From Text-based to Image-based Named Entity Recognition: A Survey

Anonymous ACL submission

Abstract

Named Entity Recognition (NER) is a task to recognize mentions of entities such as person, location, drug, time, biological protein, etc. NER serves as a key component for a number of Natural Language Processing applications including machine translation, entity linking, information retrieval, question answering, etc. Traditional NER is limited to identifying and categorizing entities in text-based data. In recent decades, as Document Image Understanding emerges as a new research area, recognizing entities from image-based documents becomes a new goal in Artificial Intelligence. This paper investigates both text-based and image-based NER through reviewing a series of significant and relevant tasks, datasets, methods, and evaluations, with the goal to present a clear overview of the field. Further, the survey provides a reflection on the field by discussing the challenges and future directions in NER.

1 Introduction

In the 1980s, enabling computers to understand documents has become an imperative goal in Artificial Intelligence. In order to distinguish the elementary building blocks of the information contained within these documents, the large and complex task has soon been decomposed into smaller tasks with the objective to first identify and categorize these information units, the named entities (NEs). The concept therefore first emerged, named entities are linguistic objects following the need expressed above (Nouvel et al., 2016). The term "Named Entity" (NE) was first proposed at the sixth Message Understanding Conference (MUC-6) (Grishman and Sundheim, 1996), with the aim to identify names of organizations, people, and geographic locations in the text, as well as currency, time and percentage expressions. The task of Named Entity Recognition (NER) is to recognize these mentions of entities from text belonging to a predefined typology. In addition to standing alone as an independent tool for Information Extraction, NER systems play key roles in the pipeline of other NLP tasks such as entity linking (Prabhakar Kannan Ravi et al., 2021), information retrieval (Guo et al., 2009), machine translation (Babych and Hartley, 2003), question answering (Mollá et al., 2006), and text summarization (Aone, 1999), etc.

Conventional NER tasks solely based on text input can be broadly divided into two sub-categories: generic NER in the general context (non-specialist language) to recognize person, location, organization, number, date, and time, etc and domain-specific NER to identify entities in specialized fields such as medicine and law, e.g. drug names, enzymes, jurisdictions, and legal institutions, etc.

Over decades of development of traditional NER on pure text input for document understanding, a more ambitious goal of extracting entities directly from image-based documents has recently emerged as a new research problem. In contrast to text-based NER, visual entity recognition, alternatively referred to as image-based entity recognition is a downstream task of Document AI. It defines a new form of NER to extract entities in a two-dimensional structured information space i.e. a document image instead of linear text sequences. With the increasing need in the business environment to process large amounts of digital-born, image-based business documents such as invoices and receipts, one major application domain of visual NER is to extract business objects such as invoice numbers, IBANs for bank accounts, which often appear as key-value pairs and convey critical and confidential information for businesses. Due to the complexity of the 2-D layout of these documents (e.g. the value can appear to the right or bottom of the key in these documents), a challenge is posed to NER in this problem setup.

While many surveys on NER in the general context had been conducted over the past few

years, this paper conducts its investigation from text-based to image-based NER, with the focus on generic NER for text-based entity recognition.

The remainder of this paper is organized as follows. The following two sections review a series of NER datasets and models that are significant and relevant to illustrate the broad view of the field. Section 4 briefly discusses some of the evaluation paradigms proposed through the major forums. Section 5 presents a reflection on the field of NER research in general by discussing the challenges and opportunities. Finally, section 6 concludes with final remarks. Each section starts its discussion from conventional NER on linear textual input in the general domain and extends to image-based NER.

2 NER Datasets

Over the past few decades, the development of NER has gone through stages, evolving from simple rule-based heuristics to statistical methods. In more recent practices, data-driven methods have proved more successful in the task and have been widely studied and applied. While the next section discusses these methods for NER in detail, this section first describes the annotated NER datasets developed over the recent years, which are critical for the training and evaluation of these statistical models for Named Entity Recognition. This section first reviews various text-based generic NER datasets historically and lists some of the major image-based datasets for visual entity recognition.

2.1 Text-based NER Datasets

Throughout the history of the development of NER, many of the NER datasets were created and distributed through NER shared tasks. After the first NER shared-task (Grishman and Sundheim, 1996), CoNLL-2002 (Tjong Kim Sang, 2002) and CoNLL-2003 (Tjong Kim Sang and De Meulder, 2003) distributed datasets created from newswire articles in four different languages (Spanish and Dutch in CoNLL-2002, English and German in CoNLL-2003) and focused on four types of entities: PER (person), LOC (location), ORG (organization) and MISC (miscellaneous entities that do not belong to the previous three groups). Later, various NER tasks have been organised for other languages such as Indian languages (Gali et al., 2008), Arabic (Shaalan, 2014), German (Benikova et al., 2014), and Slavic languages (Piskorski et al., 2017). Outside of shared tasks, various generic NER datasets in different languages have been created over the years. Peng and Dredze (2015), for instance, proposed a dataset containing messages from Chinese Social Media (Weibo).[1] The dataset is annotated with four types of entities: person, organization, location, and geo-political entity. Additionally, the OntoNotes (Hovy et al., 2006) project was initially launched to annotate a large corpus sourced from a variety of genres: broadcast, news, weblogs, USENET newsgroups, talk shows, and conversational telephone speech with structural information (syntax and predicate argument structure) and shallow semantics (word sense linked to an ontology and coreference). The project released 5 versions, with texts annotated with 18 entity types.

