Named Entity Recognition using Weak Supervision techniques

Vitor Vasconcelos de Oliveira
dept. of Computer Science
Universidade de Brasília (UNB)
Brasília, Brasil
vitorvasconcelos05062000@gmail.com

***Abstract*—Named Entity Recognition (NER) is a relevant task for extracting information from textual data, but traditionally NER training methods require human annotation to provide useful data for model training, however this process of manual labeling can be very much costly, in terms of money, time and effort. As an alternative to human labeling, Weak Supervision is a technique that essentially provides the possibility of machine labeled data by relying in heuristics and label functions to automatically annotate documents. This paper is an insight and an experimentation about the use of traditional Weak Supervision in Named Entity Recognition real live situations. Here, it will be set parameters to define traditional weak supervision as well as particular data to represent real world scenarios, experimentation will be conducted by testing not only Weak Supervision predicted labels but also results of a Conditional Random Field (CRF) model trained based in a Weak Supervision's resulting application dataset. By the end of the experimentation it will discussed about Weak Supervision's usefulness and unusefulness, as well as it's limitations and possibilities in Machine Learning.**

***Index Terms*—Named Entity Recognition(NER), Weak Supervision, Machine Learning, Conditional Random Field**

I. Introduction

Named Entity Recognition (NER) is one of the most acknowledged and challenging tasks in Natural Language Processing (NLP). Its main proposal is to extract and classify, by using various computation linguistics techniques, entities found in natural language texts. In the most recent years, this various computation techniques usually includes the training and testing of Deep Neural Network models. Although such models obtain promising results, they normally depend on large pools of data to provide a reliable source of information for training and to achieve the ideal behavior.

In this context, there is a growing need of reliable and labeled data to provide the models, which usually is supplied by human effort, by annotating and categorizing texts. This human participation in the labeling process is known as Human-in-the-Loop (HITL) [1], [2] and aims to speed up Machine Learning models training process. Even though, hastening is HITL's main objective, arguably in many projects this can be one of the most costly processes, in terms of time, money and effort.

Reading, searching, identifying, circumscribing and reviewing is the usual process of annotating textual data. As pointed before, this can be extremely expensive and additionally can still let slip some annotations errors, specially if the reviewing process is not properly conducted, by simply misjudging or by different approaches taken by different members of the annotation team. As said, time, funds and errors are common at human conducted annotations, but what if this process could be made by computers? Would it be better or worse? How could it be performed?

Weak Supervision is an alternative to HITL's approaches, once it focuses in bootstrapping labeled data with computer effort [3]. In better terms, weak supervision works by heuristically creating it's own categorized data by relying in label functions to automatically annotate documents. These functions can use different strategies on its labeling process such as regular expression patterns, class-indicative keywords or heuristic labeling functions [4], [5].

This paper aims at experimenting and better understanding the use of the Weak Supervision strategy in the usual Named Entity Recognition problem. For that, we will be designing, with the help of the Skweak and Spacy framework, label functions to identify our own entities in the "Contratos" dataset provided by an extraction of official documents from Diário Oficial do Distrito Federal (DODF) with the use of Brazilian Portuguese oriented textual analysis.

After training, will be compared the accuracy and the percentage precision in the entities, tokens and letters provided by the weak supervision based models to better understand the nuances in weak supervision approaches.

Overall, the main focus and contributions are the study an analysis of regular weak supervision with label functions aggregation on Brazilian Portuguese official documents, trying to understand it's efficiency, limitations, an possible improvements.

II. Related Work

Related to this line of research, there are many articles that not only implement regular weak supervision but improve on it. For example in [6] Submukhe, Zheng and Hassanam develop a framework for iterative self-training of deep neural networks with weak supervision and rule attenuation aiming to improve on the effectiveness of regular weak supervion.

In [7], Raphaelh, Clzhang, Xiaoling, Lsz and Weld presents an approach for multi-instance learning with overlapping relations which produces accurate sentence-level predictions,

decoding of individual sentences and is able to make corpus-level extractions.

But probably the most aligned inspiration to this article is [8] by Plison, Ahu, Jeremycb and Samiat, which presents weak supervision in broad spectrum and have an approach based in the aggregation of label functions using a Hidden Markov Model (HMM), that we as well will be using in this paper, and essentially achieves an successful approach with suited for sequence labelling tasks and probabilistic labelling predictions, by a vast number of different functions.

