Combining Visual and Textual Features for Information Extraction from Online Flyers

Emilia Apostolova
BrokerSavant Inc
2506 N. Clark St.
Chicago, IL 60614
emilia@brokersavant.com

Noriko Tomuro
DePaul University
243 S. Wabash Ave.
Chicago, IL 60604
tomuro@cs.depaul.edu

Abstract

Information in visually rich formats such as PDF and HTML is often conveyed by a combination of textual and visual features. In particular, genres such as marketing flyers and info-graphics often augment textual information by its color, size, positioning, etc. As a result, traditional text-based approaches to information extraction (IE) could underperform. In this study, we present a supervised machine learning approach to IE from online commercial real estate flyers. We evaluated the performance of SVM classifiers on the task of identifying 12 types of named entities using a combination of textual and visual features. Results show that the addition of visual features such as color, size, and positioning significantly increased classifier performance.

1 Introduction

Since the Message Understanding Conferences in the 1990s (Grishman and Sundheim, 1996; Chinchor and Robinson, 1997), Information Extraction (IE) and Named Entity Recognition (NER) approaches have been applied and evaluated on a variety of domains and textual genres. The majority of the work, however, focuses on the journalistic, scientific, and informal genres (newswires, scientific publications, blogs, tweets, and other social media texts) (Nadeau and Sekine, 2007) and deals with purely textual corpora. As a result, the feature space of NER systems involves purely textual features, typically word attributes and characteristics (orthography, morphology, dictionary lookup, etc.), their contexts and document features (surrounding word window, local syntax, document/corpus word frequencies, etc.) (Nadeau and Sekine, 2007).

At the same time, textual information is often presented in visually rich formats, e.g. HTML and PDF. In addition to text, these formats use a variety of visually salient characteristics, (e.g. color, font size, positioning) to either highlight or augment textual information. In some genres and domains, a textual representation of the data, excluding visual features is often not enough to accurately identify named entities of interest or extract relevant information. Marketing materials, such as online flyers or HTML emails, often contain a plethora of visual features and text-based NER approaches lead to poor results. In this paper, we present a supervised approach that uses a combination of textual and visual features to recognize named entities in online marketing materials.

2 Motivation and Problem Definition

A number of broker-based industries (e.g. commercial real estate, heavy equipment machinery, etc.) lack a centralized searchable database with industry offerings. In particular, the commercial real estate industry (unlike residential real estate) does not have a centralized database or an established source of information. Commercial real estate brokers often need to rely on networking, chance, and waste time with a variety of commercial real estate databases that often present outdated information. While brokers do not often update third party inventory databases, they do create marketing materials (usually PDF flyers) that contain all relevant listing information. Virtually all commercial real estate offerings come with publicly available marketing material that contains all relevant listing information. Our goal is to harness this source of information (the marketing flyer) and use it to extract structured listing information.

Figure 1 shows an example of a commercial real estate flyer. The commercial real estate flyers are often distributed as PDF documents, links to HTML pages, or visually rich HTML-based

Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 1924–1929,
October 25-29, 2014, Doha, Qatar.

Figure 1: An example of a commercial real estate flyer © Kudan Group Real Estate.

emails. They typically contain all relevant listing information such as the address and neighborhood of the offering, the names and contact information of the brokers, the type of space offered (building, land, unit(s) within a building), etc. Similar to other info-graphics, relevant information could be easily pinpointed by visual clues. For example, the listing street address in Figure 1 (*1629 N. Halsted St.*, upper left corner) can be quickly identified and distinguished from the brokerage firm street address (*156 N. Jefferson St.*, upper right corner) due to its visual prominence (font color, size, positioning).

In this study we explored a supervised machine learning approach to the task of identifying listing information from commercial real estate flyers. In particular, we focused on the recognition of 12 types of named entities as described in Table 1 below.

3 Related Work

Nadeau and Satoshi (2007) present a survey of NER and describe the feature space of NER research. While they mention multi-media NER in the context of video/text processing, all described features/approaches focus only on textual representation.

