A Comparison of Discriminative Classifiers for Web News Content Extraction

Alex Spengler
Université Paris 6
Paris, France
alex.spengler@lip6.fr

Antoine Bordes
Université Paris 6
Paris, France
antoine.bordes@lip6.fr

Patrick Gallinari
Université Paris 6
Paris, France
patrick.gallinari@lip6.fr

ABSTRACT

Until now, approaches to web content extraction have focused on random field models, largely neglecting large margin methods. Structured large margin methods, however, have recently shown great practical success. We compare, for the first time, greedy and structured support vector machines with conditional random fields on a real-world web news content extraction task, showing that large margin approaches are indeed competitive with random field models.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—Information filtering; I.5.1 [**Pattern Recognition**]: Models—Statistical

General Terms

Algorithms, Experimentation, Performance

Keywords

Web Content Extraction, SVM, CRF

1. INTRODUCTION

Recent years have witnessed a rapid growth of information on the web which is comprised of data covering all kinds of sources, topics and media. To cater the distinct information and dissemination needs, numerous specialized online publishing platforms have emerged, including, for example, web forums, blogs, social networking and news sites. Despite the common semantics that characterize each of those platforms, its web pages may still exhibit a breathtaking *variety* in *content*, *layout*, *structure*, and *style*. For instance, every news web page contains a title; the title's concrete representation in terms of positioning on the page, HTML code and visual style, however, varies considerably with sites *and* pages. In addition, the core information on a web page is often surrounded by a serious amount of clutter that guarantees the functioning of the site, but is non-essential to the page. Examples of such noisy content are visual and textual advertisements, links to all kinds of related and unrelated pages, menus, user polls, questionnaires, form fields and so on. Due to this *heterogeneity*, identifying the primary content in web page data is a challenging task.

In this paper, we address the issue of extracting semantically interesting content from a set of truly heterogeneous web pages, focussing on the *news* vertical. The extraction of interesting content can also be viewed as a transformation of pages from multiple schema into a set of pages which follow a single, unified schema. Such a transformation leverages a number of important applications, ranging from automatic text summarization over speech rendering for the visually impaired to information retrieval [2].

Earlier work on web content extraction (e.g. [6]) is dominated by the construction of fairly complex, yet fixed algorithmic solutions. Often these methods are template-dependent (so-called wrappers); wrappers search for patterns by means of regular expressions or properties of the tree structure of the document object model (DOM). More recent methods formulated the extraction problem as a classification task in which we seek to assign the correct semantic label (such as `Title` and `Author`, for instance) to each region of a web page. Most work [9, 4, 7] uses linear-chain conditional random fields (CRFs) [5] as classifiers. CRFs are probabilistic graphical models which take first-order Markov dependencies among the labels into account. Zhu et al. [10, 11] present models with more sophisticated structural dependencies, yet rest in the random field framework.

The discriminative approach to classification explicitly concentrates on getting the labels right, resulting in additional modeling freedom compared to generative models. However, conditional random fields are not the only possible discriminative classifier. Large margin classifiers, such as the structured support vector machine (SVM), can equally deal with interdependent labels and have recently shown great empirical success in a number of application domains.

In this paper, we present an experimental comparison of different discriminative classifiers. We confront, for the first time, conditional random fields and large margin methods on a web content extraction task. In particular, we compare a CRF classifier with its large margin counterpart, the structured SVM. In addition to this, we compare both with a simple and computationally less expensive alternative to the structured SVM, termed the greedy SVM. Our experimental evaluation is based on the new News600 data set [7]. It comprises 604 entire real-world news web pages from over 170 different domains that have been manually labeled at

RIAO'10, 2010, Paris, France

DOM node level. The results show that the structured SVM is a good choice which at least matches the overall accuracy of the CRF and significantly outperforms it on the non-frequent, but interesting labels. Surprisingly, the greedy SVM almost matches the performance of the structured SVM on these rare labels, even if its overall accuracy is slightly worse. It is thus worth considering the greedy alternative when learning from a huge number of examples.

