Leveraging HTML in Free Text Web Named Entity Recognition

Colin Ashby and **David Weir**
TAG Lab, Department of Informatics, University of Sussex, Brighton, UK
{c.ashby, d.j.weir}@sussex.ac.uk

Abstract

HTML tags are typically discarded in free text Named Entity Recognition from Web pages. We investigate whether these discarded tags might be used to improve NER performance. We compare Text+Tags sentences with their Text-Only equivalents, over five datasets, two free text segmentation granularities and two NER models. We find an increased F1 performance for Text+Tags of between 0.9% and 13.2% over all datasets, variants and models. This performance increase, over datasets of varying entity types, HTML density and construction quality, indicates our method is flexible and adaptable. These findings imply that a similar technique might be of use in other Web-aware NLP tasks, including the enrichment of deep language models.

1 Introduction

Named Entity Recognition (NER) is the identification of the proper names of objects. NER can be informational in its own right, but also serves as pre-processing for other Information Extraction (IE) tasks. The Web as a data source for NER offers opportunities as a massive corpus of structured information and free text, but also presents challenges of a far from uniform layout and variable writing style; the Web is formatted for ease of reading, not for ease of extraction.

Web page text elements can be divided into two categories: structured and unstructured free text.

Structured text is contained within records, lists and tables. NER for these elements is usually performed with Wrapper methods that identify text areas using precise contextual patterns. HTML is an essential delimiter of records in Wrapper methods. Techniques for handling HTML tags range from string matching to CNN based approaches. Wrappers are insufficient for dealing directly with free text.

Unstructured free text natural language elements, such as <p>, make up the majority of Web content. Most recent free text NER approaches make use of sentence based NLP neural network techniques (Goyal et al., 2018; Yadav and Bethard, 2019; Li et al., 2020), such as BERT (Devlin et al., 2018) or LSTM+CRF (Ma and Hovy, 2016), with pre-trained language models. HTML contained in these free text sentences is generally discarded. The reasons for this seem to be expectation, convenience and heritage. HTML in natural language is not expected to be of any benefit: "Free text exhibits no implicit structure at all" (Goebel and Ceresna, 2009). Any information to be extracted from free text is assumed to be held entirely in natural language grammar and semantics. Many NLP code libraries, such as Beautiful Soup (Richardson, 2007), strip HTML with a single function call: "Since so much text on the Web is in HTML format, we will also see how to dispense with markup." (Bird et al., 2009). A plain text approach to NER is used in various genres of Web site, from Web newspapers (Ekbal et al., 2012; Wibawa and Purwarianti, 2016), the social Web (Russell, 2013) and Websites identified through Web search (Tu et al., 2005; Bunescu, 2007; Speck and Ngomo, 2014) to elements of well-researched shared tasks, such as the scientific Web page section of BioNLP 2013 (Nédellec et al., 2013) and Weblog sections of CoNLL-2003 (Sang and De Meulder, 2003) and Ontonotes v5.0 (Weischedel et al., 2013).

In the five datasets used in this paper, HTML tags make up between 10% and 34% of tokens.

HTML tags have been used in NER, seemingly as a side effect of a combined approach to NER from both structured and free text elements (Soderland, 1999; Whitelaw et al., 2008; Mirończuk, 2018). We observe that these approaches typically employ the same split of structured and free text techniques detailed above. To our knowledge, these are the only approaches that make use of HTML tags in free text areas.

Supervised NER is the most commonly used approach and has attracted the greatest research effort (Nadeau and Sekine, 2007; Goyal et al., 2018), especially related to deep learning (Yadav and Bethard, 2019; Li et al., 2020). Supervised NER has performed better in F1 terms than semi-supervised in a direct comparison (Aryoyudanta et al., 2016) and on the CoNLL-2003 shared task.

In this paper we seek to answer the following questions, which form our contribution:

- To what extent does the inclusion of HTML tags in free text affect NER performance?
- What are the causes of this effect and which Web pages benefit the most?
- Can HTML tags be included efficiently in sentence based NER?

Our experiments use five English language datasets of varying characteristics, split into two different free text containing tagsets, split again into Text+Tags and Text-Only variants. Each HTML tag is represented by a single token. We assess these datasets and variants over BERT and LSTM+CRF models.

