2019), demonstrate their strengths of encoding semantic information into the produced text representations. Thus we apply a pre-trained language model $f(\cdot)$ (BERT in this work) to convert each utterance and each PI description into its representation vectors. As a widely accepted practice, we take the representation of the [CLS] token to represent an input text. Then, we apply a projection matrix $\mathbf{M}$ to map those vectors into a semantic space shared by utterances and PI descriptions,

$$\begin{aligned} \mathbf{r}_{i,j}^{(u)} &= \mathbf{M} \cdot f(u_{i,j}) \\ \mathbf{r}_{i,k}^{(p)} &= \mathbf{M} \cdot f(p_{i,k}) \end{aligned} \tag{1}$$

The association score between an utterance $u_{i,j}$ and a PI description $p_{i,k}$ is calculated by the cosine similarity between their representations, where $\langle \cdot, \cdot \rangle$ denotes the inner product of two vectors,

$$a_{i,j,k} = \frac{\langle \mathbf{r}_{i,j}^{(u)}, \mathbf{r}_{i,k}^{(p)} \rangle}{\|\mathbf{r}_{i,j}^{(u)}\| \|\mathbf{r}_{i,k}^{(p)}\|} \tag{2}$$

As we freeze the parameters of BERT in both training and testing, the only tunable parameters of this model is the matrix $\mathbf{M}$.

2.3 Model Training

Learning an association matrix between an utterance set and a PI description set in the weakly supervised setting imposes two challenges. First, there is no ground-truth label to guide the alignment training. Second, an utterance may indicate zero, one, or multiple PI descriptions, while a PI description may also be associated with varying number of utterances.

Loss. To address the first challenge, we observe that i) a linked utterance-PI pair has high semantic relatedness; ii) the utterances in a dialogue are much more likely to correlate with the PI of its interlocutors than that of other interlocutors. The latter observation provides set-level alignment signals for contrastive learning. In light of this, we maximize the set-level aggregated associated scores for utterance-PI pairs from the same interlocutors $\langle U_i, P_i\rangle$, while minimizing those scores for the pairs from different interlocutors $\langle U_i, \widehat{P}\rangle$ and $\langle \widehat{U}, P_i\rangle$.

The second challenge imposes sparsity over the links in alignments. As it is difficult to enforce representation based cosine similarity values to approach zero, we introduce an alignment weight $w_{i,j,k}$ for each utterance-PI pair during training. The weight matrix $\mathbf{W}_i = \{w_{i,j,k}\}_{n_i \times m_i}$ puts a focus on the more reliable utterance-PI pairs and reduces the influence from irrelevant links. Then,

the similarity between U_i and P_i is the weighted sum of all elements in $\mathbf{A}_i$.

$$sim(U_i, P_i) = \mathbf{W}_i \odot \mathbf{A}_i = \sum_j \sum_k w_{i,j,k} a_{i,j,k} \tag{3}$$

where $\odot$ denotes hadamard product. High weights in $\mathbf{W}_i$ will enhance the corresponding association scores during training, while low weights or zeros in $\mathbf{W}_i$ discourage participation of those corresponding scores.

By putting two ideas together, the loss for the ith training sample is defined as:

$$\begin{aligned}\mathcal{L}(U_i, P_i) = \\ \max\{0, \alpha - sim(U_i, P_i) + sim(U_i, \widehat{P})\} + \\ \max\{0, \alpha - sim(U_i, P_i) + sim(\widehat{U}, P_i)\}\end{aligned} \tag{4}$$

where $\widehat{U}$ and $\widehat{P}$ are randomly sampled from $\mathcal{D}$, α is a hyper-parameter controlling the margin of the loss. Then the loss on training set is the sum of all example losses $\mathcal{L}(\mathcal{D}) = \sum_{i=1}^{N} \mathcal{L}(U_i, P_i)$.

Sparsity. The two models SHARP-MAX and SPARSE-MAX differ in the regularizers used in $sim(U_i, P_i)$ for learning *sparse* weight matrices $\mathbf{W}_i$. The matrices $\mathbf{W}_i$ are expected to assign zeros or low weights to irrelevant pairs, while assigning high weights to the aligned pairs. They are formulated as a constrained optimization problem of the following form,

$$\begin{aligned}sim(S_i, P_i) = \max_{\mathbf{W}_i}\{\mathbf{W}_i \odot \mathbf{A}_i + \gamma H(\mathbf{W}_i)\} \\ s.t. \sum_j \sum_k w_{i,j,k} = 1, \forall j, k; w_{i,j,k} \in [0, 1]\end{aligned} \tag{5}$$

where $H(\cdot)$ is a regularization term that determines the sparsity of $\mathbf{W}_i$, and $\gamma \in \mathbb{R}^+$ adjusts the degree of regularization. If $\gamma \to 0$, the solution of the above problem is to assign the weight 1 to the maximal value in $\mathbf{A}_i$. As we expect more than one links in an alignment, the regularizer should encourage more non-zero entries in $\mathbf{W}_i$. If $\gamma \to +\infty$, the solution is weights with equal values, which aggregates $\mathbf{A}_i$ by averaging all association scores.

SHARP-MAX utilizes entropy as the regularizer because uniform distribution achieves the maximum of entropy. In another words, this term encourages similar entries in $\mathbf{W}_i$.

Proposition 1. Let $\gamma \in \mathbb{R}^+$

$$H(\mathbf{W}_i) = -\sum_{j,k} w_{i,j,k} \log w_{i,j,k}$$

in Eq. (5), the solution of $\mathbf{W}_i$ is the following softmax function with temperature γ,

$$w_{i,j,k} = \frac{\exp(a_{i,j,k}/\gamma)}{\sum_j \sum_k \exp(a_{i,j,k}/\gamma)} \tag{6}$$

Proof Idea: The solution is derived by solving the Lagrangian of Eq. (5):

$$\begin{aligned}\mathcal{L}(\mathbf{W}_i, \lambda) = &\sum_j \sum_k w_{i,j,k} a_{i,j,k} \\ &- \gamma \sum_j \sum_k w_{i,j,k} \log w_{i,j,k} \\ &+ \lambda(1 - \sum_j \sum_k w_{i,j,k})\end{aligned} \tag{7}$$

Note that, when the temperature with $\gamma < 1$ is sufficiently small, the optimal $\mathbf{W}_i$ enlarges the differences of the values in $\mathbf{A}_i$ (SHARP-MAX). If $\gamma = 1$, we got the conventional softmax, which is also referred to as **SOFT-MAX** in our experiments.