2. WEB CONTENT EXTRACTION AS A CLASSIFICATION PROBLEM

We formulate the web content extraction task as a *classification problem* in which we seek the correct semantic label y_r for each region r of a web page. y_r takes values in a predefined and application-dependent set of semantic labels $\mathcal{Y}$, e.g. $\mathcal{Y} = \{\texttt{Title}, \texttt{Author}, \texttt{Paragraph}\ldots\}$ for news content extraction. We use $x_r \in \mathcal{X}$, on the other hand, to describe a region r in terms of certain characteristics or *features* (cf. Section 4.1), such as r's position on the page, its text content or font style. For a given web page with features $\boldsymbol{x} = (x_1, \ldots, x_\mathrm{R})^\mathrm{T}$, the general goal is thus to find all correct semantic labels $\boldsymbol{y} = (y_1, \ldots, y_\mathrm{R})^\mathrm{T}$.

Due to the considerable *inter-* and *intra-page variation*, globally consistent characteristics or prototypical representations of a given semantic label y_r do, in general, not exist. In fact, the semantics of a region are predominantly defined *relative* to its neighbouring regions and their respective semantics. For example, the title of a web page is generally positioned *above* the principal text content and its font size is typically *larger* than the font size on the rest of the page. It is essential to take into account not only the local features of a web page region, but also the relations between them.

Instead of classifying each region in isolation, *structured prediction* allows to jointly infer the semantic labels of all regions of a web page. By taking dependencies among the labels into account, structured methods often outperform classifiers that split the joint classification problem into a number of independent tasks. However, the true dependencies among the regions of a web page are unknown to us. The simplest and most natural hypothesis concerning the dependency structure in web content extraction links the semantic labels y_r of a page sequentially, traversing the leaf nodes of the DOM tree from the top of the web page to its bottom. Such a linear chain is a reasonable choice, since the content of a web page is concentrated in the (visible) leaf nodes of the DOM tree. It is thus often sufficient to consecutively visit these leaf nodes, mimicking the natural order of information on a web page. Moreover, this order of regions in the DOM tree frequently translates into a horizontally or vertically aligned region layout. In particular, regions in a very regular DOM subtree tend to have a sequential page layout (menus, for example). A study by Zhu et al. [10] underpins this observation.

More formally, we encode the interdependencies in the input-output space through a vector-valued function ϕ which acts on both the features $\boldsymbol{x}$ and the semantic labels $\boldsymbol{y}$. Then, the models used in this paper all belong to the family of *linear classifiers*:

$$h_{\boldsymbol{\theta}}(\boldsymbol{x}) = \arg\max_{\boldsymbol{y}\in\mathcal{Y}^\mathrm{R}} \langle \boldsymbol{\theta}, \phi(\boldsymbol{x}, \boldsymbol{y})\rangle, \qquad (1)$$

where the parameters $\boldsymbol{\theta} \in \Theta$ designate a particular member of the family and $\langle\cdot,\cdot\rangle$ denotes the inner product. The term $\langle \boldsymbol{\theta}, \phi(\boldsymbol{x}, \boldsymbol{y})\rangle$ is called the score of $\boldsymbol{\theta}, \boldsymbol{x}$ and $\boldsymbol{y}$.

Instead of using a fixed combination or a greedy subset of the available characteristics, we *learn* the importance θ_k of each ϕ_k from a concrete sample $\mathcal{D} = \{(\boldsymbol{x}^{(i)}, \boldsymbol{y}^{(i)})\}_{i=1}^N$ of N labeled web pages. This adaptability is a crucial advantage when facing the uncertainties caused by the heterogeneity of web data. To learn $\boldsymbol{\theta}$, we avail ourselves of a *loss function* $\ell(\boldsymbol{x}, \boldsymbol{y}, h(\boldsymbol{x}))$ which quantifies the discrepancy of the prediction $h(\boldsymbol{x})$ of classifier h and the correct labeling $\boldsymbol{y}$, given the corresponding input $\boldsymbol{x}$. Statistical learning theory shows that the classifier h that minimizes the *regularized empirical risk*

$$\mathcal{R}^\ell_\mathrm{reg}[h] = \frac{\lambda}{2}\|\boldsymbol{\theta}\|^2 + \frac{1}{N}\sum_{i=1}^N \ell(\boldsymbol{x}^{(i)}, \boldsymbol{y}^{(i)}, h(\boldsymbol{x}^{(i)})) \qquad (2)$$

also minimizes the true mean error on a previsously unseen document $\boldsymbol{x}$.