Broker Name Broker Email Broker Phone Company Phone	The contact information of all listing brokers, including full name, email address, phone number. The brokerage company phone number.
Street City Neighborhood State Zip	The address information of the listing address including street or intersection, city, neighborhood, state, and zip code.
Space Size Space Type	Size and attributes of relevant spaces (e.g. *27,042 SF building, 4.44 acres site*, etc.); Mentions of space type descriptors, e.g. building, land/lot, floor, unit. This excludes space type and size information of non-essential listing attributes (e.g. basement size or parking lot size).
Confidential	Any mentions of confidentiality.

Table 1: Types and descriptions of named entities relevant to extracting listing information from commercial real estate flyers.

The literature on Information Extraction from HTML resources is dominated by various approaches based on wrapper induction (Kushmerick, 1997; Kushmerick, 2000). Wrapper inductions rely on common HTML structure (based on the HTML DOM) and formatting features to extract structured information from similarly formatted HTML pages. This approach, however, is not applicable to the genres of marketing materials (PDF and HTML) since they typically do not share any common structure that can be used to identify relevant named entities. Laender et al. (2002) present a survey of data extraction techniques and tools from structured or semi-structured web resources.

Cai et al. (2003) present a vision-based segmentation algorithm of web pages that uses HTML layout features and attempts to partition the page at the semantic level. In (Burget and Rudolfova, 2009) authors propose web-page block classification based on visual features. Yang and Zhang (2001) build a content tree of HTML documents based on *visual consistency* inferred semantics. Burget (2007) proposes a layout based information extraction from HTML documents and states that this visual approach is more robust than traditional DOM-based methods.

Changuel et al.(2009a) describe a system for automatically extracting author information from web-pages. They use spatial information based on the depth of the text node in the HTML DOM tree. In (Changuel et al., 2009b) and (Hu et al., 2006),

the authors proposed a machine learning method for title extraction and utilize format information such as font size, position, and font weight. In (Zhu et al., 2007) authors use layout information based on font size and weight for NER for automated expense reimbursement.

While the idea of utilizing visual features based on HTML style has been previously suggested, this study tackles a non-trivial visually rich dataset that prevents the use of previously suggested simplistic approaches to computing HTML features (such as relying on the HTML DOM tree or simplistic HTML style rendering). In addition, we introduce the use of RGB color as a feature and normalize it approximating human perception.

4 Dataset and Method

The dataset consists of 800 randomly selected commercial real estate flyers spanning 315 US locations, 75 companies, and 730 brokers. The flyers were collected from various online sources and were originally generated using a variety of HTML and PDF creator tools. The collection represents numerous flyer formats and layouts, commercial real estate property types (*industrial, retail, office, land*, etc.), and transactions (*investment, sale, lease*).

All flyers were converted to a common format (HTML)[1]. The HTML versions of all documents were then annotated by 2 annotators. Figure 2 shows an example of an annotated flyer. Annotation guidelines were developed and the 2 annotators were able to achieve an inter-annotator agreement of 91%[2]. The named entities with lowest inter-annotator agreement were entities describing Space Size and Type because of the somewhat complex rules for determining essential listing space information. For example, one of the space size/type rules reads as follows: *If the listing refers to a building and mentions the lot size, include both the land size, the building size, and corresponding space types. Do not include individual parts of the building (e.g. office/basement) as separate spaces. If the listing refers to a UNIT within the building, not the whole building, then DO NOT include the land site as a separate space.*

A supervised machine learning approach was

[1]PDFs were converted to HTML using the PDFTOHTML conversion program `http://pdftohtml.sourceforge.net/`.

[2]The inter-annotator agreement was measured as F1-score using one of the annotator's named entities as the gold standard set and the other as a comparison set.