SPARSE-MAX considers the squared loss on $\mathbf{W}_i$ as the regularizer, as it controls the sparsity of the matrix by encouraging equal contributions.

Proposition 2. Let $\gamma = 1$,

$$H(\mathbf{W}_i) = -\frac{1}{2} \sum_{j,k} w_{i,j,k}^2$$

in Eq. (5), the solution of $\mathbf{W}_i$ is the *sparsemax* of $\mathbf{A}_i$ (SPARSE-MAX) (Martins and Astudillo, 2016).

$$w_{i,j,k} = [a_{i,j,k} - \tau(\mathbf{A}_i)]_+ \tag{8}$$

where $\tau(\cdot)$ is a dynamic threshold function and $[t]_+ = \max\{0, t\}$.

3 Experimental Setup

3.1 PERSONA-LEAKAGE Dataset

In order to evaluate models under the weakly supervised setting, we constructed a dataset PERSONA-LEAKAGE as the test set by annotating the test set of the personalized dialogue corpus PERSONA (Zhang et al., 2018). In that corpus, each dialogue is conversed between two human interlocutors, where each interlocutor is characterized by three to five descriptions of PI. A description of PI describes one aspect of that person, e.g., 'I am a handyman'. For each dialogue, we collected link candidates by pairing each utterance of a interlocutor to each description of his PI. As a result, we constructed a set of link candidates for each

interlocutor in a dialogue. For each link candidate, we asked three annotators to judge if the utterance indicates the corresponding PI description. A candidate was considered as *aligned* if at least two annotators agreed on that decision. In total, we annotated alignments for 968 dialogues, in which there are 6,894 aligned utterance-PI pairs out of 67,601 candidate pairs.

Moreover, in order to understand the user perception on sensitivity of PI, we collected a set of all possible PI descriptions in test and dev set of PERSONA, and asked five annotators to judge if the descriptions were sensitive or not. A PI description is considered as sensitive if annotators would suggest not to share it with strangers, given that it describes their friends. We collected 306 descriptions (31.48% among all 972 descriptions) with more than 2 sensitive annotations[3].

3.2 Baselines

We apply the scoring function of two widely used information retrieval (IR) methods **TF-IDF** and **BM25** (Manning et al., 2008; Robertson and Zaragoza, 2009), and the most recent **BERT**-based IR (Dai and Callan, 2019) to measure the association between a PI description and an utterance.

We also consider the following competitive alignment models proposed in recent works.

- **MEAN** averages the contribution of association matrix, namely uniform weights ($\frac{1}{n_i \cdot m_i}$). We consider MEAN as the solution of a special case of our optimization problem with $\gamma \to +\infty$.
- **Avg-Max** (Lee et al., 2018) uses the average of the maximum similarity scores for all PI descriptions (**Avg-Max-P**) or utterances (**Avg-Max-U**).
- **LSAP** (Linear sum assignment problem) (Hessel et al., 2019) optimizes hard alignments, where each row and column has less or equal than one link, i.e. $\forall j, \sum_k w_{i,j,k} \leq 1; \forall k, \sum_j w_{i,j,k} \leq 1; \forall j,k, w_{i,j,k} \in \{0,1\}$.
- **OPT** (Optimal Transport) (Kusner et al., 2015) optimizes soft alignments, where weights are in $[0,1]$ and sums of the weights on each column and row are less or equal to one, i.e. $\forall j, \sum_k w_{i,j,k} \leq 1; \forall k, \sum_j w_{i,j,k} \leq 1; \forall j,k, w_{i,j,k} \in [0,1]$.

[3]Appendix B describes more details about data collection.

The weights of all alignment models are normalized to the sum of one.

3.3 Model Setting

In order to have a fair comparison, all alignment models share the same deep learning architecture which is composed of i) a pre-trained text representation model (BERT), ii) a learnable linear transformation layer, and iii) a weight computation module without back-propagation. The dimensions of pre-trained and final text representations are 768 and 256, respectively. We use Adam as optimizer for all experiments that require training. According to our preliminary experiments, we set learning rate to 0.01, batch size to 128 and train 200 epochs for all experiments.[4] We consider the hyper-parameters $\alpha \in \{0.0, 0.1, 0.2, 0.4, 0.8\}$ for all models and $\gamma \in \{1/4, 1/5, 1/6, 1/7, 1/8\}$ for Sharp-Max.

We evaluate the models by testing whether the alignment links between sets are correctly retrieved from all candidates links, following (Hessel et al., 2019). Given the ground-truth alignment between two sets, we evaluate the association matrix $\mathbf{A}_i$, by using precision at K (P@K)[5], R-Precision (Rprec), normalized discounted cumulative gain (NDCG) and mean average precision (MAP)[6]. In addition, we use Hellinger Distance (H-Dist) (Oosterhoff and van Zwet, 2012) $\frac{1}{N}\sum_i \frac{1}{2}\sum_{j,k}(\sqrt{w_{i,j,k}} - \sqrt{g_{i,j,k}})^2$ to quantify the matching rate of alignment weights $\mathbf{W}_i$ with ground-truth alignment weights $\mathbf{G}_i = \{g_{i,j,k}\}_{n_i \times m_i}$, where $g_{i,j,k}$ is normalized over j,k to sum to one.

3.4 Collection of Human-Bot Dialogue

To evaluate the performance of our model on chatbots, we collect human-bot dialogues using SOTA personalized chatbots and their competitors:

- $\mathcal{P}^2$ **Bot** (Liu et al., 2020) achieved SOTA performance on automatic metrics by incorporating mutual persona perception. $\mathcal{P}^2$ **Bot (w/ Persona)** and $\mathcal{P}^2$ **Bot (w/o Persona)** are models with and without personal information as input when generating responses.
- **Lost-In-Conversation** (Dinan et al., 2019)

[4]We have explored learning rate in $\mathcal{R} = \{0.1, 0.01, 0.001\}$ and number of training epochs in $\mathcal{E} = \{25, 50, 100, 200, 400\}$.

[5]As the average and maximum number of alignment links are 3.56 and 9 in our corpus, we choose $K \in \{1, 3, 5\}$.

