Structured Output Learning with Abstention: Application to Accurate Opinion Prediction

Alexandre Garcia [1] Slim Essid [1] Chloé Clavel [1] Florence d'Alché-Buc [1]

Abstract

Motivated by Supervised Opinion Analysis, we propose a novel framework devoted to Structured Output Learning with Abstention (SOLA). The structure prediction model is able to abstain from predicting some labels in the structured output at a cost chosen by the user in a flexible way. For that purpose, we decompose the problem into the learning of a pair of predictors, one devoted to structured abstention and the other, to structured output prediction. To compare fully labeled training data with predictions potentially containing abstentions, we define a wide class of asymmetric abstention-aware losses. Learning is achieved by surrogate regression in an appropriate feature space while prediction with abstention is performed by solving a new pre-image problem. Thus, SOLA extends recent ideas about Structured Output Prediction via surrogate problems and calibration theory and enjoys statistical guarantees on the resulting excess risk. Instantiated on a hierarchical abstention-aware loss, SOLA is shown to be relevant for fine-grained opinion mining and gives state-of-the-art results on this task. Moreover, the abstention-aware representations can be used to competitively predict user-review ratings based on a sentence-level opinion predictor.

1. Introduction

Up until recent years, opinion analysis in reviews has been commonly handled as a supervised polarity (positive vs. negative) classification problem. However, understanding the grounds on which an opinion is formed is of highest interest for decision makers. Aligned with this goal, the emerging field of aspect-based sentiment analysis (Pontiki et al., 2016) has evolved towards a more involved machine learning task where opinions are considered to be structured objects—typically hierarchical structures linking polarities to aspects and relying on different units of analysis (*i.e.* sentence-level and review-level) as in (Marcheggiani et al., 2014). While this problem has attracted a growing attention from the structured output prediction community, it has also raised an unprecedented challenge: the human interpretation of opinions expressed in the reviews is subjective and the opinion aspects and their related polarities are sometimes expressed in an ambiguous way and difficult to annotate (Clavel & Callejas, 2016; Marcheggiani et al., 2014). In this context, the prediction error should be flexible and should integrate this subjectivity so that, for example, mistakes on one aspect do not interfere with the prediction of polarity.

In order to address this issue, we propose a novel framework called Structured Output Learning with Abstention (SOLA) which allows for abstaining from predicting parts of the structure, so as to avoid providing erroneous insights about the object to be predicted, therefore increasing reliability. The new approach extends the principles of learning with abstention recently introduced for binary classification (Cortes et al., 2016) and generalizes surrogate least-square loss approaches to Structured Output Prediction recently studied in (Brouard et al., 2016; Ciliberto et al., 2016; Osokin et al., 2017). The main novelty comes from the introduction of an asymmetric loss, based on embeddings of desired outputs and outputs predicted with abstention in the same space. Interestingly, similarly to the case of Output Kernel Regression (Brouard et al., 2016) and appropriate inner product-based losses (Ciliberto et al., 2016), the approach relies on a simple surrogate formulation, namely a least-squares formulation followed by the resolution of a new pre-image problem. The paper is organized as follows. Section 2 introduces the problem to solve and the novel framework, SOLA. Section 3 provides statistical guarantees about the excess risk in the framework of Least Squares Surrogate Loss while section 4 is devoted to the pre-image developed for hierarchical output structures. Section 5 presents the numerical experiments and Section 6 draws a conclusion.

[1] LTCI, Telecom ParisTech, Paris, France. Correspondence to: Alexandre Garcia <algarcia@enst.fr>.

Proceedings of the 35^{th} International Conference on Machine Learning, Stockholm, Sweden, PMLR 80, 2018. Copyright 2018 by the author(s).

2. Structured Output Labeling with Abstention

Let $\mathcal{X}$ be the input sample space. We assume a target graph structure of interest, $\mathcal{G} = (\mathcal{V} = \{\nu_1, \dots, \nu_d\}, \mathcal{E} : \mathcal{V} \times \mathcal{V} \to \{0, 1\})$ where $\mathcal{V}$ is the set of vertices and $\mathcal{E}$ is the edge relationship between vertices. A legal *labeling* or *assignment* of $\mathcal{G}$ is a d-dimensional binary vector, $y \in \{0, 1\}^d$, that also satisfies some properties induced by the graph structure, *i.e.* by $\mathcal{E}$. We call $\mathcal{Y}$ the subset of $\{0, 1\}^d$ that contains all possible legal labelings of $\mathcal{G}$. Given $\mathcal{G}$, the goal of Structured Output Labeling is to learn a function $f : \mathcal{X} \to \mathcal{Y}$ that predicts a legal labeling $\hat{y}$ given some input x. Let us emphasize that x does not necessarily share the same structure $\mathcal{G}$ with the outputs objects. For instance, in Supervised Opinion Analysis, the inputs are reviews in natural language described by a sequence of feature vectors, each of them representing a sentence. Extending Supervised Classification with Abstention (Cortes et al., 2016), Structured Output Learning with Abstention aims at learning a pair of functions (h, r) from $\mathcal{X}$ to $Y^{H,R} \subset \{0, 1\}^d \times \{0, 1\}^d$ composed of a predictor h that predicts the label of each component of the structure and an abstention function r that determines on which components of the structure $\mathcal{G}$ to abstain from predicting a label. If we note $\mathcal{Y}^\star \subset \{0, 1, a\}^d$, the set of legal labelings with abstention where a denotes the abstention label, then the abstention-aware predictive model $f^{h,r} : \mathcal{X} \to \mathcal{Y}^\star$ is defined from h and r as follows:

$$\begin{aligned} f^{h,r}(x)^T &= [f_1^{h,r}(x), \dots, f_d^{h,r}(x)], \\ f_i^{h,r}(x) &= 1_{h(x)_i=1} 1_{r(x)_i=1} + a 1_{r(x)_i=0}. \end{aligned} \tag{1}$$