Most of the generic datasets share a typology similar to that in CoNLL-2003 (i.e. person, location, organization, other), with several exceptions such as Gali et al. (2008), which differentiates more entity types: person, designation, temporal expressions, abbreviations, object number, brand, etc. The source of texts varies vastly across datasets, with the majority collected from news articles, social media platforms, and Wikipedia.

Amongst these generic NER datasets developed over the years, CoNLL-2003 and OntoNotes have been most commonly used as benchmarks to report the performance of a new NER system.

Besides generic NER datasets, a variety of domain-specific NER datasets in specialized fields such as biomedicine and material science have been developed over the years. As mentioned in the beginning section, this survey does not focus on domain-specific NER, thus here list only a few such datasets. For instance, i2b2-2010 (Uzuner et al., 2011) is a clinical dataset that focuses on the identification of problem, treatment, and test from patient reports. MaSciP (Mysore et al., 2019) is material science dataset that contains synthesis procedures annotated with synthesis operations and their typed arguments (e.g., Material, Synthesis-Apparatus, etc.). Due to the requirement of a large amount of annotated training data for deep learning models and the unavailability of such datasets, efforts on data augmentation for NER datasets have been made. Dai and Adel (2020), for instance, compared a number of augmentation techniques for NER[2] and proved the effectiveness through ex-

[1] http://www.weibo.com/

[2] Label-wise token replacement (LwTR), synonym replace-

periments.

Although a large number of generic and domain-specific NER datasets have been proposed over the years, there is still the need to create NER datasets on occasions. Here briefly illustrates the possibility to create a NER dataset for special purposes from the ground up. With various APIs such as Twitter developer portal[3] and Google BigQuery,[4] it is possible to collect raw text data in generic or specified domain by setting restrictions in the data collection process. Afterward, raw text data can be annotated using annotation tools such as Label Studio,[5] which provides a simple and intuitive GUI for users to annotate named entities and export labeled data in a parsable format.[6]

2.2 Image-based Entity Extraction Datasets

In contrast to text-based NER datasets, image-based NER datasets compromise document images and their corresponding OCR annotations, which provide not only text information, but also visual and layout information. Table 1 lists some of the major benchmark datasets for visual entity recognition. Amongst the benchmarks listed, FUNSD (Jaume et al., 2019), SROIE (Huang et al., 2019), CORD (Park et al., 2019) and Kleister (Stanisławek et al., 2021) have been most commonly used to evaluate the performance of new developed systems.

The FUNSD dataset was originally developed for the form understanding task. The dataset contains 199 noisy scanned documents with 9707 semantic entities labeled "question", "answer", "header" or "other". SROIE is a dataset of receipts with entities annotated with company, date, address, or total. Each receipt is organized as a list of text lines with bounding boxes. Similarly, CORD is also a receipt key information extraction dataset with 30 semantic labels defined under 4 categories. The dataset provides both bounding boxes and OCR annotations. Kleister consists of two datasets: Kleister NDA and Kleister Charity, with Kleister NDA more commonly used as a benchmark to evaluate new systems. The Kleister NDA dataset contains legal NDA (Non-disclosure Agreement) documents with party (ORG/PER), jurisdiction (LOCATION), effective_date (DATE), and term (DURATION) entities labeled.

Dataset	Language
FUNSD (Jaume et al., 2019)	En
SROIE (Huang et al., 2019)	En
CORD (Park et al., 2019)	En
EATEN (Guo et al., 2019)	Zh
EPHOIE (Wang et al., 2021)	Zh
Deepform (Stray and Svetlichnaya, 2020)	En
Kleister (Stanisławek et al., 2021)	En
XFUND (Xu et al., 2021)	Zh/Ja/Es/ Fr/It/De/Pt

Table 1: Benchmark datasets for visual entity recognition.

In addition to the datasets listed above, datasets such as IIT-CDIP Test Collection (Lewis et al., 2006), although not specifically for the visual entity recognition downstream task, are relevant to the problem at hand. As such datasets are often used to pretrain general-purpose Document AI models, which are to be fine-tuned for various downstream tasks including image-based entity extraction. The next section will describe these pretraining methods for visual entity recognition in more detail.

3 Approaches to NER

This section discusses the approaches developed over the years to both text-based and image-based NER.

3.1 Approaches to Text-based NER

With years of development of text-based NER, a large number of approaches had been proposed, evolving from rule-based solutions to data-driven solutions including both supervised and unsupervised methods.

3.1.1 Rule-based Techniques

In the early stage of the development of NER, approaches were mainly dependent on rules written by humans. These rules are often based on syntactic-lexical patterns and domain-specific knowledge. Early works such as LaSIE-II (Humphreys et al., 1998), NetOwl (Krupka and IsoQuest, 2005), Facile (Black et al., 1998), SRA (Aone et al., 1998), FASTUS (Appelt et al., 1995), and LTG (Mikheev et al., 1999) were developed according to hand-crafted lexical, semantic and syntactic rules to identify and classify entities. Rule-based systems also perform well when it comes to NER in a specialised area based on the features de-

ment (SR), mention replacement (MR), shuffle within segments (SiS).

[3] https://developer.twitter.com

[4] https://cloud.google.com/bigquery

[5] https://labelstud.io/

[6] Code repository for an example workflow is not anonymous, shall be relaesed upon acception or rejection of the paper.

veloped according to domain-specific gazetteers.[7] Quimbaya et al. (2016), for instance, proposed a dictionary-based approach for NER in electronic health records.