[6]https://trec.nist.gov/trec_eval/

Model	P@1	P@3	P@5	Rprec	NDCG	MAP	H-Dist
RANDOM	0.1124	0.1050	0.1099	0.1107	0.4349	0.1919	N/A
TF-IDF	0.6716	0.5434	0.4294	0.5088	0.7548	0.5832	N/A
BM25	0.6824	0.5364	0.4207	0.4988	0.7535	0.5785	N/A
BERT	0.5923	0.4149	0.3257	0.3762	0.6789	0.4677	N/A
MEAN ($\alpha = 0.1$)	0.7573	0.6361	0.5230	0.6178	0.8331	0.7097	0.6801
Avg-Max-P ($\alpha = 0.4$)	0.7856	0.6748	0.5545	0.6566	0.8561	0.7486	0.3797
Avg-Max-U ($\alpha = 0.2$)	0.7785	0.6647	0.5452	0.6467	0.8493	0.7369	0.4680
OPT ($\alpha = 0.2$)	0.7725	0.6605	0.5448	0.6434	0.8470	0.7340	0.4822
LSAP ($\alpha = 0.4$)	0.7780	0.6670	0.5495	0.6522	0.8529	0.7434	0.4084
SOFT-MAX ($\alpha = 0.1$)	0.7676	0.6554	0.5341	0.6350	0.8421	0.7247	0.6042
SHARP-MAX ($\alpha = 0.4, \gamma = 1/6$)	0.7942	0.6763	0.5517	0.6618	0.8577	0.7499	0.3208
SPARSE-MAX ($\alpha = 0.4$)	**0.7970**	**0.6839**	**0.5597**	**0.6695**	**0.8612**	**0.7562**	**0.3032**

Table 1: Experimental results of random guess (RANDOM), unsupervised IR models (TF-IDF, BM25, and BERT), baseline alignment models (MEAN, Avg-Max-U, Avg-Max-P, OPT and LSAP), and our proposed models (Soft-Max, Sparse-Max and Sharp-Max).

topped the human evaluations in ConvAI2 by fine-tuning a pre-trained language model GPT.

- **Seq2Seq-Attn** (Zhang et al., 2018) is an LSTM-based sequence-to-sequence model incorporateing persona via an attention module.
- **Language Model** (Zhang et al., 2018) is an LSTM-based language module for dialogue.

For each chatbot, we provided interlocutor A's dialogue history as input and bot responded as interlocutor B. We performed 60 dialogues and collected 770 utterances for each chatbot. The responses by those chatbots are analyzed in three dimensions.

- Personal Information Engagement (**PIE**) is the proportion of the utterances leaking PI,

  $$|\text{Utterances have PI Leakage}|/|\text{All Utterances}|$$

- Disclosed PI Sensitivity (**DPS**) is the ratio of sensitive PI descriptions to the leaked ones,

  $$\frac{|\text{Sensitive Disclosed PI descriptions}|}{|\text{Disclosed PI descriptions}|}$$

  .

- Hits-at-K (**Hits@K**) is the percentile of the leaked PI that can be retrieved from top $K = 5/10$ results using alignment models.

Perplexity (**PPL**) and uni-gram **F1** are supplementary metrics that reflect the performance of bots (Liu et al., 2020).

4 Empirical Results and Analysis

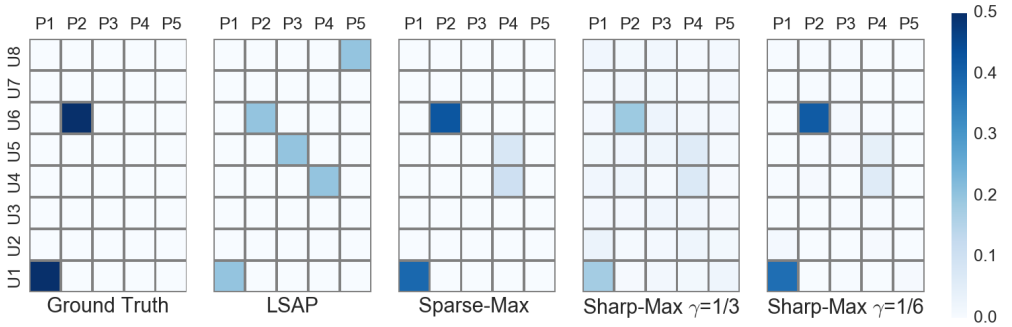
In this section, we analyze our experimental results. Our experiments are designed to answer the following research questions (**RQ**s),

- **RQ1:** How well do our alignment models perform, in comparison with the competitive baselines?
- **RQ2:** Why do our alignment models outperform the baselines?
- **RQ3:** Do the SOTA chatbots disclose PI in dialogues, and are they sensitive? Can we use our alignment models to capture the leakage?

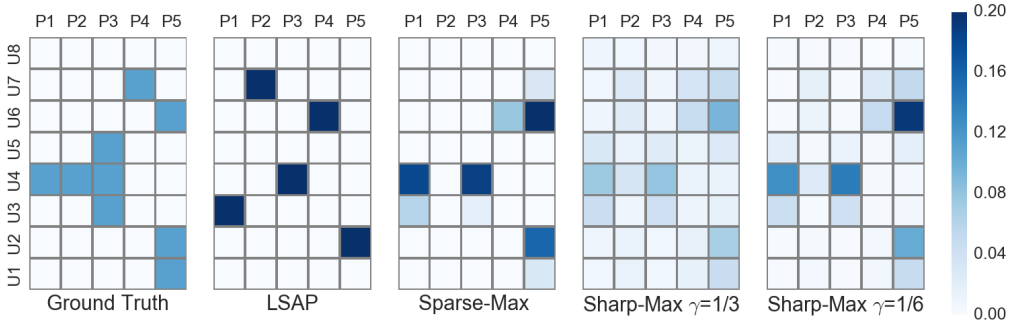
4.1 Model Comparison on PERSONA-LEAKAGE

We compare our alignment models, SHARP-MAX and SPARSE-MAX, with IR baselines and alignment baselines, in Table 1[7]. The proposed model consistently outperform baseline methods, indicating the effectiveness of our methods. H-dist is strongly correlated to other metrics, because better alignments lead to better H-dist. IR models significantly outperform random guess, showing that semantic information provided in utterances and descriptions provides strong guidance on inference. Although the naive MEAN does not enforce sparsity during training, it outperforms the unsupervised IR models with a large margin, more than

[7] Appendix C provides more details about hyper-parameter selection.



(a) A case with sparse alignment.



(b) A case with dense alignment.

Figure 3: Comparison of weights assigned to candidates between utterances (U1-U8) and personal information descriptions (P1-P5). (a) case 12 and (b) case 85 are test cases with sparse and dense alignments, respectively. The alignment weights of Ground Truth and LSAP are all normalized to the sum of one for each case.