Now, assuming we have a random variable (X, Y) taking its values in $\mathcal{X} \times \mathcal{Y}$ and distributed according to a probability distribution $\mathcal{D}$. Learning the predictive model raises the issue of designing an appropriate abstention-aware loss function to define a learning problem as a risk minimization task. Given the relationship in Eq. (1), a risk on $f^{h,r}$ can be converted into a risk on the pair (h, r) using an abstention-aware loss $\Delta_a : \mathcal{Y}^{H,R} \times \mathcal{Y} \to \mathbb{R}^+$:

$$R(h, r) = \mathbb{E}_{x,y \sim \mathcal{D}}\, \Delta_a(h(x), r(x), y). \tag{2}$$

In this paper, we propose a family of abstention-aware losses that both generalizes the abstention-aware loss in the binary classification case (see (Cortes et al., 2016)) and extends the scope of hierarchical losses previously proposed by (Cesa-Bianchi et al., 2006) for Hierarchical Output Labeling tasks. An abstention-aware loss is required to deal asymmetrically with observed labels which are supposed to be complete and predicted labels which may be incomplete due to partial abstention. We thus propose the following general form for the Δ_a function:

$$\Delta_a(h(x), r(x), y) = \langle \psi_{wa}(y), C\psi_a(h(x), r(x)) \rangle, \tag{3}$$

relying on a bounded linear operator (a rectangular matrix) $C : \mathbb{R}^p \to \mathbb{R}^q$ and two bounded feature maps: $\psi_a : \mathcal{Y}^{H,R} \to \mathbb{R}^p$ devoted to outputs with abstention and $\psi_{wa} : \mathcal{Y} \to \mathbb{R}^q$, devoted to outputs without abstention. The three ingredients of the loss Δ_a must enable the loss to be non negative. This is the case for the following examples.

In **Binary classification with abstention**, we have $\mathcal{Y} = \{0, 1\}$ and the abstention-aware loss Δ_a^{bin} is defined by :

$$\Delta_a^{bin}(h(x), r(x), y) = \begin{cases} 1 \text{ if } y \neq h(x) \text{ and } r(x) = 1 \\ 0 \text{ if } y = h(x) \text{ and } r(x) = 1 \\ c \text{ if } r(x) = 0 \end{cases},$$

where $c \in [0, 0.5]$ is the rejection cost; with $r(x) = 0$, in case of abstention and 1, otherwise. This can be written with the corresponding functions ψ_{wa} and ψ_a defined as:

$$\psi_{wa}(y) = \begin{pmatrix} y \\ 1 - y \end{pmatrix},\ C = \begin{pmatrix} 0 & 1 & c \\ 1 & 0 & c \end{pmatrix},$$

$$\psi_a(h(x), r(x)) = \begin{pmatrix} h(x)r(x) \\ (1 - h(x))r(x) \\ 1 - r(x) \end{pmatrix}.$$

H-loss (hierarchical loss): now we assume that the target structure $\mathcal{G}$ is a hierarchical binary tree. Then, $\mathcal{E}$ is now the set of directed edges, reflecting a *parent* relationship among nodes (each node except the root has one parent). Regarding the labeling, we impose the following property : if an oriented pair $(\nu_i, \nu_j) \in \mathcal{E}$, then $y_i \geq y_j$, meaning that a child node cannot be greater that his parent node. The H-loss (Cesa-Bianchi et al., 2006) which measures the length of the common path from the root to the leaves between these assignments is defined as follows:

$$\Delta_H(h(x), y) = \sum_{i=1}^{d} c_i 1_{h(x)_i \neq y_i} 1_{h(x)_{p(i)} = y_{p(i)}},$$

where $p(i)$ is the index of the parent of i according to the set of edges $\mathcal{E}$, and c_i is a set of positive constants non-increasing on paths from the root to the leaves.

Such a loss can be rewritten under the form: $\Delta_H(h(x), y) = \langle \psi_{wa}(y), C\psi_{wa}(h(x)) \rangle$

$$\psi_{wa}(z) = \begin{pmatrix} z \\ Gz \end{pmatrix},\ C = \begin{pmatrix} -2diag(c) & diag(c) \\ diag(c) & 0 \end{pmatrix},$$

G is the adjacency matrix of the underlying binary tree structure and c the vector of weights defined above. The case of the Hamming loss can also be recovered by choosing:

$$\psi_{wa}(y) = \begin{pmatrix} y \\ 1 - y \end{pmatrix},\ \psi_a(h(x), r(x)) = \begin{pmatrix} 1 - h(x) \\ h(x) \end{pmatrix},$$

$$C = I_{2d},$$

where I_{2d} is the $2d$ identity matrix.

Abstention-aware H-loss (Ha-loss): By mixing the H-loss and the abstention-aware binary classification loss, we get the novel Ha-loss which we define as follows:

$$\Delta_{Ha}(h(x), r(x), y) = \sum_{i=1}^{d} \underbrace{c_{Ai} 1_{\{f_i^{h,r}=a, f_{p(i)}^{h,r}=y_{p(i)}\}}}_{\text{abstention cost}} \tag{4}$$

$$+ \underbrace{c_{A_c i} 1_{\{f_i^{h,r} \neq y_i, f_{p(i)}^{h,r}=a\}}}_{\text{abstention regret}} + \underbrace{c_i 1_{\{f_i^{h,r} \neq y_i, f_{p(i)}^{h,r}=y_{p(i)}, a \neq f_i^{h,r}\}}}_{\text{misclassification cost}},$$

where c_{Ai} and $c_{A_c i}$ can be chosen as constants or be function of the predictions. Thus, we have designed this loss so it is adapted to hierarchies where some nodes are known to be hard to predict whereas their children are easy to predict. In this case, the abstention choice can be used at a particular node to pay the *cost* c_A for predicting its child. If this prediction is still a mistake, the price $c_{A_c i}$ is additionally paid and acts as a *regret* cost penalizing the unnecessary abstention chosen at the parent. Acting on c_A and c_{Ac} provides a way to control the number of abstentions not only through the risk taken by predicting a given node but also its children. For sake of space, the dot product representation with ψ_{wa} and ψ_a of this loss is detailed in the supplementary material.

