Large Sequence Representation Learning via Multi-Stage Latent Transformers

Ionut-Catalin Sandu **Daniel Voinea** **Alin-Ionut Popa**

Amazon Inc.

{sanion, dvoinea, popaaln}@amazon.com

Abstract

We present **LANTERN**, a multi-stage transformer architecture for named-entity recognition (NER) designed to operate on indefinitely large text sequences (*i.e.* $\gg 512$ elements). For a given image of a form with structured text, our method uses language and spatial features to predict the entity tags of each text element. It breaks the quadratic computational constraints of the attention mechanism by operating over a learned latent space representation which encodes the input sequence via the cross-attention mechanism while having the multi-stage encoding component as a refinement over the NER predictions. As a proxy task, we propose **RADAR**, an LSTM classifier operating at character level, which predicts the relevance of a word with respect to the entity-recognition task. Additionally, we formulate a challenging novel NER use case, nutritional information extraction from food product labels. We created a dataset with $11,926$ images depicting food product labels entitled **TREAT** dataset, with fully detailed annotations. Our method achieves superior performance against two competitive models designed for long sequences on the proposed **TREAT** dataset.

I. Introduction. Information extraction from images and unstructured text plays a key role in natural language understanding (NLU) with direct impact in domains such as news content synthesis (Ozsoy et al., 2011; Moratanch and Chitrakala, 2017; Foong et al., 2015), query search (Choi et al., 2018) or knowledge-base systems (Agirre et al., 2018). The transformer architecture (Vaswani et al., 2017) played a pivotal role in advancing the state-of-the art due to its robustness with respect to tasks related to sequential data manipulation and understanding. Its success relies mostly in the attention mechanism, considered as a key component for its versatility and its ability to self-specialize and filter the input information flow.

At the same time, the attention is a curse for transformers, as it scales quadratically with respect to the size of the input sequence. To overcome this limitation, we propose a sequence parsing framework for textual information understanding allowing for indefinitely large input sequences, with focus on the NER task. Our method is entitled **LANTERN** (*i.e.* **LA**rge seque**N**ce **T**ransform**ER** for **NER**). This is achieved via a multi-stage transformer approach over a latent space representation of the input sequence inspired from (Jaegle et al., 2021a) and adapted to NER. The key innovation we propose is a framework which analyses the entire sequence via a latent space attention-based modelling and leverages a multi-stage prediction setup.

II. Related Work. The transformer based architecture led to significant advances across major areas of NLU such as text classification (Lin et al., 2021), question answering (Choi et al., 2018), sequence to sequence translation (Rae et al., 2019; Lewis et al., 2019) and sentiment analysis (Zhang et al., 2018).

Pretraining (Yang et al., 2019; Devlin et al., 2018; Brown et al., 2020) proved extremely useful for generic language understanding tasks. These are unsupervised learning models which produce a vocabulary representation of the unlabelled training corpus via a masked language modelling task. One such model is Bidirectional Encoder Representation from Transformers (BERT) (Devlin et al., 2018). Given the pretrained weights of the model obtained from a large unlabelled data corpus, it has proven quite impactful in terms of peak performance and versatility for different downstream linguistic tasks, such as question answering (Garg et al., 2020; Laskar et al., 2020), NER (Xu et al., 2020b,a; Li et al., 2021; Zhang et al., 2020; Hwang et al., 2020; Lee et al., 2022; Huang et al., 2022; Sandu et al., 2022) or sentence / document classification (Lin et al., 2021).

Most of these approaches are constrained by

Proceedings of the 29th International Conference on Computational Linguistics, pages 4633–4639
October 12–17, 2022.