3. DISCRIMINATIVE CLASSIFIERS

Unlike generative models, discriminative classifiers learn a direct map from features $\boldsymbol{x}$ to the labels $\boldsymbol{y}$. They are hence particularly successful in situations in which it is difficult to properly specify class-conditional densities. This is the case in web content extraction in which we wish to incorporate a large variety of interdependent and long-range features of the data. The discriminative approach to classification therefore provides crucial modeling freedom in the web setting.

In this paper, we compare several flavours of discriminative classifiers, contrasting local with global maximization schemes. Due to the lack of space, however, we limit ourselves to a description of the principal differences and refer the reader to the original publications for more details.

3.1 Local Classifiers

Local classifiers split the maximization over all possible semantic labels $\boldsymbol{y} \in \mathcal{Y}^\mathrm{R}$ in eq. (1) into a series of successive maximizations of regional scores. Although local methods infer the semantic label y_r of each region r of a page independent of the other labels, i.e. *greedily*, they might take preceding predictions into account by augmenting the feature vector x_r. Let $h_\mathrm{local}(\boldsymbol{x}) = \big(h_1(\boldsymbol{x}), \ldots, h_\mathrm{R}(\boldsymbol{x})\big)^\mathrm{T}$ be the collection of the R *local* or *greedy* classifiers $h_r(\boldsymbol{x})$ defined recursively as

$$h_r(\boldsymbol{x}) = \arg\max_{y\in\mathcal{Y}} \langle \boldsymbol{\theta}, \phi\big(x_r, h_{r-c}(\boldsymbol{x}), \ldots, h_{r-1}(\boldsymbol{x}), y\big)\rangle, \qquad (3)$$

where the predictions $h_{r-c}(\boldsymbol{x}), \ldots, h_{r-1}(\boldsymbol{x})$ can be interpreted as a semantic aggregation of the data $x_{r-c}, \ldots, x_{r-1}$. The context size $c \in \{0, \ldots, \mathrm{R}-1\}$ determines how many of these prior predictions are considered. Please note that the recursion in eq. (3) hinges on the particular order among the web page regions (cf. Section 2).

Learning the classifier h_local is carried out by minimizing the risk (2) with a loss defined as a sum of local terms, one for each web page region r:

$$\ell(\boldsymbol{x}, \boldsymbol{y}, h_\mathrm{local}(\boldsymbol{x})) = \sum_{r=1}^\mathrm{R} \ell_\mathrm{r}(\boldsymbol{x}, \boldsymbol{y}, h_\mathrm{local}(\boldsymbol{x})).$$

Following Bordes et al. [1], we consider the hinge loss for ℓ_{r}:

$$\begin{aligned}\ell_{\mathrm{r}}(\boldsymbol{x}, \boldsymbol{y}, h_{\mathrm{local}}(\boldsymbol{x})) = 1 - \delta_{y_r, h_r(\boldsymbol{x})} \\ + \langle \boldsymbol{\theta}, \phi\big(x_r, h_{r-c}(\boldsymbol{x}), \ldots, h_r(\boldsymbol{x})\big)\rangle \\ - \langle \boldsymbol{\theta}, \phi\big(x_r, y_{r-c}, \ldots, y_r\big)\rangle\end{aligned}$$

where δ denotes the Kronecker delta, i.e. $\delta_{y,\hat{y}} = 1$ if $y = \hat{y}$ and 0 otherwise. The evaluation of the loss $\ell(\boldsymbol{x}, \boldsymbol{y}, h_{\mathrm{local}}(\boldsymbol{x}))$ is hence equally determined by the recursion order, since the parameters $\boldsymbol{\theta}$ are generally updated after each application of a local classifier $h_r(\boldsymbol{x})$. Note also that for $c = 0$ the recursion disappears and we recover the standard multi-class SVM [3] setting. The two principal advantages of greedy classifiers are

(i) a *fast* inference process with a total computational complexity of $O(\mathrm{R}|\mathcal{Y}|)$ and

(ii) the possibility of including approximated *long-range* dependencies through large context sizes (previously predicted semantic labels are assumed immutable).