2.1. Empirical risk minimization for SOLA

The goal of SOLA is to learn a pair (h, r) from a i.i.d. (training) sample drawn from a probability distribution $\mathcal{D}$ that minimizes the true risk:

$$\begin{aligned} \mathcal{R}(h, r) &= \mathbb{E}_{x,y \sim \mathcal{D}} \; \Delta_a(h(x), r(x), y), \\ &= \mathbb{E}_{x,y \sim \mathcal{D}} \; \langle \psi_{wa}(y), C\psi_a(h(x), r(x)) \rangle. \end{aligned}$$

We notice that this risk can be rewritten as an expected valued over the input variables only:

$$\mathcal{R}(h, r) = \mathbb{E}_x \; \langle \mathbb{E}_{y|x} \psi_{wa}(y), C\psi_a(h(x), r(x)) \rangle.$$

This pleads for considering the following surrogate problem:

- Step 1: we define $g^*(x) = \mathbb{E}_{y|x}\psi_{wa}(y) = \min_{g \in (\mathcal{X} \to \mathbb{R}^q)} \underbrace{\mathbb{E}_{x,y} \|\psi_{wa}(y) - g(x)\|^2}_{\text{surrogate risk}}$. g^* is then the minimizer of a square surrogate risk.
- Step 2: we solve the following pre-image or decoding problem:

$$(\hat{h}(x), \hat{r}(x)) = \underset{(y_h, y_r) \in \mathcal{Y}^{H,R}}{\arg\min} \; \langle g^*(x), C\psi_a(y_h, y_r) \rangle,$$

Solving directly the problem above raises some difficulties:

- In practice, as usual, we do not know the expected value of $\psi_{wa}(y)$ conditioned on x: $\mathbb{E}_{y|x}\psi_{wa}(y)$ needs to be estimated from the training sample $\{(x_i, y_i), i = 1, \dots, n\}$. This simple regression problem is referred to as the learning step and will be solved in the next subsection.
- The complexity of the $\arg\min$ problem will depend on some properties of ψ_a. We will refer to this problem as the pre-image and show how to solve it practically at a later stage.

These pitfalls, common to all structured output learning problems, can be overcome by substituting a surrogate loss to the target loss and proceeding in two steps:

1. Solve the surrogate penalized empirical problem (learning phase):

$$\min_g \frac{1}{n} \sum_{i=1}^{n} \|\psi_{wa}(y_i) - g(x_i)\|^2 + \lambda \Omega(g), \tag{5}$$

   where Ω is a penalty function and λ a positive parameter. Thus, get a minimizer $\hat{g}$ which is an estimate of $\mathbb{E}_{y|x}\psi_{wa}(y)$.

2. Solve the pre-image or *decoding* problem:

$$(\hat{h}(x), \hat{r}(x)) = \underset{(h(x), r(x)) \in \mathcal{Y}^{H,R}}{\arg\min} \; \langle \hat{g}(x), C\psi_a(h(x), r(x)) \rangle. \tag{6}$$

2.2. Estimation of the conditional density $\mathbb{E}_{y|x}\psi_{wa}(y)$ from training data

We choose to solve this problem in $\mathcal{H} \subset \mathcal{F}(\mathcal{X}, \mathbb{R}^q)$, a vector-valued Reproducing Kernel Hilbert Space associated to an operator-valued kernel $K : \mathcal{X} \times \mathcal{X} \to \mathcal{L}(\mathbb{R}^q)$. For the sake of simplicity, K is chosen as a decomposable operator-valued kernel with identity: $K(x, x') = Ik(x, x')$ where k is a positive definite kernel on $\mathcal{X}$ and I is the $q \times q$ identity matrix. The penalty is chosen as $\Omega(g) = \|g\|^2_{\mathcal{H}}$. This choice leads to the ridge regression problem:

$$\underset{g \in \mathcal{H}}{\arg\min} \sum_{i=1}^{n} \|g(x_i) - \psi_{wa}(y_i)\|^2 + \lambda \|g\|^2_{\mathcal{H}}, \tag{7}$$

that admits a unique and well known closed-form solution (Micchelli & Pontil, 2005; Brouard et al., 2016).

As $\hat{g}(x)$ is only needed at the prediction stage, within the pre-image to solve, it is important to emphasize the dependency of $\hat{g}(x)$ on the feature vectors $\psi_{wa}(y_i)$:

$$\hat{g}(x) = \sum_{i=1}^{n} \alpha_i(x) \psi_{wa}(y_i), \tag{8}$$

where $\alpha(x)$ is the following vector:

$$\alpha(x) = K_x(\mathbf{K} + \lambda I_{qn})^{-1}, \tag{9}$$

where $K_x = [K(x, x_1), \dots, K(x, x_n)]$. $\mathbf{K}$ is the $qn \times qn$ block matrix such that $\mathbf{K}_{i,j} = K(x_i, x_j)$ and I_{qn} is the identity matrix of the same size. $\alpha_i(x)$ is the block i of $\alpha(x)$.

3. Learning guarantee for structured losses with abstention

In this section, we give some statistical guarantees when learning predictors in the framework previously described. To this end, we build on recent results in the framework of Least Squares Loss Surrogate (Ciliberto et al., 2016) that are extended to abstention-aware prediction.