10% for all scores, showing that coarse grain signal is effective for learning semantic relevant for the PI leakage. Avg-Max, OPT and LSAP further outperform MEAN with a margin more than 2% for most of the metrics, as they apply the sparsity constraints in order to focus on aligned utterances and PI descriptions during training. Although these approaches set up competitive baselines on our task, SHARP-MAX and SPARSE-MAX achieve consistent improvement on all evaluation metrics. As SPARSE-MAX cuts off the weights of irrelevant pairs, it performs the best.

4.2 Analysis on Alignment Model

We visualize the association scores of each alignment model in Figure 3, in order to qualitatively demonstrate the strengths of our models. LSAP attempts to assign a fixed number of aligned pairs, i.e. $\min\{n_i, m_i\}$, which will lead to unavoidable false positive alignment for sparse cases (U8-P5, U5-P3 and U4-P4, in Figure 3a LSAP) and false negative alignment for dense cases (U4-P1 and U4-P2, in Figure 3b LSAP). Avg-Max-P and Avg-Max-U also hold the similar drawback as the number of aligned pairs is exact the number of columns or rows, while does not depend on the cases. In contrast, SPARSE-MAX and SHARP-MAX manage to adapt the number of 'aligned pairs' (deep colored), therefore achieve alignments closer to the ground truth. For SHARP-MAX, we can adjust the sharpness of the weight matrix using sharpness parameter γ. Using sharper model with lower γ manages to alleviate the influence of the pairs with relatively low similarity scores. For SPARSE-MAX, more deterministic alignments are achieved by cutting off pairs with low association scores. Although SHARP-MAX and SPARSE-MAX do not differ much in terms of empirical performance, they are driven by different theories of regularization. The comparison between these two solutions proposed by us helps draw a conclusion that the similarity function should be designed to find a proper degree of sparsity, which does not depend much on a particular choice of regularizer.

4.3 Analysis on Personalized Chatbots

In this section, we analyze the engagement and sensitivity of chatbots in human-bot conversations. The experiments are designed to show the risk of privacy leakage when using current chatbot models. For all generated utterances, we retrieved top 10 relevant PI using SPARSE-MAX. Then we asked an-

Model	PIE	DPS	Hits@5/10	PPL ↓	F1 ↑
Language Model	02.13	06.45	29.03 / 32.26	51.61	13.59
Seq2Seq-Attn	04.39	06.54	18.64 / 22.03	39.54	15.52
$\mathcal{P}^2$ **Bot (w/o Persona)**	08.94	10.77	51.54 / 56.15	-	17.77
Lost-In-Conversation	14.68	09.39	79.34 / 82.63	-	16.83
$\mathcal{P}^2$ **Bot (w/ Persona)**	37.19	16.86	73.62 / 77.04	18.89	19.08
Human	43.83	27.75	55.07 / 66.52	-	-

Table 2: Analysis on the responses of personalized chatbots and human interlocutors.

notators to select the leaked ones from the retrieved PI descriptions. Three annotators are asked to indicate if the utterances leak those PI descriptions. The results are summarized in Table 2. Compared with bots *without PI* as inputs, such as **Language Model** and $\mathcal{P}^2$ **Bot (w/o Persona)**, the bots *with PI* as input, namely **Lost-In-Conversation** and $\mathcal{P}^2$ **Bot (w/ Persona)**, tend to acquire higher PIE with significantly higher magnitude. PIE of $\mathcal{P}^2$ **Bot** even approaches that score of human interlocutors. DPS is correlated to PIE showing that bots with higher PIE generally disclose higher portions of sensitive PI. Although higher PIE and DPS for the chatbots *with PI* as input is expected, there is also a significant proportion of leakage for the bots without PI as input, e.g., $\mathcal{P}^2$ **Bot (w/o Persona)**. This raises serious privacy concerns in future research on PAs.

Furthermore, Hit@K measures the ability of our system for detecting PI leakage. As a warning module, our model SPARSE-MAX manages to detect most of the utterances leaking PI[8]. Our system achieves around 80% of Hit@10 on the responses generated by the two most recent and advanced chatbots, **Lost-In-Conversation** and $\mathcal{P}^2$ **Bot (w/ Persona)**.

5 Related Work

Recently, privacy and fairness started to attract more and more attention from NLP community. Sensitive information was removed from latent representations via adversarial training (Li et al., 2018; Elazar and Goldberg, 2018) and differential privacy (Fernandes et al., 2019), achieving fair decisions. Privacy-aware text rewriting methods suggested to generate new sentences with less sensitive information (Xu et al., 2019a; Emmery et al., 2018; Xu et al., 2019b; Strengers et al., 2020). Our work serves as a component that detects the sentences requires rewriting. Another line of research aims to identify mentions of pre-defined semantic categories indicating sensitive information in text (Microsoft; Bevendorff et al., 2019), such as bank account and phone number. In our setting, sensitive information can be expressed in any syntactic structures, including events conveyed in whole sentences, such as "I have got less than 5 hours of sleep each night for years" is associated with the persona "I have sleep disorders for many years.". The setting of our work is more general, as we focus on open-domain personal information written in natural language, which is not limited to mentions of fixed semantic categories in sentences or sensitivity labels of sentences.

Our work places an emphasis on privacy concerns in conversations (Huang et al., 2020; Ischen et al., 2019; Gao et al., 2018; Tur et al., 2018; Muthukrishnan et al., 2017). In recent research, several works have attempted to improve the engagement and diversity of chit-chat dialogue system (Liu et al., 2020; Tigunova et al., 2019; Wolf et al., 2019; Zhang et al., 2018) and goal-oriented dialogue system (Luo et al., 2019; Zhang et al., 2019). With the rapid development of personalized dialogue systems, PILD module is expected to address the privacy concerns (Ischen et al., 2019). Welleck et al. (2019) improved the coherence and consistency of a dialogue using Natural Language Inference (NLI) (Bowman et al., 2015). Dialogue-NLI dataset could be utilized to train retrieval models, however, it does not directly address the privacy concerns. In contrast, our dataset i) considers all possible leakage pairs, and ii) includes sensitivity annotations of all PI descriptions.

Our problem setting was inspired by an image-sentence alignment problem, given pairs of image sets and documents (Hessel et al., 2019). Similar problems were also explored in the context of align-

[8] According to our preliminary experiments in Appendix D, SPARSE-MAX achieves the best Hits on the whole test set of PERSONA-LEAKAGE.

ing image fragments with words (Lee et al., 2018; Jiang et al., 2015). In this paper, we considered utterances and PI descriptions from the *same* interlocutor as coarse-grained alignment signals, which are in the same modality.