III. Proposed Method

As previously said, this paper aims to analyse the use of regular weak supervision label functions into Brazilian Portuguese official documents. By regular label functions it is understood simpler and more traditional label functions, that do not involve the text context, semantics or syntax but only it's structure and word sequences. Because of that, firstly, it is essential to understand the documents in study structures and the entities that need to be extracted.

The documents that are subjects of this research are the "Contratos" dataset, extracted from "Diário Oficial do Distrito Federal" (DODF). DODF is a public document from the Brazilian capital Brasília and it's region the Distrito Federal, in which are contained all the acts of public administration and services conducted. "Contratos" is a specific type of act in the document, in which corresponds to a regulated Contract between companies and the public state.

In "Contratos" there are many possible and important entities to be extracted. Hereafter, it will be presented next a list of the chosen entities and an image with an example act of "Contrato" 1:

- "Número do ajuste": Contract's number
- "Órgão contratante": Contracting body
- "Entidade contratada": Contracted entity
- "Entidades convenentes": Convening entities
- "Processo do GDF": Process number
- "Objeto do ajuste": Object to which the contract refers
- "Data de assinatura do ajuste": Contract's signature date (Data de assinatura)
- "Vigência do ajuste": Term of validity of the contract
- "Valor do ajuste" : Estimated final value of the contract
- "Código da unidade orçamentária": Contract's budget union (União orçamentária)
- "Programa de trabalho": Contract's work program (Programa de trabalho)
- "Natureza da despesa": Nature of contract's expenses (Natureza de despesa)
- "Nota de empenho": Contract's commitment note (Nota de empenho)

To perform the identification of these entities, there were designed two types of label functions for each. The first one is a Regular Expression (regex) oriented label function, it uses regex to specify and identify a search pattern in the act's text. The second is keyword detection oriented, that is, it uses the occurrence of specific words, punctuation and symbols to establish starting and ending patterns for the entity. Both types of labeling functions were designed and applied for each entity and had their results merged with one another by the use of the Hidden Markov Model (HMM) [9] aggregation method.

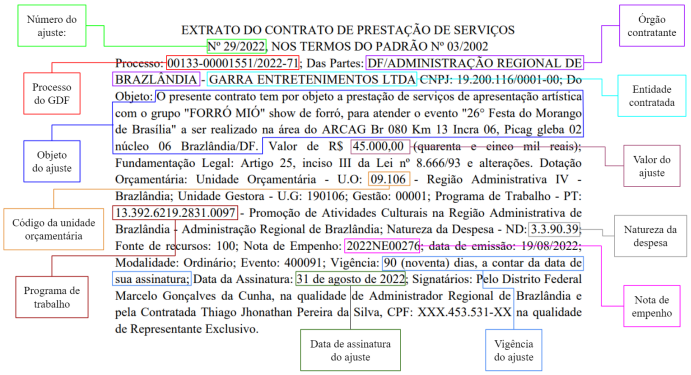

Fig. 1. "Contrato" act labeling example

All weak supervisions labeling functions were formulated and designed with the aid of the Skweak [10] framework for python, wich is a toolkit to easily define, apply and aggregate label functions. Skweak is also tightly integrated with SpaCy [11], another python framework disigned to help and solve NLP problems and tokenization of texts. Finally all Label functions were then incorporated in the dodfSkweak.py, a script that applies everything specified above and returns the result of weak supervision in "Contratos" acts in a .csv file and IOB format.

Also it was produced another python script, this time to generate and instantiate a Conditional Random Field (CRF) [12] model. dodfCRF.py uses the sklearn-crfsuite python framework to create and save a CRF model able to classify "Contratos" entities based in their neighbouring words and its characteristics, such as, the word itself, word length, capital letters, ponctuations and beginning or end of sentence.

With all the tools and methods above, everything was arranged for the study's application and experimentation. The flowchart in 1 summarizes the processes applied:

1) Dataset Gathering: extracting both, training and testing, datasets of "Contratos" from many DODF documents by using regex. The training dataset is made up by 48.982 "Contratos" acts, and has somewhat a big variability in structure and entities regularity to simulate a not perfect scenario. And the test or Golden pattern dataset made up by 456 acts manually picked and labeled, with human effort, to maintain correctness and precision.
2) dodfSkweak application in training Dataset: dodfSkweak.py was applied into the 48.982 training dataset and extracted it's entities with weak supervision.
3) dodfCRF application into resulting entities from dodfSkweak: dodfCRF.py trained a Conditional Random Field (CRF) model from the dodfSkweak.py resulting entities from the previous step.
4) Evaluation of Labels resulting from both dodfSkweak