Theorem 1. *Given the definition of Δ_a in* (3)*, let us denote* (h, r)*, the pair of predictor and reject functions associated to the estimate $\hat{g}$ obtained by solving the learning problem stated in Eq.* (7)*:*

$$(h(x), r(x)) = \underset{(y_h, y_r) \in \mathcal{Y}^{H,R}}{\arg\min} \langle C\psi_a(y_h, y_r), \hat{g}(x)\rangle.$$

Its true risk with respect to Δ_a writes as:

$$\mathcal{R}(h, r) = \mathbb{E}_x \langle C\psi_a(h(x), r(x)), \mathbb{E}_{y|x}\psi_{wa}(y)\rangle.$$

The optimal predictor (h^*, r^*) is defined as:

$$(h^*(x), r^*(x)) = \underset{(y_h, y_r) \in \mathcal{Y}^{H,R}}{\arg\min} \langle C\psi_a(y_h, y_r), \mathbb{E}_{y|x}\psi_{wa}(y)\rangle.$$

The excess risk of an abstention aware predictor (h, r): $\mathcal{R}(h, r) - \mathcal{R}(h^\star, r^\star)$ is linked to the estimation error of the conditional density $\mathbb{E}_{y|x}\psi_{wa}(y)$ by the following inequality:

$$\mathcal{R}(h, r) - \mathcal{R}(h^\star, r^\star) \leq 2c_l\sqrt{\mathcal{L}(\hat{g}) - \mathcal{L}(\mathbb{E}_{y|x}\psi_{wa}(y))}, \tag{10}$$

where $\mathcal{L}(g) = \mathbb{E}_{x,y}\|\psi_{wa}(y) - g(x)\|^2$, and $c_l = \|C\| \max_{y_h, y_r \in \mathcal{Y}^{H,R}} \|\psi_a(y_h, y_r)\|_{\mathbb{R}^p}$.

The full proof is given in the Supplements. Close to the one in (Ciliberto et al., 2016), it is extended by taking the sup of the norm of ψ_a over $\mathcal{Y}^{H,R}$. Moreover when the problem (7) is solved by Kernel Ridge Regression, (Ciliberto et al., 2016) have shown the universal consistency and have obtained a generalization bound that still holds in our case since it relies on the result of Theorem 1 only. As a consequence the excess risk of predictors built in the SOLA framework is controlled by the risk suffered at the learning step for which we use off the shelf vector valued regressors with their own convergence guarantees.

In the following, we specifically study the pre-image problem in the SOLA framework for a class of output structures that we detail hereafter.

4. Pre-image for hierarchical structures with Abstention

In what follows we focus on a class of structured outputs that can be viewed as hierarchical objects for which we show how to solve the pre-image problems involved for a large class of losses.

4.1. Hierarchical output structures

Definition 1. *A HEX graph $G = (V, E_h, E_e)$ is a graph consisting of a set of nodes V = $\{v_1, \dots, v_n\}$, directed edges $E_h \subset V \times V$, and undirected edges $E_e \subset V \times V$, such that the subgraph $G_h = (V, E_h)$ is a directed acyclic graph (DAG) and the subgraph $G_e = (V, E_e)$ has no self loop.*

Definition 2. *An assignment (state) $y \in \{0,1\}^d$ of labels V in a HEX graph $G = (V, E_h, E_e)$ is legal if for any pair of nodes labeled $(y_{(i)}, y_{(j)}) = (1, 1)$, $(v_i, v_j) \notin E_e$ and for any pair $(y_{(i)}, y_{(j)}) = (0, 1)$, $(v_i, v_j) \notin E_h$.*

Definition 3. *The state space $SG \subseteq \{0,1\}^d$ of graph G is the set of all legal assignments of G.*

Thus a HEX graph can be described by a pair of (1) a directed graph over a set of binary nodes indicating that any child can be labeled 1 only if its parent is also labeled 1 and (2) an undirected graph of exclusions such that two nodes linked by an edge cannot be simultaneously labeled 1. Note that HEX graphs can represent any type of binary labeled graph since E_h and E_e can be empty sets. In previous works, they have been used to model some coarse to fine ontology through the hierarchy G_h while incorporating some prior known labels exclusions encoded by G_e (Deng et al., 2014; BenTaieb & Hamarneh, 2016)

While the output data we consider consists of HEX graph assignments , our predictions with abstention $(h(x), r(x))$ belong to another space $\mathcal{Y}^{H,R} \subseteq \{0,1\}^d \times \{0,1\}^d$ for which we do not restrict $h(x)$ to belong to $\mathcal{Y}$ but rather allow for other choices detailed in the next section.

4.2. Efficient solution for the preimage problem

The complexity of the preimage problem is due to two aspects: i) the space in which we search the solution ($\mathcal{Y}^{H,R}$) can be hard to explore; and ii) the ψ_a function can lead to high dimensional representations for which the minimization problem is harder.

The pre-image problem involves a minimization over a constrained set of binary variables. For a large class of abstention-aware predictors we propose a branch-and-bound formulation for which a nearly optimal initialization point can be obtained in a polynomial time. Following the line given by the form of our abstention aware predictor $f^{h,r}$ defined in Section 2, we consider losses involving binary interaction between the predict function $h(x)$ and the reject function $r(x)$, and suppose that there exists a rectangular matrix M such that $\psi_a(h(x), r(x)) = M \begin{pmatrix} h(x) \\ r(x) \\ h(x) \otimes r(x) \end{pmatrix}$ where $\otimes$ is the Kronecker product between vectors. Such a class takes as special cases the examples presented in Section 2. We state the following linearization theorem under binary interaction hypothesis:

Theorem 2. *Let l_{ha} be an abstention-aware loss defined by its output mappings ψ_{wa}, ψ_a and the corresponding cost matrix C.*

If the ψ_a mapping is a linear function of the binary interactions of $h(x)$ and $r(x)$ i.e. there exists a matrix M such that $\forall (h(x), r(x)) \in \mathcal{Y}^{H,R}$ $\psi_a(h(x), r(x)) = M \begin{pmatrix} h(x) \\ r(x) \\ h(x) \otimes r(x) \end{pmatrix}$, then there exists a bounded linear operator A and a vector b such that $\forall \psi_x \in \mathbb{R}^p$ the pre-image problem:

$$(\hat{h}(x), \hat{r}(x)) = \underset{(y_h, y_r) \in \mathcal{Y}^{H,R}}{\arg\min} \langle \psi_a(y_h, y_r), \psi_x \rangle,$$

has the same solutions as the linear program:

$$\hat{h}(x), \hat{r}(x) = \underset{(y_h, y_r) \in \mathcal{Y}^{H,R}}{\arg\min} [y_h^T y_r^T c^T] M^T \psi_x$$
$$\textit{s.t. } A \begin{pmatrix} y_h \\ y_r \\ c \end{pmatrix} \leq b.$$

Where c is a d^2 dimensional vector constrained to be equal to $y_h \otimes y_r$.