As much as rule-based NER systems show good performance when the lexicon is exhaustive, these systems have the problem of being highly domain-specific and non-portable. Additionally, designing rules for the system often require human expertise in linguistics as well as knowledge in specific domains, which makes the systems expensive to develop and maintain. In general, rule-based NER systems show high precision and low recall (more False Negatives than False Positives), due to the closed-set lexicon and domain-specific rules.

3.1.2 Data-Driven Methods

Data-driven systems have several advantages over rule-based ones. This section reviews various learning-based methods for the NER task. The following discusses both unsupervised and supervised methods, with feature-based and neural supervised learning approaches in separate discussions.

Unsupervised Methods

Unsupervised learning methods require data neither classified nor labeled. The goal is to generate a model that grasps the structural and distributional features of the unlabelled training data and to make predictions on future unseen data. Unsupervised learning approaches are applied in NER, typically through clustering with associated resources/knowledge bases (e.g., WordNet (Miller, 1995)) (Nadeau and Sekine, 2007).

While a number of unsupervised methods for NER have been developed over the years, here lists only a few. Collins and Singer (1999) used unlabelled data and 7 “seed" rules such as orthography (e.g., capitalization), context (of the entity), words composed of named entities, etc. to infer mentions of named entities. Alfonseca and Manandhar (2002) takes named entity types from the WordNet and labels each input word with an NE type. The method is to assign a topic signature to each WordNet synset according to frequent co-occurrence in a large corpus. Then each word is labeled with the most similar type signature to its context in the document it belongs to. Etzioni et al. (2005) uses Pointwise Mutual Information and Information Retrieval (PMI-IR) to measure the association between an entity and an entity type based on the theory that expressions with high PMI-IR are more likely to co-occur. Shinyama and Sekine (2004) proposed a method to identify named entities based on an observation that a named entity is strongly correlated with appearing punctually and simultaneously in multiple news articles. The approach is especially effective to detect named entities with high rarity.

Feature-based Supervised Learning Methods

Compared to unsupervised learning methods for NER, supervised methods are less dependent on “rules/theory/observations/knowledge". Before the popularity of various neural models for NLP tasks, early works on data-driven NER combine text features and classic machine learning models to enable recognizing similar patterns in unseen data.

In feature-based learning, feature engineering is critical for NER systems based on classic models. These feature representations include word-level features (e.g. POS tag, case, and morphological features) (Zhou and Su, 2002), list lookup features (e.g. DBpedia and Wikipedia gazetteer) (Mikheev, 1999) and document-level features (e.g. local syntax and multiple occurrences) (Ravin and Wacholder, 1997).[8]

The feature vector representations will then be fed into models for training and inference. In feature-based supervised learning, the NER task is formalized either as a sequential labeling problem or a multi-class classification problem. Thus the typical models include sequence-to-sequence models: contextual models such as Hidden Markov Models (HMMs) (Eddy, 1996), multiple feature models such as Maximum Entropy Models (Kapur, 1989), Conditional Random Fields (CRFs) (Lafferty et al., 2001), and classifiers such as Decision Trees (Quinlan, 1986) and Support Vector Machines (SVMs) (Cortes and Vapnik, 1995).

Bikel et al. (1998), for instance, first employed HMMs for the NER task. While systems proposed in Bender et al. (2003), Chieu and Ng (2002), Curran and Clark (2003) used MaxEnt Models to identify and categorize entities. In order to take better account of context information, McCallum and Li (2003) used CRFs to perform the task. It is worth noting that CRFs are widely used in named entity recognition and show advanced performance for this task according to metrics reported in several experiments from different works (Nouvel et al.,

[7] A gazetteer consists of a set of lists containing names of entities such as cities, organisations, days of the week, etc.

[8] Nadeau and Sekine (2007) discusses feature engineering for NER in more details.

2016). As for classifier-based systems, Szarvas et al. (2006) trained multiple decision tree models using different sets of features to perform the task through a majority voting system. McNamee and Mayfield (2002) used orthography and punctuation features to train SVM classifiers. Each classifier makes a binary decision whether the current token belongs to one of the eight classes, i.e., B (Beginning) and I (Inside) for each of the four NE tags.

Neural Network Methods

In contrast to classic models, neural network models are less dependent on feature engineering and more reliant on large amount of training data and computing power. Since Collobert and Weston (2008) proposed one of the first neural network methods for NER, many more works followed. Neural architectures for NER have been classified into four categories by previous surveys according to word representation: Word level, Character level, Word+Character level, and Word+Character+Affix level models. In word level architectures, each word is represented by its embedding, and the input to the Recurrent Neural Networks (RNNs) is a sentence represented by a sequence of words. Huang et al. (2015) presented a word level neural architecture with LSTM layers and a CRF layer on top. The model showed promising performance on the benchmark dataset CoNLL-2003. While Kuru et al. (2016) proposed CharNER, which is a character level RNN model for multilingual NER on 7 languages. In this model, the sentence is represented as a sequence of characters to be fed into the RNN model. The tags predicted for each character were converted to word tags using Viterbi decoder (Forney, 1973). CharNER also presented good performance on the CoNLL datasets (2002 & 2003). Character+Word level architectures combine both character and word embeddings in two different ways: 1) words are represented as a combination of a word embedding and a convolution over the word characters, followed by a Bi-LSTM layer over the word representations of a sentence, and finally, a softmax or CRF layer over the Bi-LSTM to generate labels (implemented in Ma and Hovy (2016), Chiu and Nichols (2016), etc.) 2) word embeddings are concatenated with (Bi-directional) LSTMs over word characters, passed to sentence-level Bi-LSTM and predicting the final tags using a final softmax or CRF layer (Lample et al., 2016). Character+Word level models in general show very strong performance on benchmark datasets. Later, Yadav et al. (2018) introduced affixes to the Word+Character models in Lample et al. (2016). Affix is one of the most successful features from feature engineering. The models were extended to learn affix embeddings alongside the word embeddings and character RNNs. The Word+Character+Affix model achieved even better performance on the CoNLL datasets in four languages.