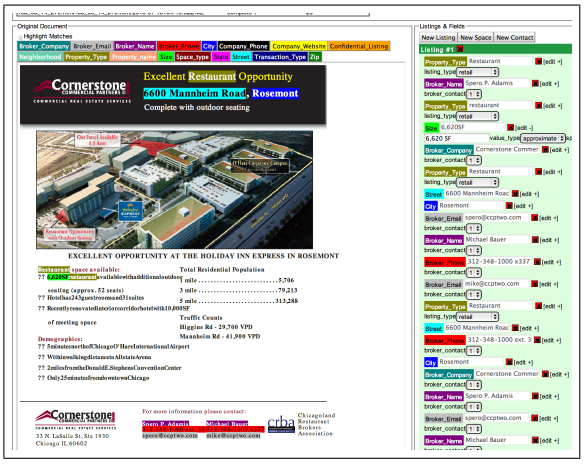

Figure 2: The HTML versions of the flyers were annotated by 2 annotators using a custom web-based annotation tool.

then applied to the task of identifying the 12 named entities shown in Table 1. Flyers were converted to text using an HTML parser while preserving some of the white space formatting. The text was tokenized and the task was then modeled as a **BIO** classification task, classifiers identify the **B**eginning, the **I**nside, and **O**utside of the text segments. We first used a traditional set of text-based features for the classification task. Table 2 lists the various text-based features used. In all cases, a sliding window including the 5 preceding and 5 following tokens was used as features.

Feature Name	Description
Token	A normalized string representation of the token. All tokens were converted to lower case and all digits were converted to a common format.
Token Orth	The token orthography. Possible values are lowercase (all token characters are lower case), all capitals (all token characters are upper case), upper initial (the first token character is upper case, the rest are lower case), mixed (any mixture of upper and lower case letters not included in the previous categories).
Token Kind	Possible values are word, number, symbol, punctuation.
Regex type	Regex-based rules were used to mark chunks as one of 3 regex types: email, phone number, zip code.
Gazetteer	Text chunks were marked as possible US cities or states based on US Census Bureau city and state data. www.census.gov/geo/maps-data/data/gazetteer2013.html.

Table 2: List of text-based features used for the NER task. A sliding window of the 5 preceding and 5 following tokens was used for all features.

As noted previously, human annotators were able to quickly spot named entities of interest solely because of their visual characteristics. For example, a text-only version of the flyer shown in Figure 1, stripped of all rich formatting, will make it quite difficult to distinguish the listing address (shown in prominent size, position, and color) from the brokerage company address, which is rarely prominent as it is not considered important information in the context of the flyer. Similarly, the essential size information for the listing shown on Figure 2 appears prominently on the first page (square footage of the offered restaurant), while non-essential size information, such as the size of the adjacent parking lot or basement, tend to appear in smaller font on subsequent flyer pages.

To account for such visual characteristics we attempted to also include visual features associated with text chunks. We used the computed HTML style attributes for each DOM element containing text. Table 3 lists the computed visual features.

Feature Name	Description
Font Size	The computed *font-size* attribute of the surrounding HTML DOM element, normalized to 7 basic sizes (*xx-small, x-small, small, medium, large, x-large, xx-large*).
Color	The computed *color* attribute of the surrounding HTML DOM element. The RGB values were normalized to a set of 100 basic colors. We converted the RGB values to the YUV color space, and then used Euclidian distance to find the most similar basic color approximating human perception.
Y Coordinate	The computed *top* attribute of the surrounding HTML DOM element, i.e. the y-coordinate in pixels. The pixel locations was normalized to 150 pixel increments (roughly 1/5th of the visible screen for the most common screen resolution.)

Table 3: List of visual features used for the NER task. A sliding window of 5 preceding and 5 following DOM elements were used for all features.

Computing the HTML style attributes is a complex task since they are typically defined by a combination of CSS files, in-lined HTML style attributes, and browser defaults. The complexities of style definition, inheritance, and overwriting are handled by browsers[3]. We used the Chrome browser to compute dynamically the style of each DOM element and output it as inline style attributes. To achieve this we programmatically inserted a javascript snippet that inlines the computed style and saves the new version of the HTML on the local file system utilizing the HTML5 *saveAs* interface[4]. Details on how we normalized the style attribute values for font size, RGB color, and Y coordinate are shown in Table 3.