The proof is detailed in the supplementary material.

The problem above still involves a minimization over the structured binary set $\mathcal{Y}^{H,R}$. Such a set of solutions encodes some predefined constraints:

- Since the objects we intend to predict are HEX graph assignments, the vectors of the output space $y \in \mathcal{Y}$ should satisfy the hierachical constraint : $y_i \leq y_{p(i)}$ with $p(i)$ the index of the parent of i according to the hierarchy. When predicting with abstention we relax this condition since we suppose that a descendant node can take the value $y_i = 1$ if its parent was active $y_{p(i)} = 1$ or if we abstained from predicting it $r_{p(i)} = 0$. Such a condition is equivalent to the constraint
$$y_i r_{p(i)} \leq y_{p(i)} r_{p(i)}. \tag{11}$$
- A second condition we used in practice is the restriction of the use of abstention for two consecutive nodes: structured abstention at a layer must be used in order to reveal a subsequent prediction which is known to be easy. Such a condition can be encoded through the inequality:
$$r_i + r_{p(i)} \leq 1. \tag{12}$$

In our experiments, the structured space $\mathcal{Y}^{H,R}$ has been chosen as the set of binary vectors $(h(x), r(x)) \in \mathcal{Y}^{H,R}$ that respect the two above conditions. These choices are motivated by our application but note that any subset of $\{0,1\}^d \times \{0,1\}^d$ can be built in a similar way by adding some inequality constraints: $A_{\mathcal{Y}^{H,R}} \begin{pmatrix} h(x) \\ r(x) \\ h(x) \otimes r(x) \end{pmatrix} \leq b_{\mathcal{Y}^{H,R}}$. Consequently, the $\mathcal{Y}^{H,R}$ constraints can be added to the previous minimization problem to build the canonical form:

$$(\hat{h}(x), \hat{r}(x)) = \underset{(y_h, y_r)}{\arg\min} [y_h^T y_r^T c^T] M^T \psi_x$$
$$\text{s.t. } A_{\text{canonical}} \begin{pmatrix} y_h \\ y_r \\ c \end{pmatrix} \leq b_{\text{canonical}},$$
$$(y_h, y_r) \in \{0,1\}^d \times \{0,1\}^d,$$

where $A_{\text{canonical}} = \begin{pmatrix} A \\ A_{\mathcal{Y}^{H,R}} \end{pmatrix}$ and $b_{\text{canonical}} = \begin{pmatrix} b \\ b_{\mathcal{Y}^{H,R}} \end{pmatrix}$.

The complexity of the problem above is linked to some properties of the $A_{\text{canonical}}$ operator. (Goh & Jaillet, 2016) have shown that in the case of the minimization of the H-loss with hierarchical constraints, the linear operator $A_{\text{canonical}}$ satisfies the property of total unimodularity (Schrijver, 1998) which is a sufficient condition for the problem above to have the same solutions as its continuous relaxation leading to a polynomial time algorithm. In the more general case of the Ha-loss, solving such an integer program is NP-hard and the optimal solution can be obtained using a branch-and-bound algorithm. When implementing this type of approach, the choice of the initialization point can strongly influence the convergence time. As in practical applications, we expect the number of abstentions to remain low, such a point can be chosen as the solution of the original prediction problem without abstention (Goh & Jaillet, 2016). Moreover since the abstention mechanism should modify only a small subset of the predictions, we expect this solution to be close to the abstention aware one.

5. Numerical Experiments

We study three subtasks of opinion mining, namely sentence-based aspect prediction, sentence-based joint prediction of aspects and polarities (possibly with abstention) and full review-based star rating. We show that these tasks can be linked using a hierarchical graph similar to the probabilistic model of (Marcheggiani et al., 2014) and exploit the abstention mechanism to build a robust pipeline: based on the opinion labels available at the sentence-level, we build a two-stage predictor that first predicts the aspects and polarities at the sentence level, before deducing the corresponding review-level values.

5.1. Parameterization of the Ha-loss

In all our experiments, we rely on the expression of the Ha-loss presented in 4. The linear programming formulation of the pre-image problem used in the branch-and-bound solver is derived in the supplementary material and involves a decomposition similar to the one described in Section 2 for the H-loss. Implementing the Ha-loss requires choosing the weights c_i, c_{Ai} and c_{A_ci}. We first fix the c_i weights in the following way :

$$
\begin{aligned}
c_0 &= 1 \\
c_i &= \frac{c_{p(i)}}{|\text{siblings(i)}|} \; \forall i \in \{1, \dots, d\}.
\end{aligned}
$$

Here, 0 is assumed to be the index of the root node. This weighting scheme has been commonly used in previous studies (Rousu et al., 2006; Bi & Kwok, 2012) and is related to the minimization of the Hamming Loss on a vectorized representation of the graph assignment. As far as the abstention weights c_{Ai} and c_{A_ci} are concerned, making an exhaustive analysis of all the possible choices is impossible due to the number of parameters involved. Therefore, our experiments focus on weighting schemes built in the following way:

$$
\begin{aligned}
c_{Ai} &= K_A c_i \\
c_{A_ci} &= K_{A_c} c_i
\end{aligned}
$$

The effect of the choices of K_A and K_{A_c} will be illustrated below on the opinion prediction task. We also ran a set of experiments on a hierarchical classification task of MRI images from the IMAGECLEF2007 dataset reusing the setting of (Dimitrovski et al., 2008) where we show the results obtained for different c_i weighting schemes. The settings and the results have been placed in the supplementary material.