In addition to the traditional neural network models, transformer-based models proposed more recently have presented advanced performance on various NLP tasks including NER. Transformer (Vaswani et al., 2017) utilizes stacked self-attention and point-wise, fully connected layers to build basic blocks for encoder and decoder, disregarding recurrence and convolutions completely. Transformer-based methods such as BERT (Devlin et al., 2018), GPT (Radford et al., 2018) and ELMo (Peters et al., 2018) show even stronger performance on NER compared to non-transformer neural models. More recent works such as SpanBERT (Joshi et al., 2020) and LUKE (Yamada et al., 2020) extended BERT in order to see better performance on entity-related tasks including NER. Both LUKE and SpanBERT outperformed general-purposed BERT baselines on several NER benchmarks.

In addition to the typical neural methods reviewed above, it is worth noting that in more recent years, applied deep learning techniques such as deep active learning (Shen et al., 2017), reinforcement learning (Narasimhan et al., 2016), adversarial learning (Cao et al., 2019) have also been introduced to the NER task. A more thorough review on various more deep learning techniques can be found in Li et al. (2020).

3.2 Approaches to Image-based Entity Extraction

Various approaches proposed to image-based NER treat the task as different problems since document images convey both textual and visual information.

A number of recent research works consider visual entity recognition as a Computer Vision problem, and perform the task through semantic segmentation or text box detection (Cui et al., 2021). Given the significance of the text information contained in the document images, typical frameworks represent these document images as a pixel grid with text features added to the visual feature map. The approaches to represent text infor-

mation evolved from character-level to word-level, and then to context-level. Chargrid (Katti et al., 2018) uses a convolution-based encoder-decoder network to fuse text information into images using one-hot encoded characters. VisualWordGrid (Kerroumi et al., 2021) replaced character-level text information with word-level representation in Chargrid, and improved model performance. In order to take better account of contextual information, BERTgrid (Denk and Reisswig, 2019) uses BERT to obtain context representation and increased recognition accuracy. ViBERTgrid (Lin et al., 2021) further built on BERTgrid using image features from the CNN model and presented better model performance.

While some other research consider the task as a special natural language understanding task. Majumder et al. (2020), for instance, generates extraction candidates based on the knowledge of the types of the target fields and uses a neural network architecture to learn a dense representation of each candidate based on its context. The approach proved useful in solving the extraction task.

On top of the two ways discussed above to formalise the task, unstructured visually-rich documents can be naturally well represented by Graph Neural Network (GNN) since they are often composed of multiple adjacent text fragments: the text fragments can be abstracted as nodes, with the relationship between the text fragments modeled as edges. This enabled a number of research works for visual entity recognition based on GNNs. In Hwang et al. (2020), the document is modeled as a directed graph, which enables information extraction through dependency analysis. Many works proposed to use GNN-based methods for visual extraction tasks, these include Cheng et al. (2020), Riba et al. (2019), Wei et al. (2020).

In addition to the methods discussed above, general-purpose multi-modal pretraining approaches for Document Understanding can also be applied to the downstream image-based entity extraction task. This type of approach involves two stages of learning: 1) pretraining model on a large-scale Document AI dataset for general purpose 2) fine-tuning the pretrained model on a task-specific dataset for each downstream task such as visual entity extraction, document image classification, etc. LayoutLM (Xu et al., 2020b), which is a pioneer work using this approach, uses text, image, and layout information to jointly pretrain an extended BERT model in order to make better use of all relevant information conveyed in document images for various Document AI tasks. While LayoutLM achieved good performance on the downstream tasks already, subsequent works such as LayoutLMv2 (Xu et al., 2020a), LayoutLMv3 (Huang et al., 2022), LayoutXLM (Xu et al., 2021) and DocFormer (Appalaraju et al., 2021) followed and further improved from LayoutLM by showing stronger performance and enabling multilinguality on various downstream tasks including visual NER. These subsequent works mostly built themselves on the basis of LayoutLM through redesigning the architecture and pre-training objectives. LayoutLM has gradually become the basic unit for building more complex algorithms.

4 NER Evaluations

Both text-based and image-based NER are often formalised as a sequential labeling problem. The metrics used to evaluate a NER system are typically: precision, recall, and F-score.

It is worth noting that in text-based NER, there have been two typical evaluations. CoNLL first introduced **Exact-Match Evaluation**, in which case a named entity is considered correctly recognized only if its both boundaries and type match ground truth. While for **Relaxed-Match Evaluation** defined in MUC-6, a correct type is credited if an entity is assigned its correct type regardless of its boundaries as long as there is an overlap with ground truth boundaries; a correct boundary is credited regardless of an entity's type assignment. Most works employ the exact-match evaluation to measure their model performance since relaxed-match evaluation is complex and causes difficulty to error analysis.