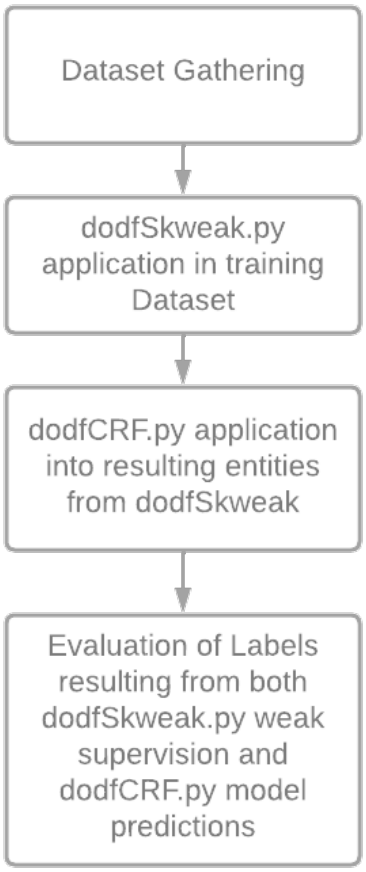

Fig. 2. Flowchart

weak supervision and dodfCRF model predictions: As a final step, both dodfSkweak.py label functions resulting entities and the CRF model, originated form dodfCRF.py, predictions will be tested and analysed in the Golden Pattern testing dataset.

IV. Experimental Results

After applying weak supervision into the 48.982 training dataset, a CRF model was trained based on the labeling. In this section it will be presented the results of this training and Weak Supervision's application, they will be shown as the custom accuracy of the model predictions and of the label functions "pure" predictions into the Golden Pattern testing dataset. Also a CRF model evaluation will be performed in the testing dataset to visualize IOB labels accuracy, model precision, recall and F1-score.

The testing will focus on how accurate the CRF model and the label function predictions are, by using four types of analysis:

1) EQUAL type: testing if the predicted entity is exactly equal to the actual entity accordingly to the Golden Pattern.
2) IN type: testing only if the predicted entity is included in the actual entity accordingly to the Golden Pattern. Here is relevant to iterate that cases where predicted entities extrapolate real entities size in characters are considered wrong predictions.
3) CONTAIN type: testing only if the predicted entity contains the actual entity accordingly to the Golden Pattern as a part of their structure. Here is relevant to iterate that cases where real entities extrapolate predicted entities size in characters are considered wrong predictions.
4) 80% type: testing if the predicted entity has at least 80% similarity of characters to the actual entity accordingly to the Golden Pattern.

In the tables bellow it will be presented the results of each type of analysis. the table's columns meaning are:

- Correct: Number of correct predictions.
- Percentage: Percentage of correct predictions on the total number of occurrences of this entity.
- Null: Number of times this entity doesn't occur in the Golden Pattern Dataset.
- Total: Total number of times this entity occur in the Golden Pattern Dataset.

The next four tables, EQUAL, IN, CONTAIN and 80%, in this order, represent the results of the Weak supervision label functions direct application under the golden dataset:

TABLE I
EQUAL type table for Weak supervision results

	Correct	Percentage	Null	Total
Número do ajuste:	350	82.94%	34	422
Órgão contratante:	4	0.47%	74	844
Entidade contratada:	15	1.79%	72	840
Entidades convenentes	0	0.00%	439	31
Processo do GDF:	357	80.95%	18	441
Objeto do ajuste:	254	56.82%	11	447
Data de assinatura do ajuste:	297	76.74%	70	387
Vigência do ajuste:	188	45.63%	46	412
Valor do ajuste:	266	75.57%	104	352
Código da unidade orçamentária:	160	83.33%	271	192
Programa de trabalho:	211	73.78%	197	286
Natureza da despesa:	177	81.57%	252	217
Nota de empenho:	178	73.86%	253	241

TABLE II
IN type table for Weak supervision results

	Correct	Percentage	Null	Total
Número do ajuste:	354	83.89%	34	422
Órgão contratante:	425	50.36%	74	844
Entidade contratada:	445	52.98%	72	840
Entidades convenentes	3	9.68%	439	31
Processo do GDF:	375	85.03%	18	441
Objeto do ajuste:	392	87.70%	11	447
Data de assinatura do ajuste:	332	85.79%	70	387
Vigência do ajuste:	317	76.94%	46	412
Valor do ajuste:	325	92.33%	104	352
Código da unidade orçamentária:	184	95.83%	271	192
Programa de trabalho:	238	83.22%	197	286
Natureza da despesa:	189	87.10%	252	217
Nota de empenho:	194	80.50%	253	241