We show that the inclusion of HTML tags improves NER F1 performance over every dataset, tagset and model. Pages with a higher tag density correlate fairly well with higher performance gains. The inclusion of tags helps delimit entities, can perform well with any semi-structured elements that might be present and appears to mitigate annotation noise to some extent.

These results point to some intriguing future possibilities. HTML might be included in other Web-sourced NLP tasks, such as Knowledge Base Population or integrated into deep language models trained on the Web.

2 Related Work

To extract free text from a Web page, the level of granularity of free text containing elements must be decided. Approaches to free text segmentation range from indiscriminate whole-page approaches to stripping HTML (Bird et al., 2009; Richardson, 2007). A more granular approach is to use specific HTML tags that are deemed text containing (Bunescu, 2007; Sarode et al., 2019), further filtering may then be applied (Kohlschütter et al., 2010; Kim, 2017).

To include HTML tags in sentence based NER, we have a range of options. Blohm (2011) and Mirończuk (2018) simplify tags, removing attributes and style, including the tag as one standard token. Soderland (1999) takes a similar approach, including attributes, splitting into tokens by whitespace. Freitag (1998) engineers features to represent aspects of HTML positions relative to a token. Apostolova and Tomuro (2014) combine visual and textual page content into engineered features for NER. Gogar et al. (2016) use a CNN to automatically extract similar features.

3 Approach

We conduct NER experiments on the five datasets detailed below. These datasets allow sentences, including HTML tags, to be extracted from the Web while applying a supplied gold standard set of entity label annotations or Distant Supervision.

We extract two sets of free text elements and their inner HTML contents from the `<body>` of raw Web pages, removing script and comment blocks. Set1 contains `<p>` tags and contents, Set2 contains `<p>` and `<h...>` tags and contents. We ignore any structured elements contained in a table record, list item or option. The contents of these elements are then tokenized on whitespace and HTML tags, including *all* tags. Each tag is simplified into a single token, removing attributes and style. These free text sequences are then sentence segmented using NLTK (Bird et al., 2009) to create a Text+Tags variant. HTML is stripped from each sentence to create a directly comparable Text-Only variant. "`<h3>` `<a>` Australia `</a>` and the world `</h3>`" is a typical sentence.

	Entity		Sentence count		Tag density%		Avg. sentence len.	
Dataset	Types	Categories	Set1	Set2	Set1	Set2	Set1	Set2
OrgPersons	1	1	8198	10901	11	13	24.6	20.1
Persons	1	1	121598	186523	16	21	22.4	15.9
RE3D	14	7	1393	2528	11	19	15.6	11.3
SWDE	16	8	49849	113902	10	12	22.0	19.4
WEIR	14	4	3796	10858	30	34	22.6	6.4

Table 1: Dataset and tagset summary. Tag density% is HTML tag tokens as a proportion of all tokens. Average sentence length includes HTML tag tokens.

3.1 Datasets

RE3D[1] (Science and Technology Laboratory, 2017) contains entities relevant to the role of a defence and security intelligence analyst. We generated this dataset from the live pages of seven sites using the gold standard. **SWDE**[2] (Hao et al., 2011) contains entities from eight semantically diverse categories for Structured Web Data Extraction testing. We generated this dataset from the supplied cached pages using the gold standard. **WEIR**[3] (Bronzi et al., 2013) contains entities from four categories and 40 sites. We generated this dataset from the supplied cached pages using the gold standard. We include an additional free text containing tag, <a>, in tag Set2 due to a sparseness of <h...> tags. These three sets are evaluated using stratified cross validation. **Persons** was constructed to test Distant Supervision (Mintz et al., 2009), for the task of extracting person attributes from organisation Websites. This set is the first stage of this process and contains only person name entity types. The construction process extracts all persons from DBpedia (Bizer et al., 2009), then uses the top-10 Web search results from each person name as a page corpus. Each page has free text areas extracted, then direct string matches for forename and surname, plus possessive apostrophes, are labelled. Distant supervision is a noisy process (Roth et al., 2013); we exclude noisy sentences by the presence of a non-labelled forename or title present from two lookup lists. This set is evaluated against a hand labelled set of 1,214 sentences extracted from 30 Websites. Annotation was performed by the authors, with an inter-annotator agreement of 98.5%. Disagreements were due to incorrectly labelled titles, these were corrected to not include titles. **OrgPersons**, constructed in a similar way to Persons, is person-only but is oriented toward a person's employing organisation, rather than the person themselves. Organisations and corresponding key persons are extracted from DBpedia, with the top-5 Web search results from the organisation name processed. Each result is processed three pages deep, with the same matching and exclusion criteria as Persons. This set is evaluated using the same set as Persons. This is a smaller task-focussed set; the three page deep processing is likely to hit more of the types of pages that were labelled in our evaluation set.