3.2 Global Classifiers

Unlike local classifiers, global classification infers the semantic labels $\hat{\boldsymbol{y}} = h(\boldsymbol{x})$ of a previously unseen web page $\boldsymbol{x}$ by maximizing the sum of local scores over *all* possible label configurations $\boldsymbol{y} \in \mathcal{Y}^{\mathrm{R}}$:

$$\begin{aligned}h_{\mathrm{global}}(\boldsymbol{x}) &= \arg\max_{\boldsymbol{y} \in \mathcal{Y}^{\mathrm{R}}} \langle \boldsymbol{\theta}, \phi(\boldsymbol{x}, \boldsymbol{y})\rangle \\ &= \arg\max_{\boldsymbol{y} \in \mathcal{Y}^{\mathrm{R}}} \langle \boldsymbol{\theta}, \big(\sum_{r=1}^{\mathrm{R}} \phi(x_r, y_r), \sum_{r=2}^{\mathrm{R}} \phi(y_{r-1}, y_r)\big)\rangle,\end{aligned}$$

where we hypothesized $\phi(\boldsymbol{x}, \boldsymbol{y})$ to decompose into local features and pairwise label transitions. Such a decomposition of ϕ is equivalent to making a first-order Markov assumption for the semantic labels $\boldsymbol{y}$ of a page $\boldsymbol{x}$, connecting the individual region labels y_r through a linear chain (see Section 2 for a description of the underlying assumptions).

A naïve approach to the exact maximization has complexity $O(|\mathcal{Y}|^{\mathrm{R}})$, since it requires to iterate over all label configurations $\boldsymbol{y} \in \mathcal{Y}^{\mathrm{R}}$. By exploiting the linear-chain dependency structure in $\boldsymbol{y}$ using dynamic programming, the Viterbi algorithm computes the *exact* maximum over all possible configurations in $O(\mathrm{R}|\mathcal{Y}|^2)$ steps. The global maximization therefore comes at the price of increased computational complexity when compared with local classifiers. Moreover, interactions among regions are only among direct neighbours.[1]

In this paper, we compare two global classifiers that share the same inference process, but employ different loss functions ℓ in training. The *structured SVM* [8] implements a generalization of the hinge loss to an entire web page:

$$\begin{aligned}\ell(\boldsymbol{x}, \boldsymbol{y}, h_{\mathrm{global}}(\boldsymbol{x})) &= \Delta(\boldsymbol{y}, h_{\mathrm{global}}(\boldsymbol{x})) \\ &\quad + \langle \boldsymbol{\theta}, \phi(\boldsymbol{x}, h_{\mathrm{global}}(\boldsymbol{x}))\rangle - \langle \boldsymbol{\theta}, \phi(\boldsymbol{x}, \boldsymbol{y})\rangle \\ &= \max_{\hat{\boldsymbol{y}} \in \mathcal{Y}^{\mathrm{R}}} \Big\{\Delta(\boldsymbol{y}, \hat{\boldsymbol{y}}) + \langle \boldsymbol{\theta}, \phi(\boldsymbol{x}, \hat{\boldsymbol{y}})\rangle\Big\} \\ &\quad - \langle \boldsymbol{\theta}, \phi(\boldsymbol{x}, \boldsymbol{y})\rangle\end{aligned}$$

where $\Delta(\boldsymbol{y}, \hat{\boldsymbol{y}}) = \mathrm{R} - \sum_{r=1}^{\mathrm{R}} \delta_{y_r, \hat{y}_r}$ is the Hamming loss.

Conditional random fields[2] [5], in contrast, use the differentiable log loss (replacing the max with a soft-max):

$$\begin{aligned}\ell(\boldsymbol{x}, \boldsymbol{y}, h_{\mathrm{global}}(\boldsymbol{x})) &:= - \log p(\boldsymbol{y}|\boldsymbol{x}) \\ &= \log \sum_{\hat{\boldsymbol{y}} \in \mathcal{Y}^{\mathrm{R}}} \exp\{\langle \boldsymbol{\theta}, \phi(\boldsymbol{x}, \hat{\boldsymbol{y}})\rangle\} - \langle \boldsymbol{\theta}, \phi(\boldsymbol{x}, \boldsymbol{y})\rangle.\end{aligned}$$

Minimizing the risk in eq. (2) thus corresponds to maximizing the conditional log posterior under an isotropic Gaussian prior. We see that the log loss is in fact never zero, but that it gets smaller once the scores for all incorrect labels $\hat{\boldsymbol{y}}$ get very small.