TABLE III
CONTAIN TYPE TABLE FOR WEAK SUPERVISION RESULTS

	Correct	Percentage	Null	Total
Número do ajuste:	412	97.63%	34	422
Órgão contratante:	399	47.27%	74	844
Entidade contratada:	359	42.74%	72	840
Entidades convenentes	24	77.42%	439	31
Processo do GDF:	367	83.22%	18	441
Objeto do ajuste:	298	66.67%	11	447
Data de assinatura do ajuste:	326	84.24%	70	387
Vigência do ajuste:	263	63.83%	46	412
Valor do ajuste:	270	76.70%	104	352
Código da unidade orçamentária:	166	86.46%	271	192
Programa de trabalho:	236	82.52%	197	286
Natureza da despesa:	190	87.56%	252	217
Nota de empenho:	187	77.59%	253	241

TABLE IV
80% TYPE TABLE FOR WEAK SUPERVISION RESULTS

	Correct	Percentage	Null	Total
Número do ajuste:	383	90.76%	34	422
Órgão contratante:	34	4.03%	74	844
Entidade contratada:	134	15.95%	72	840
Entidades convenentes	2	6.45%	439	31
Processo do GDF:	364	82.54%	18	441
Objeto do ajuste:	385	86.13%	11	447
Data de assinatura do ajuste:	302	78.04%	70	387
Vigência do ajuste:	264	64.08%	46	412
Valor do ajuste:	291	82.67%	104	352
Código da unidade orçamentária:	163	84.90%	271	192
Programa de trabalho:	241	84.27%	197	286
Natureza da despesa:	193	88.94%	252	217
Nota de empenho:	214	88.80%	253	241

The next four tables, EQUAL, IN, CONTAIN and 80%, in this order, represent the results of the Conditional random field (CRF) direct application under the golden dataset:

TABLE V
EQUAL TYPE TABLE FOR CONDITIONAL RANDOM FIELD (CRF) RESULTS

	Correct	Percentage	Null	Total
Número do ajuste:	234	55.45%	34	422
Órgão contratante:	1	0.12%	74	844
Entidade contratada:	1	0.12%	72	840
Entidades convenentes	0	0.00%	439	31
Processo do GDF:	313	70.98%	18	441
Objeto do ajuste:	37	8.28%	11	447
Data de assinatura do ajuste:	289	74.68%	70	387
Vigência do ajuste:	40	9.71%	46	412
Valor do ajuste:	242	68.75%	104	352
Código da unidade orçamentária:	153	79.69%	271	192
Programa de trabalho:	197	68.88%	197	286
Natureza da despesa:	174	80.18%	252	217
Nota de empenho:	167	69.29%	253	241

It can be seen in the results above the behavior of each entity in each scenario. Most entities in all eight tables presents percentage of correctness between 70% and 80%, also it is notable that the scores correspondent to the direct Weak supervision applications usually presents higher values than the Conditional Random Field results, which can possibly mean some kind of misrepresentation during model training,

TABLE VI
IN TYPE TABLE FOR CONDITIONAL RANDOM FIELD (CRF) RESULTS

	Correct	Percentage	Null	Total
Número do ajuste:	355	84.12%	34	422
Órgão contratante:	459	54.38%	74	844
Entidade contratada:	465	55.36%	72	840
Entidades convenentes	9	29.03%	439	31
Processo do GDF:	331	75.06%	18	441
Objeto do ajuste:	70	15.66%	11	447
Data de assinatura do ajuste:	331	85.53%	70	387
Vigência do ajuste:	125	30.34%	46	412
Valor do ajuste:	293	83.24%	104	352
Código da unidade orçamentária:	178	92.71%	271	192
Programa de trabalho:	220	76.92%	197	286
Natureza da despesa:	188	86.64%	252	217
Nota de empenho:	187	77.59%	253	241

TABLE VII
CONTAIN TYPE TABLE FOR CONDITIONAL RANDOM FIELD (CRF) RESULTS

	Correct	Percentage	Null	Total
Número do ajuste:	239	56.64%	34	422
Órgão contratante:	219	25.95%	74	844
Entidade contratada:	295	35.12%	72	840
Entidades convenentes	14	45.16%	439	31
Processo do GDF:	335	75.96%	18	441
Objeto do ajuste:	47	10.51%	11	447
Data de assinatura do ajuste:	337	87.08%	70	387
Vigência do ajuste:	63	15.29%	46	412
Valor do ajuste:	296	84.09%	104	352
Código da unidade orçamentária:	165	85.94%	271	192
Programa de trabalho:	226	79.02%	197	286
Natureza da despesa:	192	88.48%	252	217
Nota de empenho:	204	84.65%	253	241