6 Conclusions and Future Work

We formulate protection of personal information in conversations as a weakly supervised alignment between personal information and dialogue utterances. To tackle this task, we proposed two new alignment models and created a dataset PERSONA-LEAKAGE for evaluation. Our experimental results demonstrate the effectiveness of our methods in comparison with the competitive baselines on that dataset. Further analysis on human-bot dialogue performance demonstrated the potential privacy risks with advanced personalized dialogue techniques. This work is the first step towards fully preventing leakage of privacy in text, which still requires PAs or users to select and hide sensitive information. We hope this work and the dataset will pave the way for the research on privacy leakage in conversations. In the future, we will explore full-fledged solutions to address the privacy concerns of both humans and dialogue systems.

Acknowledgments

This work is partially supported by an ARC Future Fellowship (FT190100039) to G. H.

References

Janek Bevendorff, Martin Potthast, Matthias Hagen, and Benno Stein. 2019. Heuristic authorship obfuscation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1098–1108.

Samuel Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642.

Bo Dai and Dahua Lin. 2017. Contrastive learning for image captioning. In *Advances in Neural Information Processing Systems*, pages 898–907.

Zhuyun Dai and Jamie Callan. 2019. Deeper text understanding for ir with contextual neural language modeling. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 985–988.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 NAACL*, pages 4171–4186.

Emily Dinan, Varvara Logacheva, Valentin Malykh, Alexander Miller, Kurt Shuster, Jack Urbanek, Douwe Kiela, Arthur Szlam, Iulian Serban, Ryan Lowe, et al. 2019. The second conversational intelligence challenge (convai2). *arXiv preprint arXiv:1902.00098*.

Yanai Elazar and Yoav Goldberg. 2018. Adversarial removal of demographic attributes from text data. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 11–21.

C Emmery, E Manjavacas, and G Chrupała. 2018. Style obfuscation by invariance. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 984–996.

Natasha Fernandes, Mark Dras, and Annabelle McIver. 2019. Generalised differential privacy for text document processing. In *International Conference on Principles of Security and Trust*, pages 123–148. Springer.

Jianfeng Gao, Michel Galley, and Lihong Li. 2018. Neural approaches to conversational ai. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pages 1371–1374.

Jack Hessel, Lillian Lee, and David Mimno. 2019. Unsupervised discovery of multimodal links in multi-image, multi-sentence documents. In *Proceedings of the 2019 EMNLP-IJCNLP*, pages 2034–2045.

Minlie Huang, Xiaoyan Zhu, and Jianfeng Gao. 2020. Challenges in building intelligent open-domain dialog systems. *ACM Transactions on Information Systems (TOIS)*, 38(3):1–32.

Carolin Ischen, Theo Araujo, Hilde Voorveld, Guda van Noort, and Edith Smit. 2019. Privacy concerns in chatbot interactions. In *International Workshop on Chatbot Research and Design*, pages 34–48. Springer.

Xinyang Jiang, Fei Wu, Xi Li, Zhou Zhao, Weiming Lu, Siliang Tang, and Yueting Zhuang. 2015. Deep compositional cross-modal learning to rank via local-global alignment. In *Proceedings of the 23rd ACM international conference on Multimedia*, pages 69–78.

Matt Kusner, Yu Sun, Nicholas Kolkin, and Kilian Weinberger. 2015. From word embeddings to document distances. In *Proceedings of the 33rd ICML*, pages 957–966.

Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. 2018. Stacked cross attention for image-text matching. In *Proceedings of the European Conference on Computer Vision*, pages 201–216.

Yitong Li, Timothy Baldwin, and Trevor Cohn. 2018. Towards robust and privacy-preserving text representations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 25–30.

Qian Liu, Yihong Chen, Bei Chen, Jian-Guang Lou, Zixuan Chen, Bin Zhou, and Dongmei Zhang. 2020. You impress me: Dialogue generation via mutual persona perception. *arXiv preprint arXiv:2004.05388*.

Liangchen Luo, Wenhao Huang, Qi Zeng, Zaiqing Nie, and Xu Sun. 2019. Learning personalized end-to-end goal-oriented dialog. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6794–6801.

Christopher D Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to information retrieval*. Cambridge university press.

Andre Martins and Ramon Astudillo. 2016. From softmax to sparsemax: A sparse model of attention and multi-label classification. In *International Conference on Machine Learning*, pages 1614–1623.

Microsoft. Presidio - data protection and anonymization api.

Tehila Minkus, Kelvin Liu, and Keith W Ross. 2015. Children seen but not heard: When parents compromise children's online privacy. In *Proceedings of the 24th International Conference on World Wide Web*, pages 776–786.

Andriy Mnih and Koray Kavukcuoglu. 2013. Learning word embeddings efficiently with noise-contrastive estimation. In *Advances in neural information processing systems*, pages 2265–2273.

Muthu Muthukrishnan, Andrew Tomkins, Larry Heck, Alborz Geramifard, and Deepak Agarwal. 2017. The future of artificially intelligent assistants. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 33–34.

Patricia A Norberg, Daniel R Horne, and David A Horne. 2007. The privacy paradox: Personal information disclosure intentions versus behaviors. *Journal of consumer affairs*, 41(1):100–126.

Jacobus Oosterhoff and Willem R van Zwet. 2012. A note on contiguity and hellinger distance. In *Selected Works of Willem van Zwet*, pages 63–72. Springer.

Stephen Robertson and Hugo Zaragoza. 2009. *The probabilistic relevance framework: BM25 and beyond*. Now Publishers Inc.

Yolande Strengers, Lizhen Qu, Qiongkai Xu, and Jarrod Knibbe. 2020. Adhering, steering, and queering: Treatment of gender in natural language generation. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–14.

Anna Tigunova, Andrew Yates, Paramita Mirza, and Gerhard Weikum. 2019. Listening between the lines: Learning personal attributes from conversations. In *The World Wide Web Conference*, pages 1818–1828.

Gokhan Tur, Asli Celikyilmaz, Xiaodong He, Dilek Hakkani-Tür, and Li Deng. 2018. Deep learning in conversational language understanding. In *Deep Learning in Natural Language Processing*, pages 23–48. Springer.

Sean Welleck, Jason Weston, Arthur Szlam, and Kyunghyun Cho. 2019. Dialogue natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3731–3741.

Thomas Wolf, Victor Sanh, Julien Chaumond, and Clement Delangue. 2019. Transfertransfo: A transfer learning approach for neural network based conversational agents. *ArXiv*, abs/1901.08149.