We then applied Support Vector Machines (SVM) (Vapnik, 2000) on the NER task using the LibSVM library (Chang and Lin, 2011). We chose SVMs as they have been shown to perform well on a variety of NER tasks, for example (Isozaki and Kazawa, 2002; Takeuchi and Collier, 2002; Mayfield et al., 2003; Ekbal and Bandyopadhyay, 2008). We used a linear kernel model with the default parameters. The multi-class problem was converted to binary problems using the one-vs-others scheme. 80% of the documents were used for training, and the remaining 20% for testing.

5 Results

Results are shown in Table 4. We compared classifier performance using only textual features (first 3 columns), versus performance using both textual and visual features (next 3 columns). Results were averaged over 2 runs of randomly selected training/test documents with 80%/20% ratio. We used an exact measure which considers an answer to be correct only if both the entity boundaries and entity type are accurately predicted.

The addition of visual features significantly[5] increased the overall F1-score from 83 to 87%. As expected, performance gains are more significant for named entities that are typically visually salient and are otherwise difficult (or impossible) to identify in a text-only version of the flyers. Named Entities referring to listing address information showed the most significant improvements. In particular, the F1-score for mentions of *Neighborhoods* (typically prominently shown on the first page of the flyers) improved by 19%; F1-score for mentions of the listing *State* improved by 9%; and *Street, City, Zip* by roughly 4% each, all

[3]We attempted to use an HTML renderer from the Cobra java toolkit `http://lobobrowser.org/cobra.jsp` to compute HTML style attributes. However, this renderer produced poor results on our dataset and failed to accurately compute the pixel location of text elements.

[4]`https://github.com/eligrey/FileSaver.js`

[5]The difference is statistically significant with p value $<$ 0.0001% using Z-test on two proportions.

Named Entity	Pt	Rt	Ft	Pv+t	Rv+t	Fv+t	S
Broker Name	82.7	91.7	87.0	95.0	91.6	93.2	Y
Broker Email	92.3	92.8	92.6	97.2	90.2	93.6	N
Broker Phone	90.2	86.1	88.1	94.7	85.2	89.7	N
Company Ph.	95.2	67.4	78.9	89.8	65.4	75.7	N
Street	87.4	70.5	78.1	87.3	77.3	82.0	Y
City	92.5	88.5	90.5	94.9	92.8	93.8	Y
Neighborhood	68.2	52.8	59.5	85.3	72.9	78.6	Y
State	77.4	97.5	86.3	95.8	95.0	95.4	Y
Zip	89.7	94.5	92.1	96.1	97.1	96.6	Y
Space Size	80.2	65.0	71.8	87.0	70.6	77.9	Y
Space Type	76.0	74.7	75.3	78.6	72.2	75.3	N
Confidential	100	60.0	75.0	75.0	85.7	79.9	N
OVERALL	84.8	81.3	83.0	91.2	83.2	87.0	Y

Table 4: Results from applying SVM using the textual features described in Table 2, as well as both the textual and visual features described in Tables 2 and 3. t=textual features only, v+t=visual + textual features, P=Precision, R=Recall, F=F1-score, S=Significant Difference

statistically significant. Visual clues are also typically used when identifying relevant size information and, as expected, performance improved significantly by roughly 6%. The difference in performance for mentions used to describe confidential information is not statistically significant[6] because such mentions rarely occurred in the dataset. Similarly, performance differences for Company Phone, Broker Phone, Broker Email, and Space Type are not statistically significant. In all of these cases, visual features did not influence performance and text-based features proved adequate predictors.