TABLE VIII
80% TYPE TABLE FOR CONDITIONAL RANDOM FIELD (CRF) RESULTS

	Correct	Percentage	Null	Total
Número do ajuste:	235	55.69%	34	422
Órgão contratante:	19	2.25%	74	844
Entidade contratada:	85	10.12%	72	840
Entidades convenentes	2	6.45%	439	31
Processo do GDF:	372	84.35%	18	441
Objeto do ajuste:	381	85.23%	11	447
Data de assinatura do ajuste:	298	77.00%	70	387
Vigência do ajuste:	261	63.35%	46	412
Valor do ajuste:	261	74.15%	104	352
Código da unidade orçamentária:	157	81.77%	271	192
Programa de trabalho:	236	82.52%	197	286
Natureza da despesa:	186	85.71%	252	217
Nota de empenho:	180	74.69%	253	241

being it the poor quality of the training data provided by the weak supervision or even model overfitting.

Now focusing in the non ”regular” entities, those which presented the most contrasting results. ”Órgão contratante”, ”Entidade contratada”, ”Entidades convenentes” presented lower results in EQUAL, IN and 80% analysis but considerate higher results in the CONTAIN analysis, this behavior can be understood by the similarity of the three entities and the way they appear in textual data usually clumped in an single ”PARTES” contract section, representing contracting parts,

which made very complex the differentiation between them by the labeling functions. Because of this the decision was to unite this three entities into one only "PARTES" entity and to possibly separate them previously by human effort or other techniques.

By analyzing the tables in comparison to each other, it is possible to infer that, especially in the Weak supervision case, the CONTAIN table has the best results or higher percentages ending in a average of 74.91%. This can possibly mean that the Weak Supervision applied may not be 100% precise in it's predictions, but it is still extracting the right entities even if sometimes accompanied by other characters.

Finally, in the next table it's possible to visualize CRF's precision, recall and F1-score for each label in IOB format, also the overall value of the model score was 0.9689.

TABLE IX
CONDITIONAL RANDOM FIELD (CRF) EVALUATION TABLE FROM GOLDEN DATASET APPLICATION

	Precision	Recall	F1-score	Support
B-CONTRATO	0.896	0.631	0.740	122
I-CONTRATO	0.973	0.486	0.649	74
B-PARTES	0.973	0.985	0.979	395
I-PARTES	0.939	0.981	0.960	6067
B-OBJETO	0.984	0.969	0.976	445
I-OBJETO	0.977	0.980	0.978	14733
B-VIGENCIA	0.989	0.981	0.985	361
I-VIGENCIA	0.977	0.988	0.982	5639
B-UNI_ORC.	1.000	0.994	0.997	167
B-NOTA_EMP.	0.996	0.974	0.985	229
B-VALOR	0.878	0.949	0.912	409
I-NOTA_EMP.	0.950	0.826	0.884	23
B-PROCESSO	0.974	0.904	0.938	491
I-PROCESSO	0.992	0.910	0.949	133
B-DATA_ASS	0.982	0.974	0.978	383
I-DATA_ASS	0.988	0.966	0.977	958
B-PROG_TRAB	0.952	0.968	0.960	247
B-NAT_DESP	0.955	0.946	0.950	202
I-PROG_TRAB	0.898	0.964	0.930	55
I-NAT_DESP	0.947	0.750	0.837	24
I-UNI_ORC	1.000	0.444	0.615	9
I-VALOR	0.000	0.000	0.000	0
Micro-Avg	**0.967**	**0.972**	**0.970**	**31496**
Macro-Avg	**0.919**	**0.844**	**0.871**	**31496**
Weighted-Avg	**0.967**	**0.972**	**0.969**	**31496**

V. CONCLUSION

The high cost and labeling effort is a known problem of Human-in-the-Loop textual data annotation. Weak supervision is also a known alternative to this costly process, but as shown here, having it's limitations intertwined with textual structure, on how standardized and well defined are the labeled entities, and label function adaptability into each entity structure.

This paper method functionates by standardizing weak supervision labeling functions into two simpler types and textual data into regular yet sometimes flawed dataset. This process aims to approximate paper results to real life applications in a relatable manner, making possible to better understand weak supervision's worth in real life applications.