4. EXPERIMENTAL EVALUATION

We compare the performance of local and global discriminative classifiers on the task of web news content extraction.

4.1 Experimental Setup

The NEWS600 web page corpus [7] consists of 604 web pages from 177 different domains. It therefore is truly heterogeneous. In total, the corpus contains 165,654 manually annotated, individual regions (visible leaf nodes), which is equivalent to an average of 274 regions per web page (the maximum is 1255). There are nine semantic labels: `Advertisement`, `Author`, `Caption`, `Date`, `Multimedia`, `None`, `Paragraph`, `Subtitle` and `Title`. `None` is the label that subsumes all regions which are uninteresting to us here, such as navigation menus. `Multimedia` marks images and animations which are tightly linked with the primary textual content (labeled `Paragraph`) and `Caption` points to captions of `Multimedia`. The 604 web pages in the NEWS600 data set are divided uniformly at random into 300 pages for training, 100 pages for validation and 204 pages for testing purposes. For each visible leaf node in the pages' DOM trees we collect both its label y_r and its features x_r. We remove all features that appear less than three times in the training set.

We use a large variety of atomic features x_r to characterize a web page region r. We combine cues from all available sources, including text content, DOM structure, CSS style

Table 1: Examples of feature functions for web news content extraction, split into four categories.

Text Features
all tokens in text content, has-text-content, all-capitals, all-capitals-firstword, text-content-string-length, text-content-word-count, output-position, output-siblings-next,. . .
DOM Features
all attributes, all tokens in attribute values for `class` and `name`, tokens in next & previous comments, dom-depth, dom-name, previous-depth-difference, node-dom-siblings-next, dom-parent-name, node-in-iframe, p-average-depth-difference,. . .
Style & Layout Features
text-color, background-color, top, bottom, left, right, height, width, surface, opacity, border-left-color, border-bottom-width, max-font-size-difference, next-font-size-difference, margin-left, next-left-difference, previous-height-difference, padding-bottom,. . .
Task-Specific Features
title-word-match, title-levenshtein-distance, author-cue, contributor-cue, date-cue, email-cue, months-cue, number-cue paragraph-cue, publisher-cue, time-cue, weekday-cue,. . .

[1] Higher-order interactions are possible, but very costly.

[2] Although generally the case, we require a CRF to be a conditional distribution in the exponential family.

Table 2: Experimental results (in per cent) comparing four different discriminative classifiers.

Model	Micro-averaged F_1-measure	Macro-averaged F_1-measure	Overall Accuracy
Multi-class SVM	95.19	84.33	95.20 $\star$
Greedy SVM	95.04	86.41	95.02 $\star$
Linear-chain CRF	95.72	85.63	95.72
Structured SVM	**96.06**	**86.64**	**96.06**

as well as the positioning information recovered from the browser's rendering engine. Table 1 shows a subset of the $106,436$ atomic observational features we extracted.

We compare two local classifiers, a multi-class SVM ($c = 0$) and a greedy SVM ($c = 10$), and two global classifiers, a structured SVM and a CRF. All reported results are based on the best model-specific hyper-parameter settings as determined on the validation set. In particular, the SVMs worked best with $\lambda = 0.1$, whereas for the CRF we used $\lambda = 1$. For the greedy SVM, we tested the context sizes $1, 2, 5, 7, 10, 13, 15, 25$, finding $c = 10$ to give the best results. All models have been run to convergence.

We measure the performance of our model on the *individual* labels using precision, recall and F_1-measure. Although appropriate when judging the overall performance, the accuracy on *individual* labels is dominated by true negatives (due to the significant imbalance among labels), which easily leads to misinterpretation.

4.2 Results

Table 2 shows the overall performance of the four classifiers, revealing three main results:[3]

(i) Global classifiers outperform local classifiers. Not only is the maximization of a global joint score theoretically exact, it is also practically preferable to a greedily maximized series of local scores and at a reasonable supplementary cost.