3.2 Models

We use two recent state-of-the-art NLP models. A Bi-LSTM+CNN+CRF based on Ma and Hovy (2016), with Word2Vec skip-gram embeddings (Mikolov et al., 2013), generated with Gensim (Rehurek and Sojka, 2010) over the five datasets. We generated four variants of embeddings to cover tag Set1 and Set2, Text+Tags and Text-Only, using 100 dimensions, 20 iterations and window 5. We also included pre-trained Stanford GloVe embeddings (Pennington et al., 2014) as a baseline. For our second model, we fine-tuned a BERT (Devlin et al., 2018) transformer using the Hugging Face (Wolf et al., 2019) bert-base-cased language model. These models and embeddings provide a good contrast for our experiments.

4 Results and Analysis

Our results in Table 2 show increased F1 performance for Text+Tags over every dataset, tagset and model.

[1] https://github.com/dstl/re3d

[2] https://archive.codeplex.com/?p=swde

[3] http://www.dia.uniroma3.it/db/weir/

	Text-Only LSTM		Text-Only	Text+Tags LSTM		Text+Tags	Text+Tags
Dataset.Tagset	GloVe	W2V	BERT	GloVe	W2V	BERT	Improvement
OrgPersons.1	85.5	84.0	80.5		86.1	**88.1**	+2.6
OrgPersons.2	85.1	82.7	81.2		**89.0**	87.1	+3.9
Persons.1	70.9	67.3	69.0	**73.8**	70.3	72.8	+2.9
Persons.2	74.1	70.6	70.3	**77.2**	71.0	76.8	+3.1
RE3D.1	72.4	71.6	71.9		72.8	**73.4**	+1.0
RE3D.2	74.7	74.3	73.6		**75.6**	75.0	+0.9
SWDE.1	61.7	64.0	74.8		**76.6**	76.0	+1.8
SWDE.2	66.8	68.0	82.6		84.6	**86.6**	+4.0
WEIR.1	75.4	72.9	87.5		89.4	**91.5**	+4.0
WEIR.2	64.1	65.2	70.5		**83.7**	73.4	+13.2

Table 2: Full F1 results. Improvement of best Text+Tags approach over best Text-Only approach.

Sentence characteristics

This improvement correlates fairly well (Pearson correlation coefficient of 0.72) with the tag densities in Table 1, suggesting dataset tag density is applicable for assessing a future dataset for our Text+Tags technique. Analysis of the distribution of sentence tag densities, reveals two types of sentence: natural language sentences containing some tags and repetitive tag dense patterns, for example, "<h1> John Smith </h1>". We find Text+Tags performs slightly better for variants that contain a mix of sentence types and performs much better for WEIR.2 which is pattern dominant.

Entity delimitation

We analysed LSTM results, looking at ratios between occurrences of tags that delimit successful and unsuccessful entity labels. We find that HTML tags delimit between 16% and 31% of entities and that entity-closing tags have a success ratio between 122% and 251% better than entity-opening tags. An example is the sentence "<h2> Traveler Advice on Little Delhi Restaurant </h2>" where the labelled entity is Little Delhi Restaurant. When we look at success ratios for individual HTML tags, we find opening tags </em>, <h...>, <em>, <strong>, <a>, <br/>, <span>, <div> and closing tags <br/>, <em>, <span>, </h...>, </span>, <img>, </a>, </p>, </div> perform well, while opening tags <i>, <p>, <br> and closing tags </i>, <br>, </strong> perform poorly. These lists are ordered by best/worse performance and are not exhaustive. Interesting points are that close tags are good openers and vice versa, the variable performance of different types of <br>, and that italics performs much worse than other formatting options, perhaps indicating poor annotation quality in SWDE where this was especially prevalent. Poor performing tags might be pre-processed out in future work to further improve performance.