By analyzing the previous shown results, it's possible to infer that weak supervision, even in it's simpler forms, are indeed capable of mitigating labeling effort and is a very considerable alternative to manual labeling. It is shown that weak supervision has managed to achieve a 56.41%, 74.71%, 74.91%, 65.96% medium percentage of correctness in the the EQUAL, IN, CONTAIN and 0.8 tables, being most entities numbers in between 70% and 80%. Also the CRF model trained with the weak supervision results showed similar results being 45.08%, 65.12%, 59.53%, 60.25% medium percentage of correctness for the EQUAL, IN, CONTAIN and 0.8 tables and also most entities numbers in between 70% and 80%.

Yet, it's also possible to find it's main flaws and improvement areas, especially by identifying the issues in the lower scoring entities. The mains issues located in this study are the extrapolation of entity limits, such as marking punctuation and even words before or after entities, difficulties in finding explicit and well defined text characteristics for each entity and finally the hardship of identifying and selecting the right label in cases of occurrence of more than one possible entity, especially by context.

Studies on resolving these previously pointed issues are or even were already done. Works such as [6], [7] and [8] implement different and more complex forms of weak supervision and have shown, with the presentation of promising results, that it is indeed possible to overcome this issues. With all that said, even though, it is still not completely optimized and sometimes involves complex implementations, weak supervision is a noteworthy technique and with it's recent advances show how it has a promising future.

REFERENCES

[1] S. Zhang, L. He, E. Dragut, and S. Vucetic, "How to invest my time: Lessons from human-in-the-loop entity extraction," in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2019, pp. 2305–2313.

[2] X. Wu, L. Xiao, Y. Sun, J. Zhang, T. Ma, and L. He, "A survey of human-in-the-loop for machine learning," *Future Generation Computer Systems*, vol. 135, pp. 364–381, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167739X22001790

[3] H. Zamani and W. B. Croft, "On the theory of weak supervision for information retrieval," in *Proceedings of the 2018 ACM SIGIR International Conference on Theory of Information Retrieval*, ser. ICTIR '18. New York, NY, USA: Association for Computing Machinery, 2018, p. 147–154. [Online]. Available: https://doi.org/10.1145/3234944.3234968

[4] S. H. Bach, D. Rodriguez, Y. Liu, C. Luo, H. Shao, C. Xia, S. Sen, A. Ratner, B. Hancock, H. Alborzi, R. Kuchhal, C. Ré, and R. Malkin, "Snorkel drybell: A case study in deploying weak supervision at industrial scale," in *Proceedings of the 2019 International Conference on Management of Data*, ser. SIGMOD '19. New York, NY, USA: Association for Computing Machinery, 2019, p. 362–375. [Online]. Available: https://doi.org/10.1145/3299869.3314036

[5] H. Dai, Y. Song, and H. Wang, "Ultra-fine entity typing with weak supervision from a masked language model," *CoRR*, vol. abs/2106.04098, 2021. [Online]. Available: https://arxiv.org/abs/2106.04098

[6] G. Karamanolakis, S. Mukherjee, G. Zheng, and A. H. Awadallah, "Self-training with weak supervision," *CoRR*, vol. abs/2104.05514, 2021. [Online]. Available: https://arxiv.org/abs/2104.05514

[7] R. Hoffmann, C. Zhang, X. Ling, L. Zettlemoyer, and D. S. Weld, "Knowledge-based weak supervision for information extraction of overlapping relations," in *Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies*, 2011, pp. 541–550.

[8] P. Lison, A. Hubin, J. Barnes, and S. Touileb, "Named entity recognition without labelled data: A weak supervision approach," *CoRR*, vol. abs/2004.14723, 2020. [Online]. Available: https://arxiv.org/abs/2004.14723

[9] S. R. Eddy, "What is a hidden markov model?" *Nature biotechnology*, vol. 22, no. 10, pp. 1315–1316, 2004.

[10] P. Lison, J. Barnes, and A. Hubin, "skweak: Weak supervision made easy for nlp," *arXiv preprint arXiv:2104.09683*, 2021.

[11] Y. Vasiliev, *Natural Language Processing with Python and SpaCy: A Practical Introduction*. No Starch Press, 2020.

[12] H. Tseng, P.-C. Chang, G. Andrew, D. Jurafsky, and C. D. Manning, "A conditional random field word segmenter for sighan bakeoff 2005," in *Proceedings of the fourth SIGHAN workshop on Chinese language Processing*, 2005.