5 Reflection: Challenges and Opportunities in NER

In addition to the issues briefly mentioned in the previous sections, this survey noticed several challenges and opportunities in the field of NER.

Dataset Quality

As noted in a number of previous examinations on the data quality of NER benchmarks, the issue of annotation inconsistency has been spotted in datasets such as CoNLL-2003 (Tjong Kim Sang and De Meulder, 2003), MUC-7 (Chinchor and Robinson, 1997), and ACE (Doddington et al.,

2004). For instance, “Empire State Building” is labeled as Location in the ACE dataset, while the boundary is set at "Empire State" in CoNLL-2003. Another example of inconsistency would be that “Baltimore” in the sentence “Baltimore defeated the Yankees”, is labeled as Location in MUC-7 while in CoNLL-2003 as Organization. In addition to inconsistent annotation, the “generalizability" of NER benchmark results also brings concerns. Most of the newly proposed systems are trained and evaluated on these benchmark datasets, which may not be the best ways to reflect the general system performance in cases different from benchmarks. To illustrate this, CoNLL-2003, for instance, contains texts sourced from news article. A system that performs well on this dataset may not have comparable performance on another dataset with different data properties, for example, data from social media posts. With various issues existing with the NER benchmark datasets and the models developed based on them, future work can can consider taking a further step by examining the validity and reliability of these benchmark datasets. Dataset validity and reliability is discussed in Riezler and Hagmann (2021). In order to inspect the validity of a dataset, there are several model-based and descriptive tests such as dataset bias test, which can be used to examine whether a model learns superficial patterns in the data to perform well on training data, but does not generalize well and performs poorly on out-of-domain test data (Clark et al., 2019). While reliability tests examine how consistent is a performance evaluation if replicated under variations of meta-parameters (or varying data properties). With efforts in NER mostly put on creating new datasets and developing new systems with SOTA results on benchmarks, an examination on the reliability and validity of the existing benchmark or non-benchmark datasets, as well as the models evaluated on them would be a reasonable and meaningful next step.

Domain-specific & Low-resource Language

Another issue with NER datasets is that data for specialised domains and low-resource languages is far away from sufficient. Most NER datasets have been developed for English. Especially for image-based NER, as a relatively new area, there has been few work on datasets for other languages. Future work should consider devoting more efforts on building datasets on low-resource languages. In addition to creating new datasets, developing further data augmentation techniques can also help to resolve the issue. Furthermore, enabling better unsupervised, semi-/self-supervised, and multilingual systems is always an important future direction for NER as a solution to the problems discussed above.

Visual-specific Issues

Some of the most frequently discussed challenges and future directions in visual NER include few-shot and zero-shot learning, multi-page/cross-page problems, and uneven quality of training data. In addition to these, multi-modal pretraining, which is considered currrently the most effective approach, relies on large scale of pretraining data. These data are often generated using OCR tools to obtain text data from image-based documents for joint training. Thus, the accuracy of the employed OCR tools is a potential concern that requires attention. Additionally, the current pretrained models can benefit from more data for pretraining, which would also be based on OCR. Therefore, future research should pay more attention to the accuracy and reliability of the OCR tools for pretraining data generation, meanwhile producing more data to scale up training in order to drive SOTA results further for image-based NER.

6 Conclusions

This survey aims to investigate Named Entity Recognition, covering both text-based and image-based NER. In order to provide a good overview of the field, this paper reviews a wide range of datasets and approaches including both rule-based and learning-based methods. For soundness, this paper also briefly discusses NER evaluations. Finally, the investigation ends with a reflection on the research field by discussing the challenges and opportunities in NER. The review on NER works in this survey is not exhaustive, the intend is to illustrate a set of selected works considered most significant and relevant to describe the “lay of the land". While longer survey papers may list a number more datasets, methods and evaluation systems, and provide more details by further explaining the algorithms, and comparing the evaluation metrics of the methods discussed. This paper provides a brief overview in few pages through a preliminary survey.

References

Enrique Alfonseca and Suresh Manandhar. 2002. An unsupervised method for general named entity

recognition and automated concept discovery. In *Proceedings of the 1st international conference on general WordNet, Mysore, India*, pages 34–43.

Chinatsu Aone. 1999. A trainable summarizer with knowledge acquired from robust nlp techniques. *Advances in automatic text summarization*, pages 71–80.

Chinatsu Aone, Lauren Halverson, Tom Hampton, and Mila Ramos-Santacruz. 1998. Sra: Description of the ie2 system used for muc-7. In *Seventh Message Understanding Conference (MUC-7): Proceedings of a Conference Held in Fairfax, Virginia, April 29-May 1, 1998*.

Srikar Appalaraju, Bhavan Jasani, Bhargava Urala Kota, Yusheng Xie, and R Manmatha. 2021. Docformer: End-to-end transformer for document understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 993–1003.

Douglas Appelt, Jerry R Hobbs, John Bear, David Israel, Megumi Kameyama, Andrew Kehler, David Martin, Karen Myers, and Mabry Tyson. 1995. Sri international fastus systemmuc-6 test results and analysis. In *Sixth Message Understanding Conference (MUC-6): Proceedings of a Conference Held in Columbia, Maryland, November 6-8, 1995*.

Bogdan Babych and Anthony Hartley. 2003. Improving machine translation quality with automatic named entity recognition. In *Proceedings of the 7th International EAMT workshop on MT and other language technology tools, Improving MT through other language technology tools, Resource and tools for building MT at EACL 2003*.