Qiongkai Xu, Lizhen Qu, Chenchen Xu, and Ran Cui. 2019a. Privacy-aware text rewriting. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 247–257.

Qiongkai Xu, Chenchen Xu, and Lizhen Qu. 2019b. Alter: Auxiliary text rewriting tool for natural language generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations*, pages 13–18.

Lingjing Yu, Sri Mounica Motipalli, Dongwon Lee, Peng Liu, Heng Xu, Qingyun Liu, Jianlong Tan, and Bo Luo. 2018. My friend leaks my privacy: Modeling and analyzing privacy in social networks. In *Proceedings of the 23nd ACM on Symposium on Access Control Models and Technologies*, pages 93–104.

Bowen Zhang, Xiaofei Xu, Xutao Li, Yunming Ye, Xiaojun Chen, and Lianjie Sun. 2019. Learning personalized end-to-end task-oriented dialogue generation. In *CCF International Conference on Natural Language Processing and Chinese Computing*, pages 55–66. Springer.

Chi Zhang, Jinyuan Sun, Xiaoyan Zhu, and Yuguang Fang. 2010. Privacy and security for online social networks: challenges and opportunities. *IEEE network*, 24(4):13–18.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th ACL*, pages 2204–2213.

A Malicious Attack on Siri

We conducted an experiment using Siri installed in iPad pro, with iPadOS version 13.3.1 released in January 28 2020. An unauthorized user manage to acquire the owner's personal information by asking Siri questions. The responses by Siri are demonstrated in Figure 4. User name and home address of the facility devide owner is disclosed to the attacker, when asked *'Where do I live?'*. Name, partner and home address of the owner's parents are unveiled, when asked *'Who is my father?'*. Although Siri represented the answer in form of contact cards, we argue that such risky reactions by personal assistants could appear in natural language responses as well.

B Details for Data Collection

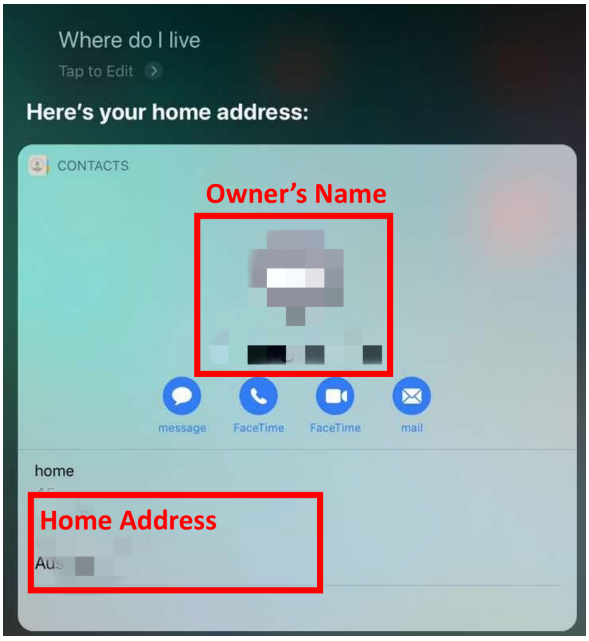
Starting from test set of PERSONA, our dataset basically tops up two annotations on test sets, alignment annotations on utterance-persona pairs and sensitivity annotations on all personal information statements. For both parts, we use Amazon Mechanical Turk (MTurk)[9] for crowdsourcing. We only accept results from the qualified annotators that i) have more than 90% HIT acceptance rate, ii) have finished more than 100 HITs, iii) locate in America. For further quality control, we reject 2.1% and 2.0% unreliable HITs for alignment annotation and sensitivity annotation respectively by automatically rejecting HITs that are i) not completed or ii) inconsistent in answers.

For alignment annotations, annotators were instructed to *"find the personal descriptions leaked in a conversation"* by *"select if the sentence indicates any of the provided personal descriptions or none of them"*, see task screenshot in Figure 5.

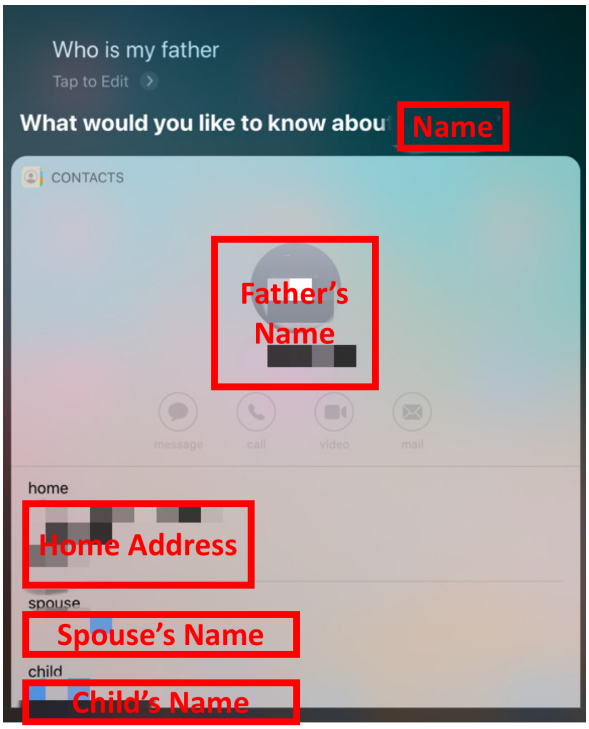
For sensitivity annotations, annotators were instructed to *"give advice to a friend who belongs to a vulnerable group"*, see task screenshot in Figure 6. Sensitive information is defined as the one that *"your friend rather not let strangers know"*.

- **Sensitive:** In most cases, your friend would rather not to tell a stranger such information. Otherwise it will do more harm than good if the information is utilized by malicious people.
- **Non-sensitive:** In most cases, it is safe for your friend to share such information with strangers.

[9] https://requester.mturk.com/



(a) Malicious Attack 1: *'Where do I live?'*.



(b) Malicious Attack 2: *'Who is my father?'*.

Figure 4: Screenshots of Siri's responses to an unauthorized user, when it is inquired *'Where do I live?'* and *'Who is my father?'*. The sensitive personal information is blurred by mosaics.