Models

BERT shows a stable Text+Tags improvement between dataset variants, where sentence tag density can vary considerably and between datasets of differing quality. We observe increased precision on the single entity datasets: OrgPersons and Persons, with increased recall for other datasets. BERT is able to adapt to Text-Only patterns in the WEIR dataset that the LSTM model fails on. Our LSTM Word2Vec Text+Tags outperforms BERT 4/10 times, which might indicate a deep language model trained on Web text including HTML tags may further improve performance.

Our own LSTM Word2Vec embeddings show a Text+Tags improvement over all datasets and variants. On the Persons dataset, GloVe embeddings outperform our Text+Tags embeddings. This prompted us to experiment with Text+Tags GloVe which unexpectedly scored best overall for this dataset. This Text+Tags GloVe performance may indicate our own embeddings would benefit from more extensive training in both volume and context window.

LSTM shows a large Text+Tags improvement for SWDE and WEIR datasets due to poor precision on Text-Only. For the WEIR dataset, this is mainly due to patterns that are almost un-differentiable without delimiting tags. For SWDE, this is due to poor annotation from the gold standard, introducing many false negatives. An example is the sentence "What initially grabs your attention in <b> The Five People You Meet in Heaven </b> ?", the book title in bold is recognised with the tags present, but not without. Text+Tags and the more complex BERT models can deal with these scenarios.

5 Conclusion

We investigated whether HTML tags might improve NER performance by comparing Text+Tags sentences with their Text-Only equivalents, over varying datasets, free text segmentation methods and NER models. We used a simplified tags as tokens approach to sentence processing, which required minimal extra pre-processing and between 3% and 11% extra processing time. We observed F1 improvements of between 0.9% and 13.2% for Text+Tags over all datasets, variants and models. These datasets, variants and models had quite different characteristics, proving the flexibility and adaptability of our approach. This performance points to the inclusion of HTML in other free text NLP tasks such as Knowledge Base Population or deep language model generation from the Web.

Acknowledgements

We would like to thank our reviewers and members of Team Tom for their helpful feedback and suggestions. This work was supported by EPSRC grant EP/M508172/1.

References

Emilia Apostolova and Noriko Tomuro. 2014. Combining visual and textual features for information extraction from online flyers. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1924–1929.

Bayu Aryoyudanta, Teguh Bharata Adji, and Indriana Hidayah. 2016. Semi-supervised learning approach for indonesian named entity recognition (ner) using co-training algorithm. In *2016 International Seminar on Intelligent Technology and Its Applications (ISITIA)*, pages 7–12. IEEE.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.".

Christian Bizer, Jens Lehmann, Georgi Kobilarov, Sören Auer, Christian Becker, Richard Cyganiak, and Sebastian Hellmann. 2009. Dbpedia-a crystallization point for the web of data. *Journal of web semantics*, 7(3):154–165.

Sebastian Blohm. 2011. *Large-scale pattern-based information extraction from the world wide web*. KIT Scientific Publishing.

Mirko Bronzi, Valter Crescenzi, Paolo Merialdo, and Paolo Papotti. 2013. Extraction and integration of partially overlapping web sources. *Proceedings of the VLDB Endowment*, 6(10):805–816.

Razvan Constantin Bunescu. 2007. *Learning for information extraction: from named entity recognition and disambiguation to relation extraction*. Ph.D. thesis.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Asif Ekbal, Sriparna Saha, and Dhirendra Singh. 2012. Active machine learning technique for named entity recognition. In *Proceedings of the International Conference on Advances in Computing, Communications and Informatics*, pages 180–186.

Dayne Freitag. 1998. Information extraction from html: Application of a general machine learning approach. In *AAAI/IAAI*, pages 517–523.

Max Goebel and Michal Ceresna. 2009. Wrapper induction.

Tomas Gogar, Ondrej Hubacek, and Jan Sedivy. 2016. Deep neural networks for web page information extraction. In *IFIP International Conference on Artificial Intelligence Applications and Innovations*, pages 154–163. Springer.

Archana Goyal, Vishal Gupta, and Manish Kumar. 2018. Recent named entity recognition and classification techniques: a systematic review. *Computer Science Review*, 29:21–43.

Qiang Hao, Rui Cai, Yanwei Pang, and Lei Zhang. 2011. From one tree to a forest: a unified solution for structured web data extraction. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*, pages 775–784.