Oliver Bender, Franz Josef Och, and Hermann Ney. 2003. Maximum entropy models for named entity recognition. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003*, pages 148–151.

Darina Benikova, Chris Biemann, and Marc Reznicek. 2014. Nosta-d named entity annotation for german: Guidelines and dataset. In *LREC*, pages 2524–2531.

Daniel M Bikel, Scott Miller, Richard Schwartz, and Ralph Weischedel. 1998. Nymble: a high-performance learning name-finder. *arXiv preprint cmp-lg/9803003*.

William J Black, Fabio Rinaldi, and David Mowatt. 1998. Facile: Description of the ne system used for muc-7. In *Seventh Message Understanding Conference (MUC-7): Proceedings of a Conference Held in Fairfax, Virginia, April 29-May 1, 1998*.

Yixin Cao, Zikun Hu, Tat-seng Chua, Zhiyuan Liu, and Heng Ji. 2019. Low-resource name tagging learned with weakly labeled data. *arXiv preprint arXiv:1908.09659*.

Mengli Cheng, Minghui Qiu, Xing Shi, Jun Huang, and Wei Lin. 2020. One-shot text field labeling using attention and belief propagation for structure information extraction. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 340–348.

Hai Leong Chieu and Hwee Tou Ng. 2002. Named entity recognition: a maximum entropy approach using global information. In *COLING 2002: The 19th International Conference on Computational Linguistics*.

Nancy Chinchor and Patricia Robinson. 1997. Muc-7 named entity task definition. In *Proceedings of the 7th Conference on Message Understanding*, volume 29, pages 1–21.

Jason PC Chiu and Eric Nichols. 2016. Named entity recognition with bidirectional lstm-cnns. *Transactions of the association for computational linguistics*, 4:357–370.

Christopher Clark, Mark Yatskar, and Luke Zettlemoyer. 2019. Don't take the easy way out: Ensemble based methods for avoiding known dataset biases. *arXiv preprint arXiv:1909.03683*.

Michael Collins and Yoram Singer. 1999. Unsupervised models for named entity classification. In *1999 Joint SIGDAT conference on empirical methods in natural language processing and very large corpora*.

Ronan Collobert and Jason Weston. 2008. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th international conference on Machine learning*, pages 160–167.

Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine learning*, 20(3):273–297.

Lei Cui, Yiheng Xu, Tengchao Lv, and Furu Wei. 2021. Document ai: Benchmarks, models and applications. *arXiv preprint arXiv:2111.08609*.

James R Curran and Stephen Clark. 2003. Language independent ner using a maximum entropy tagger. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003*, pages 164–167.

Xiang Dai and Heike Adel. 2020. An analysis of simple data augmentation for named entity recognition. *arXiv preprint arXiv:2010.11683*.

Timo I Denk and Christian Reisswig. 2019. Bertgrid: Contextualized embedding for 2d document representation and understanding. *arXiv preprint arXiv:1909.04948*.

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*.

George R Doddington, Alexis Mitchell, Mark A Przybocki, Lance A Ramshaw, Stephanie M Strassel, and Ralph M Weischedel. 2004. The automatic content extraction (ace) program-tasks, data, and evaluation. In *Lrec*, volume 2, pages 837–840. Lisbon.

Sean R Eddy. 1996. Hidden markov models. *Current opinion in structural biology*, 6(3):361–365.

Oren Etzioni, Michael Cafarella, Doug Downey, Ana-Maria Popescu, Tal Shaked, Stephen Soderland, Daniel S Weld, and Alexander Yates. 2005. Unsupervised named-entity extraction from the web: An experimental study. *Artificial intelligence*, 165(1):91–134.

G David Forney. 1973. The viterbi algorithm. *Proceedings of the IEEE*, 61(3):268–278.

Karthik Gali, Harshit Surana, Ashwini Vaidya, Praneeth M Shishtla, and Dipti Misra Sharma. 2008. Aggregating machine learning and rule based heuristics for named entity recognition. In *Proceedings of the IJCNLP-08 Workshop on Named Entity Recognition for South and South East Asian Languages*.

Ralph Grishman and Beth M Sundheim. 1996. Message understanding conference-6: A brief history. In *COLING 1996 Volume 1: The 16th International Conference on Computational Linguistics*.

He Guo, Xiameng Qin, Jiaming Liu, Junyu Han, Jingtuo Liu, and Errui Ding. 2019. Eaten: Entity-aware attention for single shot visual text extraction. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 254–259. IEEE.

Jiafeng Guo, Gu Xu, Xueqi Cheng, and Hang Li. 2009. Named entity recognition in query. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pages 267–274.

Eduard Hovy, Mitch Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. 2006. Ontonotes: the 90% solution. In *Proceedings of the human language technology conference of the NAACL, Companion Volume: Short Papers*, pages 57–60.

Yupan Huang, Tengchao Lv, Lei Cui, Yutong Lu, and Furu Wei. 2022. Layoutlmv3: Pre-training for document ai with unified text and image masking. *arXiv preprint arXiv:2204.08387*.

Zheng Huang, Kai Chen, Jianhua He, Xiang Bai, Dimosthenis Karatzas, Shijian Lu, and CV Jawahar. 2019. Icdar2019 competition on scanned receipt ocr and information extraction. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 1516–1520. IEEE.

Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional lstm-crf models for sequence tagging. *arXiv preprint arXiv:1508.01991*.