5.2. Learning with Abstention for aspect-based opinion mining

We test our model on the problem of aspect-based opinion mining on a subset of the TripAdvisor dataset released in (Marcheggiani et al., 2014). It consists of 369 hotel reviews for a total of 4856 sentences with predefined train and test sets. In addition to the review-level star ratings, the authors gathered the opinion annotations at the sentence-level for a set of 11 predefined aspects and their corresponding polarity. Similarly to them, we discard the “NOT RELATED” aspect and consider the remaining 10 aspects with the 3 different polarities (positive, negative or neutral) for each. We propose a graphical representation of the opinion structure at the sentence level (see Fig. 1). Objects in the output space $y \in \mathcal{Y}$ consist of trees of depth 3 where the first node is the root, the second layer is made of aspect labels and the third one is the polarities corresponding to each aspect. The corresponding assignments are encoded by a binary matrix $y \in \mathcal{Y}$ where y is the concatenation of the vectors indicating the presence of each aspect (depth 2) and the ones indicating the polarity.

An example of y encoding is displayed in Fig.1. Based on the recent results of (Conneau et al., 2017), we focus on the InferSent representation to encode our inputs. This dense sentence embedding corresponds to the inner representation of a deep neural network trained on a natural language inference task and has been shown to give competitive results in other natural language processing tasks.

We test our model on 3 different subtasks. In **Exp1**, we first apply our model (H Regression InferSent) to the task of opinion aspect prediction and compare it against two baselines and the original results of (Marcheggiani et al., 2014). In **Exp2**, we test our method and baselines on the problem of joint aspect and polarity prediction in order to assess the ability of the hierarchical predictor to take advantage of the output structure. On this task we additionally illustrate the behavior of abstention when varying the constants K_A and K_{A_c}. In **Exp3**, we illustrate the use abstention as a mean to build a robust pipeline on the task of star rating regression based on a sentence-level opinion predictor.

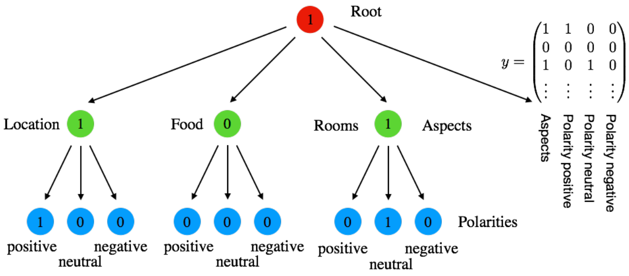

Figure 1. Graphical representation of the opinion structure

Exp1. Aspect prediction. In this first task, we aim at predicting the different aspects discussed in each sentence. This problem can be cast as a multilabel classification problem where the target is the first column of the output objects y for which we devise two baselines. The first relies on a logistic regression model (Logistic Regression InferSent) trained separately for each aspect. The second baseline (Linear chain Conditional Random Fields (CRF) (Sutton et al., 2012) InferSent) is inspired by the work of (Marcheggiani et al., 2014) who built a hierarchical CRF model based on a handcrafted sparse feature set including one-hot word encoding, POS tags and sentiment vocabulary. Since the optimization via Gibbs sampling of their model relies on the sparsity of the feature set, we could not directly use it with our dense representation. Linear chain CRF InferSent takes advantage of our input features while remaining computationally tractable. One linear chain is trained for each node of the output structures and the chain encodes the dependency between successive sentences.

Table 5.2 below shows the results in terms of micro-averaged F1 (μ-F1) score obtained on the task of aspect prediction. The three methods using InferSent give significantly better results than (Marcheggiani et al., 2014). Consequently, the next experiments will not consider them. Even though H Regression was trained in order to predict the whole structure, it obtains results similar to logistic regression and linear chain CRF.

method	μ-F1
H Regression InferSent	0.59
Logistic Regression InferSent	0.60
Linear chain CRF InferSent	0.59
Linear chain CRF sparse features Marcheggiani et al.	0.49
Hierarchical CRF sparse features Marcheggiani et al.	0.49

Table 1. Experimental results on the TripAdvisor dataset for the aspect prediction task.

Exp2. Joint polarity and aspect prediction with abstention. We take as output objects the assignments of the graph described (Fig. 1) and build an adapted abstention mechanism. Our intuition is that in some cases, the polarity might be easier to predict than the aspect to which it is linked. This can typically happen when some vocabulary linked to the current aspect has been unseen during the training or is implicit whereas the polarity vocabulary is correctly recognized. An example is the sentence " We had great views over the East River" where the aspect "Location" is implicit and where the "views" could mislead the predictor and result in a prediction of the aspect "Other". In such a case, (Marcheggiani et al., 2014) underline that the inter-annotator agreement is low. For this reason, we want that our classifier allows multiple candidates for aspect prediction while providing the polarity corresponding to them. We illustrate this behavior by running two sets of experiments in which we do not allow the predictor to abstain on the polarity.

In the first experiment, we want to analyze the influence of the parameterization of the Ha-loss. Following the parameterization of c_{Ai} and c_{A_ci} previously proposed, we generated some predictions with varying values of $K_A \in [0, 0.5]$ and $K_{A_c} \in \{0.25, 0.5, 0.75\}$. We displayed the Hamming loss between the true labels and the predictions as a function of the mean number of aspects on which the predictor abstained (Fig. 2) and handle two cases : modified : in the left figure, all nodes except the one on which we abstained were used to compute the Hamming loss. In the right one, all nodes except the aspect on which we abstained and their corresponding polarity were used to compute the Hamming loss. The H_{Strict} results correspond to a predictor for which the original hierarchical constraint is forced: $y_{(i)} \leq y_{p(i)}$ and the three other curves have been obtained with the generalized constraint hypothesis $y_{(i)} r_{p(i)} \leq y_{p(i)} r_{p(i)}$.