Maria Myung Hee Kim. 2017. Incremental knowledge acquisition approach for information extraction on both semi-structured and unstructured text from the open domain web. In *Proceedings of the Australasian Language Technology Association Workshop 2017*, pages 88–96.

Christian Kohlschütter, Peter Fankhauser, and Wolfgang Nejdl. 2010. Boilerplate detection using shallow text features. In *Proceedings of the third ACM international conference on Web search and data mining*, pages 441–450.

Jing Li, Aixin Sun, Jianglei Han, and Chenliang Li. 2020. A survey on deep learning for named entity recognition. *IEEE Transactions on Knowledge and Data Engineering*.

Xuezhe Ma and Eduard Hovy. 2016. End-to-end sequence labeling via bi-directional lstm-cnns-crf. *arXiv preprint arXiv:1603.01354*.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 1003–1011.

Marcin Michał Mirończuk. 2018. The biggrams: the semi-supervised information extraction system from html: an improvement in the wrapper induction. *Knowledge and Information Systems*, 54(3):711–776.

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

Claire Nédellec, Robert Bossy, Jin-Dong Kim, Jung-Jae Kim, Tomoko Ohta, Sampo Pyysalo, and Pierre Zweigenbaum. 2013. Overview of bionlp shared task 2013. In *Proceedings of the BioNLP shared task 2013 workshop*, pages 1–7.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Radim Rehurek and Petr Sojka. 2010. Software framework for topic modelling with large corpora. In *In Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. Citeseer.

Leonard Richardson. 2007. Beautiful soup documentation. *Dosegljivo: https://www. crummy. com/software/BeautifulSoup/bs4/doc/.[Dostopano: 7. 7. 2018]*.

Benjamin Roth, Tassilo Barth, Michael Wiegand, and Dietrich Klakow. 2013. A survey of noise reduction methods for distant supervision. In *Proceedings of the 2013 workshop on Automated knowledge base construction*, pages 73–78.

Matthew A Russell. 2013. *Mining the social web: data mining Facebook, Twitter, LinkedIn, Google+, GitHub, and more*. " O'Reilly Media, Inc.".

Erik F Sang and Fien De Meulder. 2003. Introduction to the conll-2003 shared task: Language-independent named entity recognition. *arXiv preprint cs/0306050*.

Rashmi P Sarode, Shelly Sachdeva, Wanming Chu, and Subhash Bhalla. 2019. Segment-search vs knowledge graphs: Making a key-word search engine for web documents. In *International Conference on Big Data Analytics*, pages 88–107. Springer.

The Defence Science and UK Technology Laboratory. 2017. Relationship and entity extraction evaluation dataset.

Stephen Soderland. 1999. Learning information extraction rules for semi-structured and free text. *Machine learning*, 34(1-3):233–272.

René Speck and Axel-Cyrille Ngonga Ngomo. 2014. Ensemble learning for named entity recognition. In *International semantic web conference*, pages 519–534. Springer.

Nguyen Cam Tu, Tran Thi Oanh, Phan Xuan Hieu, and Ha Quang Thuy. 2005. Named entity recognition in vietnamese free-text and web documents using conditional random fields. In *The 8th Conference on Some selection problems of Information Technology and Telecommunication*, page 12. Citeseer.

Ralph Weischedel, Martha Palmer, Mitchell Marcus, Eduard Hovy, Sameer Pradhan, Lance Ramshaw, Nianwen Xue, Ann Taylor, Jeff Kaufman, Michelle Franchini, et al. 2013. Ontonotes release 5.0 ldc2013t19. *Linguistic Data Consortium, Philadelphia, PA*, 23.

Casey Whitelaw, Alex Kehlenbeck, Nemanja Petrovic, and Lyle Ungar. 2008. Web-scale named entity recognition. In *Proceedings of the 17th ACM conference on Information and knowledge management*, pages 123–132.

Aditya Satrya Wibawa and Ayu Purwarianti. 2016. Indonesian named-entity recognition for 15 classes using ensemble supervised learning. *Procedia Computer Science*, 81:221–228.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Vikas Yadav and Steven Bethard. 2019. A survey on recent advances in named entity recognition from deep learning models. *arXiv preprint arXiv:1910.11470*.