Kevin Humphreys, Robert Gaizauskas, Saliha Azzam, Christian Huyck, Brian Mitchell, Hamish Cunningham, and Yorick Wilks. 1998. University of sheffield: Description of the lasie-ii system as used for muc-7. In *Seventh Message Understanding Conference (MUC-7): Proceedings of a Conference Held in Fairfax, Virginia, April 29-May 1, 1998*.

Wonseok Hwang, Jinyeong Yim, Seunghyun Park, Sohee Yang, and Minjoon Seo. 2020. Spatial dependency parsing for semi-structured document information extraction. *arXiv preprint arXiv:2005.00642*.

Guillaume Jaume, Hazim Kemal Ekenel, and Jean-Philippe Thiran. 2019. Funsd: A dataset for form understanding in noisy scanned documents. In *2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)*, volume 2, pages 1–6. IEEE.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S Weld, Luke Zettlemoyer, and Omer Levy. 2020. Spanbert: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, 8:64–77.

Jagat Narain Kapur. 1989. *Maximum-entropy models in science and engineering*. John Wiley & Sons.

Anoop Raveendra Katti, Christian Reisswig, Cordula Guder, Sebastian Brarda, Steffen Bickel, Johannes Höhne, and Jean Baptiste Faddoul. 2018. Chargrid: Towards understanding 2d documents. *arXiv preprint arXiv:1809.08799*.

Mohamed Kerroumi, Othmane Sayem, and Aymen Shabou. 2021. Visualwordgrid: Information extraction from scanned documents using a multimodal approach. In *International Conference on Document Analysis and Recognition*, pages 389–402. Springer.

GR Krupka and K IsoQuest. 2005. Description of the nerowl extractor system as used for muc-7. In *Proc. 7th Message Understanding Conf*, pages 21–28.

Onur Kuru, Ozan Arkan Can, and Deniz Yuret. 2016. Charner: Character-level named entity recognition. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 911–921.

John Lafferty, Andrew McCallum, and Fernando CN Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. *arXiv preprint arXiv:1603.01360*.

David Lewis, Gady Agam, Shlomo Argamon, Ophir Frieder, David Grossman, and Jefferson Heard. 2006. Building a test collection for complex document information processing. In *Proceedings of*

the 29th annual international ACM SIGIR conference on Research and development in information retrieval, pages 665–666.

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*, 34(1):50–70.

Weihong Lin, Qifang Gao, Lei Sun, Zhuoyao Zhong, Kai Hu, Qin Ren, and Qiang Huo. 2021. Vibertgrid: a jointly trained multi-modal 2d document representation for key information extraction from documents. In *International Conference on Document Analysis and Recognition*, pages 548–563. Springer.

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

Bodhisattwa Prasad Majumder, Navneet Potti, Sandeep Tata, James Bradley Wendt, Qi Zhao, and Marc Najork. 2020. Representation learning for information extraction from form-like documents. In *proceedings of the 58th annual meeting of the Association for Computational Linguistics*, pages 6495–6504.

Andrew McCallum and Wei Li. 2003. Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons.

Paul McNamee and James Mayfield. 2002. Entity extraction without language-specific resources. In *COLING-02: The 6th Conference on Natural Language Learning 2002 (CoNLL-2002)*.

Andrei Mikheev. 1999. A knowledge-free method for capitalized word disambiguation. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pages 159–166.

Andrei Mikheev, Marc Moens, and Claire Grover. 1999. Named entity recognition without gazetteers. In *Ninth Conference of the European Chapter of the Association for Computational Linguistics*, pages 1–8.

George A Miller. 1995. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

Diego Mollá, Menno Van Zaanen, and Daniel Smith. 2006. Named entity recognition for question answering. In *Proceedings of the Australasian language technology workshop 2006*, pages 51–58.

Sheshera Mysore, Zach Jensen, Edward Kim, Kevin Huang, Haw-Shiuan Chang, Emma Strubell, Jeffrey Flanigan, Andrew McCallum, and Elsa Olivetti. 2019. The materials science procedural text corpus: Annotating materials synthesis procedures with shallow semantic structures. *arXiv preprint arXiv:1905.06939*.

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

Karthik Narasimhan, Adam Yala, and Regina Barzilay. 2016. Improving information extraction by acquiring external evidence with reinforcement learning. *arXiv preprint arXiv:1603.07954*.

Damien Nouvel, Maud Ehrmann, and Sophie Rosset. 2016. *Named entities for computational linguistics*. John Wiley & Sons.

Seunghyun Park, Seung Shin, Bado Lee, Junyeop Lee, Jaeheung Surh, Minjoon Seo, and Hwalsuk Lee. 2019. Cord: a consolidated receipt dataset for post-ocr parsing. In *Workshop on Document Intelligence at NeurIPS 2019*.

Nanyun Peng and Mark Dredze. 2015. Named entity recognition for chinese social media with jointly trained embeddings. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 548–554.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Jakub Piskorski, Lidia Pivovarova, Jan Šnajder, Josef Steinberger, and Roman Yangarber. 2017. The first cross-lingual challenge on recognition, normalization and matching of named entities in slavic languages. In *Proceedings of the 6th Workshop on Balto-Slavic Natural Language Processing*. The Association for Computational Linguistics.

Manoj Prabhakar Kannan Ravi, Kuldeep Singh, Isaiah Onando Mulang, Saeedeh Shekarpour, Johannes Hoffart, and Jens Lehmann. 2021. Cholan: A modular approach for neural entity linking on wikipedia and wikidata. *arXiv e-prints*, pages arXiv–2101.