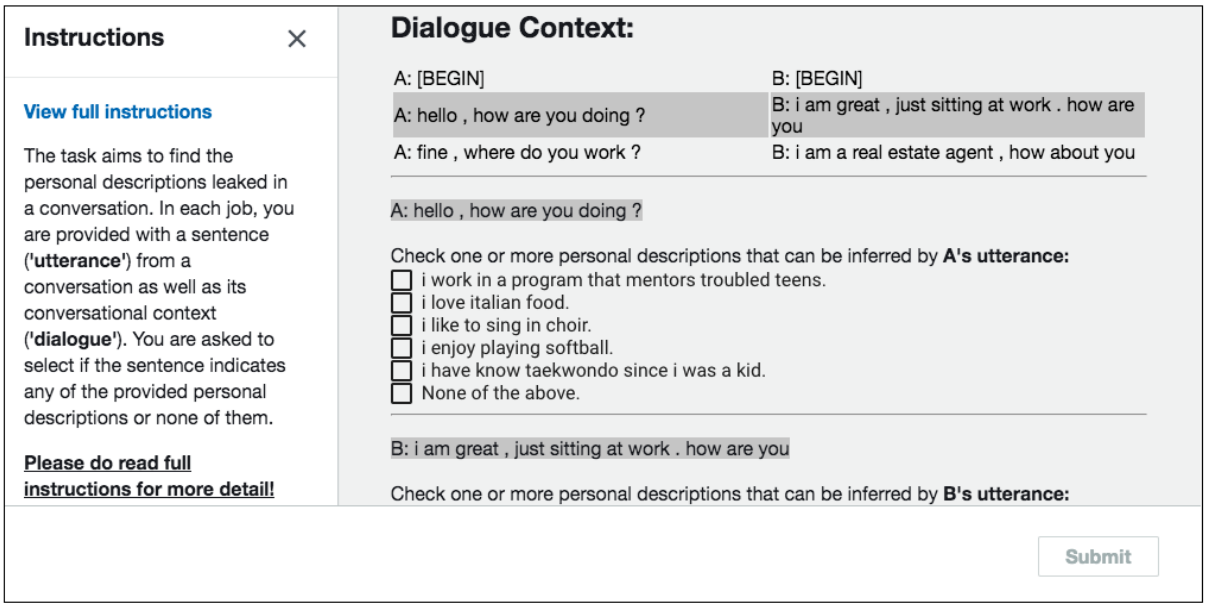


Figure 5: Task screenshot for utterance-persona alignment annotation.

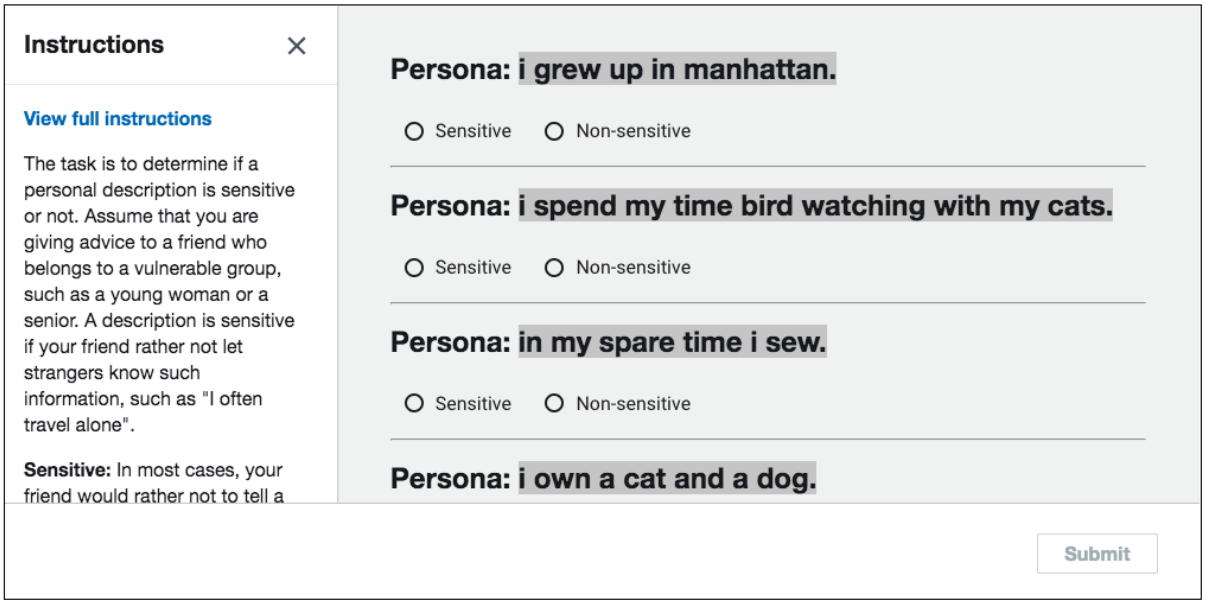


Figure 6: Task screenshot for personal information sensitivity annotation.

C Hyper-parameter Selection

We provide details about the hyper-parameter selection for baseline alignment models and our models in Table 3. More details about Sharp-Max is demonstrated in Table 4.

D Hits on human-human dialogue

We compare alignment models on Dialogue60, a subset used in our paper, and DialogueTest, the whole test set of *PERSONA-LEAKAGE*. Overall, Sparse-Max achieves the best performance.