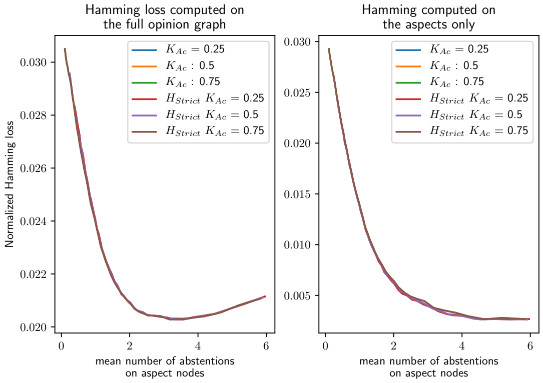

Figure 2. Hamming loss as a function of the number of aspect labels where the predictor abstained itself.

We additionally ran our model H Regression without abstention and our two baselines logistic regression for which we measured a similar Hamming loss of 0.03 (corresponding to 0 abstention on the left Figure 2). Concerning the micro-averaged F1 score, the H Regression retrieved a score of 0.54 being slightly above the logistic regression which scored 0.53 and the linear chain CRF with 0.52.

Two conclusions can be raised. Firstly, the value of K_{A_c} and the choice of the hypothesis H_{Strict} have little to no influence on the scores computed in the two cases previously described. Secondly, increasing the number of abstentions on aspects helps reducing the number of errors counted on the aspects nodes when the predictor abstains on less than 3 labels. After this point, the quality of the overall prediction decreases since the error rate on the remaining aspects selected for abstention is less than the one on the polarity labels

Subsequently, we examine the Hamming loss on the polarity predictions situated after an aspect node to understand the influence of the c_{A_c} coefficients and the relaxation of the H_{Strict} hypothesis in Figure 3. The orange curve gives the best score when the mean number of abstentions is between 2 and 4 per sentence. The only difference with the H_{strict} hypothesis is the ability to predict the polarity of an aspect candidate for abstention even if the predictor function does not select it. This behavior is made possible by the fact that

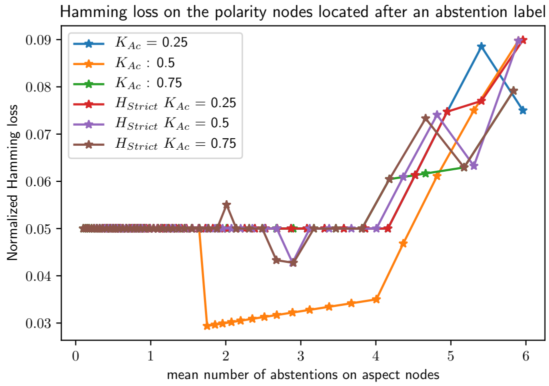

Figure 3. Hamming loss computed on polarity nodes located after an aspect for which the predictor abstained

our prediction does not respect the $\mathcal{Y}$ constraints but instead belong to the more flexible space $\mathcal{Y}^{H,R}$ Finally we show how abstention can be used to build a robust pipeline for star-rating regression.

Exp3. Star rating regression at the review level based on sentence level predictions. In the last round of experiments, we show that abstention can be used as a way to build a robust intermediate representation for the task of opinion rating regression (Wang et al., 2011) which consists in predicting the overall average star rating given by each reviewer on a subset of six predefined aspects. The figure below illustrates the different elements involved in our problem. The procedure is split in two steps. Firstly, we learn a sentence-level opinion predictor that takes advantage of the available annotations. This step corresponds to the one studied in the previous experiment. Then a vector-valued regressor (star regressor in Figure 4) is built. It takes as input the component-wise average of the sentence level opinion representations, and intends to predict the star ratings at the review level. For each of the five overall aspects a separate Ridge Regressor is trained based on the true labels available. Once learned, the regressors take as input the prediction of the first step in a pipelined way

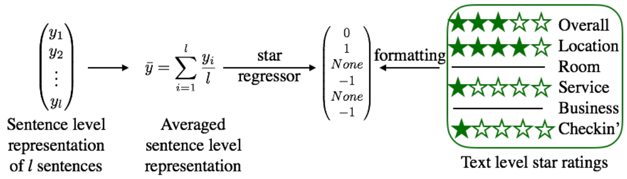
$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_l \end{pmatrix}$
Sentence level representation of l sentences
$\bar{y} = \sum_{i=1}^{l} \frac{y_i}{l}$
Averaged sentence level representation
star regressor
$\begin{pmatrix} 0 \\ 1 \\ None \\ -1 \\ None \\ -1 \end{pmatrix}$
formatting
★★★☆☆ Overall
★★★★☆ Location
Room
★☆☆☆☆ Service
Business
★☆☆☆☆ Checkin'
Text level star ratings

Figure 4. Star rating regression pipeline

Similarly to (Marcheggiani et al., 2014), we rescale the star ratings on a (-1,0,1) scale and report the macro-averaged mean average error on the test-set in Table 5.2 below under the column MAE text level. We additionally include the MAE error measured on polarity predictions at the sentence level counted when the underlying aspect predicted is a true positive. The first row is our oracle: the sentence-level opinion representations are assumed to be known on the test set and fed in the text-level opinion regressors to find back the star ratings. The Hierarchical CRF line corresponds to the best results reported by (Marcheggiani et al., 2014) on the two tasks. H Regression is our model without abstention used as a predictor of the sentence-level representation in the pipeline shown in Fig 4. Finally for the H Regression with abstention, we used as a sentence-level representation : $y_{\text{a}} = h(x) - (1 - r(x))$. Since the only non-zero components of $(1 - r(x))$ correspond to aspects on which we abstained, subtracting them from the original prediction results in a reduction of the confidence of the regressor for these aspects and biasing the corresponding polarity predictions towards 0. H Regression strongly outperforms Hierarchical CRF on both tasks. We do not report the score for H Regression with abstention since it is dependent on the number of abstentions but show that it improves the results of the H Regression model on the text-level prediction task. The significance of the scores has been assessed with a Wilcoxon rank sum test (p-value 10^{-}6).