Alexandra Pomares Quimbaya, Alejandro Sierra Múnera, Rafael Andrés González Rivera, Julián Camilo Daza Rodríguez, Oscar Mauricio Muñoz Velandia, Angel Alberto Garcia Peña, and Cyril Labbé. 2016. Named entity recognition over electronic health records through a combined dictionary-based approach. *Procedia Computer Science*, 100:55–61.

JR Quinlan. 1986. Induction of decision trees. mach. learn.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. Improving language understanding by generative pre-training.

Yael Ravin and Nina Wacholder. 1997. *Extracting names from natural-language text*. Citeseer.

Pau Riba, Anjan Dutta, Lutz Goldmann, Alicia Fornés, Oriol Ramos, and Josep Lladós. 2019. Table detection in invoice documents by graph neural networks. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 122–127. IEEE.

Stefan Riezler and Michael Hagmann. 2021. Validity, reliability, and significance: Empirical methods for nlp and data science. *Synthesis Lectures on Human Language Technologies*, 14(6):1–165.

Khaled Shaalan. 2014. A survey of arabic named entity recognition and classification. *Computational Linguistics*, 40(2):469–510.

Yanyao Shen, Hyokun Yun, Zachary C. Lipton, Yakov Kronrod, and Animashree Anandkumar. 2017. Deep active learning for named entity recognition. *CoRR*, abs/1707.05928.

Yusuke Shinyama and Satoshi Sekine. 2004. Named entity discovery using comparable news articles. In *COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics*, pages 848–853.

Tomasz Stanisławek, Filip Graliński, Anna Wróblewska, Dawid Lipiński, Agnieszka Kaliska, Paulina Rosalska, Bartosz Topolski, and Przemysław Biecek. 2021. Kleister: key information extraction datasets involving long documents with complex layouts. In *International Conference on Document Analysis and Recognition*, pages 564–579. Springer.

Jonathan Stray and Stacey Svetlichnaya. 2020. Project deepform: Extract information from documents, 2020.

György Szarvas, Richárd Farkas, and András Kocsor. 2006. A multilingual named entity recognition system using boosting and c4. 5 decision tree learning algorithms. In *International Conference on Discovery Science*, pages 267–278. Springer.

Erik F. Tjong Kim Sang. 2002. Introduction to the CoNLL-2002 shared task: Language-independent named entity recognition. In *COLING-02: The 6th Conference on Natural Language Learning 2002 (CoNLL-2002)*.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 142–147.

Özlem Uzuner, Brett R South, Shuying Shen, and Scott L DuVall. 2011. 2010 i2b2/va challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association*, 18(5):552–556.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Jiapeng Wang, Chongyu Liu, Lianwen Jin, Guozhi Tang, Jiaxin Zhang, Shuaitao Zhang, Qianying Wang, Yaqiang Wu, and Mingxiang Cai. 2021. Towards robust visual information extraction in real world: new dataset and novel solution. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 2738–2745.

Mengxi Wei, Yifan He, and Qiong Zhang. 2020. Robust layout-aware ie for visually rich documents with pre-trained language models. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2367–2376.

Yang Xu, Yiheng Xu, Tengchao Lv, Lei Cui, Furu Wei, Guoxin Wang, Yijuan Lu, Dinei Florencio, Cha Zhang, Wanxiang Che, et al. 2020a. Layoutlmv2: Multi-modal pre-training for visually-rich document understanding. *arXiv preprint arXiv:2012.14740*.

Yiheng Xu, Minghao Li, Lei Cui, Shaohan Huang, Furu Wei, and Ming Zhou. 2020b. Layoutlm: Pre-training of text and layout for document image understanding. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1192–1200.

Yiheng Xu, Tengchao Lv, Lei Cui, Guoxin Wang, Yijuan Lu, Dinei Florencio, Cha Zhang, and Furu Wei. 2021. Layoutxlm: Multimodal pre-training for multilingual visually-rich document understanding. *arXiv preprint arXiv:2104.08836*.

Vikas Yadav, Rebecca Sharp, and Steven Bethard. 2018. Deep affix features improve neural named entity recognizers. In *Proceedings of the seventh joint conference on lexical and computational semantics*, pages 167–172.

Ikuya Yamada, Akari Asai, Hiroyuki Shindo, Hideaki Takeda, and Yuji Matsumoto. 2020. Luke: deep contextualized entity representations with entity-aware self-attention. *arXiv preprint arXiv:2010.01057*.

GuoDong Zhou and Jian Su. 2002. Named entity recognition using an hmm-based chunk tagger. In *Proceedings of the 40th annual meeting of the association for computational linguistics*, pages 473–480.

7 Supplemental Material

Table 2 lists off-the-shelf NER tools (text-based & image-based) for practical usage.

Tool	Link
Text-based NER	
StanfordCoreNLP	stanfordnlp.github.io/CoreNLP
NeuroNER	neuroner.com
spaCy	spacy.io/api/entityrecognizer
NLTK	nltk.org
OpenNLP	opennlp.apache.org/
Image-based NER	
LayoutLM	huggingface.co/microsoft/layoutlm-base-uncased
LayoutLMv2	huggingface.co/microsoft/layoutlmv2-base-uncased
LayoutLMv3	huggingface.co/microsoft/layoutlmv3-base
LayoutXLM	huggingface.co/microsoft/layoutxlm-base

Table 2: Off-the-shelf NER Tools for practical usage.