Model	P@1	P@3	P@5	Rprec	NDCG	MAP
MEAN ($\alpha = 0.0$)	0.7329	0.5984	0.4780	0.5632	0.8014	0.6554
MEAN ($\alpha = 0.1$)‡	**0.7573**	**0.6361**	**0.5230**	**0.6178**	**0.8331**	**0.7097**
MEAN ($\alpha = 0.2$)	0.7096	0.5963	0.4898	0.5653	0.8029	0.6618
MEAN ($\alpha = 0.4$)	0.6080	0.5009	0.4125	0.4670	0.7395	0.5643
MEAN ($\alpha = 0.8$)	0.4973	0.4106	0.3501	0.3858	0.6797	0.4817
Avg-Max-P ($\alpha = 0.1$)	0.7769	0.6620	0.5370	0.6440	0.8470	0.7323
Avg-Max-P ($\alpha = 0.2$)	0.7823	0.6672	0.5483	0.6550	0.8526	0.7426
Avg-Max-P ($\alpha = 0.4$)‡	**0.7856**	**0.6748**	**0.5545**	**0.6566**	**0.8561**	**0.7486**
Avg-Max-P ($\alpha = 0.8$)	0.7758	0.6661	0.5505	0.6498	0.8528	0.7435
Avg-Max-U ($\alpha = 0.1$)	**0.7883**	0.6632	0.5356	0.6411	0.8471	0.7314
Avg-Max-U ($\alpha = 0.2$)‡	0.7785	**0.6647**	**0.5452**	**0.6467**	**0.8493**	**0.7369**
Avg-Max-U ($\alpha = 0.4$)	0.7617	0.6513	0.5383	0.6341	0.8425	0.7262
Avg-Max-U ($\alpha = 0.8$)	0.7416	0.6377	0.5295	0.6204	0.8342	0.7141
OPT ($\alpha = 0.1$)	0.7714	**0.6632**	0.5369	0.6400	0.8433	0.7272
OPT ($\alpha = 0.2$)‡	**0.7725**	0.6605	**0.5448**	**0.6434**	**0.8470**	**0.7340**
OPT ($\alpha = 0.4$)	0.7649	0.6495	0.5387	0.6334	0.8420	0.7256
OPT ($\alpha = 0.8$)	0.7541	0.6412	0.5315	0.6261	0.8377	0.7188
LSAP ($\alpha = 0.1$)	0.7720	0.6650	0.5392	0.6403	0.8446	0.7294
LSAP ($\alpha = 0.2$)	**0.7823**	0.6667	0.5456	0.6515	0.8512	0.7400
LSAP ($\alpha = 0.4$)‡	0.7780	**0.6670**	**0.5495**	**0.6522**	**0.8529**	**0.7434**
LSAP ($\alpha = 0.8$)	0.7709	0.6612	0.5468	0.6487	0.8506	0.7401
Soft-Max ($\alpha = 0.0$)	0.7394	0.5921	0.4818	0.5661	0.8034	0.6581
Soft-Max ($\alpha = 0.1$)‡	**0.7676**	**0.6554**	**0.5341**	**0.6350**	**0.8421**	**0.7247**
Soft-Max ($\alpha = 0.2$)	0.7421	0.6279	0.5148	0.6032	0.8256	0.6977
Sharp-Max ($\alpha = 0.2, \gamma = 1/6$)	0.7758	0.6683	0.5407	0.6489	0.8490	0.7361
Sharp-Max ($\alpha = 0.4, \gamma = 1/6$)‡	**0.7942**	**0.6763**	**0.5517**	**0.6618**	**0.8577**	**0.7499**
Sharp-Max ($\alpha = 0.8, \gamma = 1/6$)	0.7725	0.6554	0.5398	0.6384	0.8448	0.7291
Sparse-Max ($\alpha = 0.2$)	0.7763	0.6735	0.5456	0.6559	0.8512	0.7402
Sparse-Max ($\alpha = 0.4$)‡	**0.7970**	**0.6839**	**0.5597**	**0.6695**	**0.8612**	**0.7562**
Sparse-Max ($\alpha = 0.8$)	0.7828	0.6690	0.5497	0.6592	0.8537	0.7450

Table 3: Hyper-parameter Selection for MEAN, Avg-Max-P, Avg-Max-U, OPT, LSAP, Soft-Max, Sparse-Max and Sharp-Max, with various α. Best models denoted by ‡ are reported in our paper.

Model	P@1	P@3	P@5	Rprec	NDCG	MAP
Sharp-Max ($\alpha = 0.2, \gamma = 1/4$)	0.7785	0.6627	0.5410	0.6438	0.8487	0.7351
Sharp-Max ($\alpha = 0.2, \gamma = 1/5$)	0.7736	0.6636	0.5408	0.6460	0.8482	0.7345
Sharp-Max ($\alpha = 0.2, \gamma = 1/6$)	0.7758	0.6683	0.5407	0.6489	0.8490	0.7361
Sharp-Max ($\alpha = 0.2, \gamma = 1/7$)	0.7731	0.6672	0.5406	0.6482	0.8474	0.7338
Sharp-Max ($\alpha = 0.2, \gamma = 1/8$)	0.7687	0.6678	0.5379	0.6469	0.8460	0.7318
Sharp-Max ($\alpha = 0.4, \gamma = 1/4$)	0.7839	0.6672	0.5450	0.6518	0.8521	0.7409
Sharp-Max ($\alpha = 0.4, \gamma = 1/5$)	0.7883	0.6755	0.5504	0.6623	0.8563	0.7477
Sharp-Max ($\alpha = 0.4, \gamma = 1/6$)	**0.7942**	0.6763	**0.5517**	0.6618	**0.8577**	**0.7499**
Sharp-Max ($\alpha = 0.4, \gamma = 1/7$)	0.7926	0.6786	0.5510	0.6614	0.8569	0.7487
Sharp-Max ($\alpha = 0.4, \gamma = 1/8$)	0.7872	**0.6793**	0.5515	**0.6644**	0.8561	0.7482
Sharp-Max ($\alpha = 0.8, \gamma = 1/4$)	0.7470	0.6431	0.5263	0.6190	0.8327	0.7106
Sharp-Max ($\alpha = 0.8, \gamma = 1/5$)	0.7666	0.6498	0.5331	0.6282	0.8407	0.7224
Sharp-Max ($\alpha = 0.8, \gamma = 1/6$)	0.7725	0.6554	0.5398	0.6384	0.8448	0.7291
Sharp-Max ($\alpha = 0.8, \gamma = 1/7$)	0.7812	0.6596	0.5416	0.6473	0.8481	0.7338
Sharp-Max ($\alpha = 0.8, \gamma = 1/8$)	0.7818	0.6661	0.5423	0.6497	0.8489	0.7355

Table 4: Hyper-parameter Selection for Sharp-Max, with $\alpha \in \{0.2, 0.4, 0.8\}$ and $\gamma \in \{4, 5, 6, 7, 8\}$.

Model	Dialogue60		DialoguesTest	
	Hits@5	Hits@10	Hits@5	Hits@10
MEAN ($\alpha = 0.1$)	46.70	57.71	48.25	57.47
Avg-Max-P ($\alpha = 0.4$)	**56.39**	65.64	56.39	65.09
Avg-Max-U ($\alpha = 0.2$)	55.51	66.08	55.87	64.75
OPT ($\alpha = 0.2$)	52.86	61.67	54.95	63.55
LSAP ($\alpha = 0.4$)	53.74	63.44	55.33	64.03
Sharp-Max ($\alpha = 0.4, \gamma = 1/6$)	51.54	59.91	56.39	65.26
Sparse-Max ($\alpha = 0.4$)	55.07	**66.52**	**59.91**	**68.24**

Table 5: Comparison of our alignment models with baselines on human-human conversations using Hits@5/10. Dialogue60 is the subset used in our paper. DialoguesTest contains all dialogues in test set of PERSONA-LEAKAGE. Sparse-Max results on Dialogue60 is reported in our paper.