6 Conclusion

We have shown that information extraction in certain genres and domains spans different media - textual and visual. Ubiquitous online and digital formats such as PDF and HTML often exploit the interaction of textual and visual elements. Information is often augmented or conveyed by non-textual features such as positioning, font size, color, and images. However, traditionally, NER approaches rely exclusively on textual features and as a result could perform poorly in visually rich genres such as online marketing flyers or infographics. We have evaluated the performance gain on the task of NER from commercial real estate flyers by adding visual features to a set of traditional text-based features. We used SVM classifiers for the task of identifying 12 types of named entities. Results show that overall visual features improved performance significantly.

[6] p value = 0.7323% using Z-test on two proportions.

References

Radek Burget and Ivana Rudolfova. 2009. Web page element classification based on visual features. In *Intelligent Information and Database Systems, 2009. ACIIDS 2009. First Asian Conference on*, pages 67–72. IEEE.

Radek Burget. 2007. Layout based information extraction from html documents. In *Document Analysis and Recognition, 2007. ICDAR 2007. Ninth International Conference on*, volume 2, pages 624–628. IEEE.

Deng Cai, Shipeng Yu, Ji-Rong Wen, and Wei-Ying Ma. 2003. Extracting content structure for web pages based on visual representation. In *Web Technologies and Applications*, pages 406–417. Springer.

Chih-Chung Chang and Chih-Jen Lin. 2011. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

Sahar Changuel, Nicolas Labroche, and Bernadette Bouchon-Meunier. 2009a. Automatic web pages author extraction. In *Flexible Query Answering Systems*, pages 300–311. Springer.

Sahar Changuel, Nicolas Labroche, and Bernadette Bouchon-Meunier. 2009b. A general learning method for automatic title extraction from html pages. In *Machine Learning and Data Mining in Pattern Recognition*, pages 704–718. Springer.

Nancy Chinchor and Patricia Robinson. 1997. Muc-7 named entity task definition. In *Proceedings of the 7th Conference on Message Understanding*.

Asif Ekbal and Sivaji Bandyopadhyay. 2008. Named entity recognition using support vector machine: A language independent approach. *International Journal of Computer Systems Science & Engineering*, 4(2).

Ralph Grishman and Beth Sundheim. 1996. Message understanding conference-6: A brief history. In *COLING*, volume 96, pages 466–471.

Yunhua Hu, Hang Li, Yunbo Cao, Li Teng, Dmitriy Meyerzon, and Qinghua Zheng. 2006. Automatic extraction of titles from general documents using machine learning. *Information processing & management*, 42(5):1276–1293.

Hideki Isozaki and Hideto Kazawa. 2002. Efficient support vector classifiers for named entity recognition. In *Proceedings of the 19th international conference on Computational linguistics-Volume 1*, pages 1–7. Association for Computational Linguistics.

Nicholas Kushmerick. 1997. *Wrapper induction for information extraction*. Ph.D. thesis, University of Washington.

Nicholas Kushmerick. 2000. Wrapper induction: Efficiency and expressiveness. *Artificial Intelligence*, 118(1):15–68.

Alberto HF Laender, Berthier A Ribeiro-Neto, Altigran S da Silva, and Juliana S Teixeira. 2002. A brief survey of web data extraction tools. *ACM Sigmod Record*, 31(2):84–93.

James Mayfield, Paul McNamee, and Christine Piatko. 2003. Named entity recognition using hundreds of thousands of features. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4*, pages 184–187. Association for Computational Linguistics.

David Nadeau and Satoshi Sekine. 2007. A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 30(1):3–26.

Koichi Takeuchi and Nigel Collier. 2002. Use of support vector machines in extended named entity recognition. In *proceedings of the 6th conference on Natural language learning-Volume 20*, pages 1–7. Association for Computational Linguistics.

Vladimir Vapnik. 2000. *The nature of statistical learning theory*. springer.

Yudong Yang and HongJiang Zhang. 2001. Html page analysis based on visual cues. In *Document Analysis and Recognition, 2001. Proceedings. Sixth International Conference on*, pages 859–864. IEEE.

Guangyu Zhu, Timothy J Bethea, and Vikas Krishna. 2007. Extracting relevant named entities for automated expense reimbursement. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1004–1012. ACM.