(ii) The structured SVM improves on the CRF. The use of the generalized hinge loss leads to a slight increase in the overall accuracy (and the micro-averaged F_1-measure) compared with the log loss and significantly boosts the macro-averaged F_1-measure. This is an important result, meaning that we improve on average on the individual labels and in particular on the *rare* labels such as `Title` and `Author`.

(iii) Greedy SVMs perform surprisingly well on the macro-averaged F_1-measure. Including previous predictions in order to approximate long-range interactions on the labels might hence provide a computationally viable alternative to the more expensive structured SVMs, especially in large-scale settings in which we wish to learn a classifier from hundreds of thousands of examples or more.

Table 3 shows the performance of the structured SVM on the individual labels. It is the labels `Author` and `Subtitle` which are hardest to get right (due to a low recall). `Date` instead suffers from low precision.

[3] The $\star$ indicates statistical significance (from the structured SVM) according to a Wilcoxon matched-pairs signed-ranks test and significance level $p \leq 0.05$.

Table 3: Performance results (in per cent) of the structured SVM on the individual labels.

Label	n_{train}	n_{test}	Precision	Recall	F_1-measure
`Advertisement`	12452	6685	90.35	89.35	89.85
`Author`	307	221	86.96	63.35	73.30
`Caption`	127	82	92.00	84.15	87.90
`Date`	306	208	73.18	91.83	81.45
`Multimedia`	175	123	94.64	86.18	90.21
`None`	61796	39535	97.78	97.33	97.55
`Paragraph`	5194	3192	89.72	97.90	93.63
`Subtitle`	33	20	80.00	60.00	68.57
`Title`	300	204	98.98	95.59	97.26
Micro-Averages	—	—	96.11	96.06	96.06
Macro-Averages	—	—	89.29	85.07	86.64

5. CONCLUSION

We compared the performance of four discriminative classifiers for the extraction of semantically interesting content from news web pages. Our empirical evaluation shows that the CRF is not the only pertinent classification framework for such a task. Structured SVMs exhibit similar overall performance and significantly outperform the CRF on the rare labels. Another surprising result is the high macro-averaged F_1-score obtained for the greedy SVM, making it a fast alternative to the structured SVM in situations in which learning via Viterbi inference is too costly.

6. REFERENCES

[1] A. Bordes, N. Usunier, and L. Bottou. Sequence labelling SVMs trained in one pass. In *ECML/PKDD*, pages 146–161. Springer, 2008.

[2] D. Cai, S. Yu, J.-R. Wen, and W.-Y. Ma. Block-based web search. In *SIGIR*, pages 456–463, 2004.

[3] K. Crammer and Y. Singer. On the algorithmic implementation of multiclass kernel-based vector machines. *JMLR*, 2:265–292, 2001.

[4] J. Gibson, B. Wellner, and S. Lubar. Adaptive web-page content identification. In *Intl. Workshop on Web Information and Data Management*, pages 105–112, 2007.

[5] J. D. Lafferty, A. McCallum, and F. C. N. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *ICML 18*, pages 282–289, 2001.

[6] S.-H. Lin and J.-M. Ho. Discovering informative content blocks from web documents. In *SIGKDD 8*, pages 588–593, 2002.

[7] A. Spengler and P. Gallinari. Learning to extract content from news webpages. *Intl. Conf. on Advanced Information Networking and Applications Workshops*, pages 709–714, 2009.

[8] I. Tsochantaridis, T. Joachims, T. Hofmann, and Y. Altun. Large margin methods for structured and interdependent output variables. *JMLR*, 6:1453–1484, 2005.

[9] Y. Xue, Y. Hu, G. Xin, R. Song, S. Shi, Y. Cao, C.-Y. Lin, and H. Li. Web page title extraction and its application. In *Information Processing & Management*, volume 43, pages 1332–1347, 2007.

[10] J. Zhu, Z. Nie, J.-R. Wen, B. Zhang, and W.-Y. Ma. 2D conditional random fields for web information extraction. In *ICML 22*, pages 1044–1051, 2005.

[11] J. Zhu, Z. Nie, B. Zhang, and J.-R. Wen. Dynamic hierarchical Markov random fields and their application to web data extraction. In *ICML 24*, pages 1175–1182, 2007.