method	MAE sentence level	MAE text level
Oracle: regression with true sentence labels	0	0.38
Hierarchical CRF	0.50	0.50
H Regression	0.30	0.45
H Regression with Abstention	C	0.43

Table 2. Experimental result on the TripAdvisor dataset for the polarity prediction task

6. Conclusion

The novel framework, Structured Learning with Abstention, extends two families of approaches: learning with abstention and least-squares surrogate structured prediction. It is important to notice that beyond ridge regression, any vector-valued regression model that writes as (8) is eligible. This is typically the case of Output Kernel tree-based methods (Geurts et al., 2006). Also, SOLA has here been applied to opinion analysis but it could prove suitable for more complex structure-labeling problems. Concerning Opinion Analysis, we have shown that abstention can be used to build a robust representation for star rating in a pipeline framework. One extension of our work would consist in learning how to abstain by jointly predicting the aspects and polarity at the sentence and text level.

Acknowledgements

This work has been funded by the french ministry of research and by the chair Machine Learning for Big Data of Télécom ParisTech.

References

BenTaieb, A. and Hamarneh, G. *Topology Aware Fully Convolutional Networks for Histology Gland Segmentation*, pp. 460–468. Springer International Publishing, Cham, 2016. ISBN 978-3-319-46723-8. doi: 10.1007/978-3-319-46723-8_53. URL https://doi.org/10.1007/978-3-319-46723-8_53.

Bi, W. and Kwok, J. T. Hierarchical multilabel classification with minimum bayes risk. In *Data Mining (ICDM), 2012 IEEE 12th International Conference on*, pp. 101–110. IEEE, 2012.

Brouard, C., Szafranski, M., and d'Alché Buc, F. Input output kernel regression: supervised and semi-supervised structured output prediction with operator-valued kernels. *Journal of Machine Learning Research*, 17(176):1–48, 2016.

Cesa-Bianchi, N., Gentile, C., and Zaniboni, L. Hierarchical classification: combining bayes with svm. In *Proceedings of the 23rd international conference on Machine learning*, pp. 177–184. ACM, 2006.

Ciliberto, C., Rosasco, L., and Rudi, A. A consistent regularization approach for structured prediction. In Lee, D. D., Sugiyama, M., Luxburg, U. V., Guyon, I., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 29*, pp. 4412–4420. Curran Associates, Inc., 2016.

Clavel, C. and Callejas, Z. Sentiment analysis: from opinion mining to human-agent interaction. *IEEE Transactions on affective computing*, 7(1):74–93, 2016.

Conneau, A., Kiela, D., Schwenk, H., Barrault, L., and Bordes, A. Supervised learning of universal sentence representations from natural language inference data. *CoRR*, abs/1705.02364, 2017. URL http://arxiv.org/abs/1705.02364.

Cortes, C., DeSalvo, G., and Mohri, M. Boosting with abstention. In Lee, D. D., Sugiyama, M., Luxburg, U. V., Guyon, I., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 29*, pp. 1660–1668. Curran Associates, Inc., 2016.

Deng, J., Ding, N., Jia, Y., Frome, A., Murphy, K., Bengio, S., Li, Y., Neven, H., and Adam, H. Large-scale object classification using label relation graphs. In *European Conference on Computer Vision*, pp. 48–64. Springer, 2014.

Dimitrovski, I., Kocev, D., Loskovska, S., and Džeroski, S. Hierchical annotation of medical images. In *Proceedings of the 11th International Multiconference - Information Society IS 2008*, pp. 174–181. IJS, Ljubljana, 2008.

Geurts, P., Wehenkel, L., and d'Alché-Buc, F. Kernelizing the output of tree-based methods. In *Machine Learning, Proceedings of the Twenty-Third International Conference (ICML 2006), Pittsburgh, Pennsylvania, USA, June 25-29, 2006*, pp. 345–352, 2006.

Goh, C. Y. and Jaillet, P. Structured Prediction by Conditional Risk Minimization. *ArXiv e-prints*, November 2016.

Marcheggiani, D., Täckström, O., Esuli, A., and Sebastiani, F. Hierarchical multi-label conditional random fields for aspect-oriented opinion mining. In *ECIR*, pp. 273–285. Springer, 2014.

Micchelli, C. A. and Pontil, M. On learning vector-valued functions. *Neural computation*, 17(1):177–204, 2005.

Osokin, A., Bach, F. R., and Lacoste-Julien, S. On structured prediction theory with calibrated convex surrogate losses. In *Advances in Neural Information Processing Systems 30*, pp. 301–312, 2017.

Pontiki, M., Galanis, D., Papageorgiou, H., Androutsopoulos, I., Manandhar, S., Mohammad, A.-S., Al-Ayyoub, M., Zhao, Y., Qin, B., De Clercq, O., et al. Semeval-2016 task 5: Aspect based sentiment analysis. In *Proceedings of the 10th international workshop on semantic evaluation (SemEval-2016)*, pp. 19–30, 2016.

Rousu, J., Saunders, C., Szedmak, S., and Shawe-Taylor, J. Kernel-based learning of hierarchical multilabel classification models. *Journal of Machine Learning Research*, 7 (Jul):1601–1626, 2006.

Schrijver, A. *Theory of linear and integer programming*. John Wiley & Sons, 1998.

Sutton, C., McCallum, A., et al. An introduction to conditional random fields. *Foundations and Trends® in Machine Learning*, 4(4):267–373, 2012.

Wang, H., Lu, Y., and Zhai, C. Latent aspect rating analysis without aspect keyword supervision. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 618–626. ACM, 2011.