(a) (b) (c)

Figure 1: **(a) TREAT labelling sample.** Visualizations of entity-based annotations containing nutrient information. First, the annotators look for the `[PER100]` keyword, followed by the identification of the key / value pair of the nutrients of interest. **(b) Word cloud visualization of TREAT words.** The vocabulary of interest is strongly related to ingredient and nutritional information word corpus. Some of the highest occurring words are measurements such as *g* or *kcal*. **(c) Stage-wise performance of LANTERN on TREAT dataset.** The performance improves across stage and batch dimensions, with a significant gap across the first stages.

				Carbohydrates		Energy		Proteins		Salt		Sugar		Total Fat		Saturated Fat	
Statistic	**All**	**Other**	**Per 100**	**Key**	**Value**	**Key**	**Value**	**Key**	**Value**	**Key**	**Value**	**Key**	**Value**	**Key**	**Value**	**Key**	**Value**
Min / Max	4/2007	4/1916	0/24	0/44	0/11	0/22	0/17	0/26	0/10	0/20	0/10	0/31	0/10	0/49	0/19	0/14	0/7
Mean	488.63	458.69	2.01	1.36	1.00	1.36	1.92	1.18	1.00	0.98	0.81	2.12	0.93	1.33	1.03	2.11	0.64
Std.	165.8	162.24	1.94	2.12	0.82	1.8	1.68	1.6	0.8	1.5	0.8	3.26	0.75	2.23	0.98	4.86	0.79
Median	482	462	2	1	1	1	2	1	1	1	1	1	1	1	1	0	0

Table 1: **TREAT Dataset Image-Level Statistics**. Per image statistics with the words corresponding to each unique entity from **TREAT**. The `[OTHER]` entity is dominant (*i.e.* it represents $\approx 85\%$ out of the total labelled data corpus) as the nutrient section of a food label usually occupies around 10% of the total space of the label. The mean / median statistics does not reflect directly the length of the sequence inputed to **LANTERN**. It is processed via a tokeniser (Kudo, 2018) and the resulted length is usually $2.5\times$ longer.

the quadratic computational limitation of the attention mechanism (sequences of up to 512 elements), thus bounded to subcontext. Several approaches (Dai et al., 2019; Goyal et al., 2021; Beltagy et al., 2020; Zaheer et al., 2020; Ainslie et al., 2020) considered modelling the sequence as a whole from the prism of the attention mechanism, even for sequence lengths $\gg 512$.

In (Beltagy et al., 2020), the authors propose a sparse attention mechanism to analyse larger sequences of up to 4096 elements, while keeping the computational time of the attention matrix linear with respect to the input size. A similar setup is also proposed in (Zaheer et al., 2020) using an additional random iterative attention mechanism to parse the entire graph of the sequence. Different from these approaches, our proposed pipeline can process indefinitely large sequences and has a stage-wise refinement mechanism of the NER predictions.

III. Dataset. Additionally, we introduce a new language modelling problem setup from the umbrella of NER tasks, namely nutritional information extraction from images of food products. This was achieved by collecting a dataset of food product images, entitled **TREAT** (*i.e.* nu**TR**i**E**nt f**A**c**T**s) with $11,926$ images depicting food products with fully extracted text information, bounding box image localization and complete NER class label annotations with sequences of up to $\approx 2,000$ words (≈ 5000 tokens) to be parsed. Labelling was performed using a web interface highlighting words. Annotators had to individually assign classes to words of interest. The proposed NER use-case is formulated on a linguistically diverse dataset in terms of semantics and specialisation. The collected data covers multiple European languages (EN, FR, DE, ES, IT, etc.) containing the nutritional information expressed with different languages within the same input text sequence. We further explore the specialisation in the compliance domain, learning to reason within the vocabulary of text content from food product labels. They must contain information related, but not limited, to ingredients, storage instructions, manufacturer addresses, nutritional information, allergen statements, advertising and production process details. Every piece of information can be found in standalone paragraphs, lists or table structures. The general layout is rich in both structure and visual cues. In terms of textual information, the collected data has both semantic text with full paragraphs or text structures (*e.g.* instructions, tables, lists or advertisement sections) and independent sentences

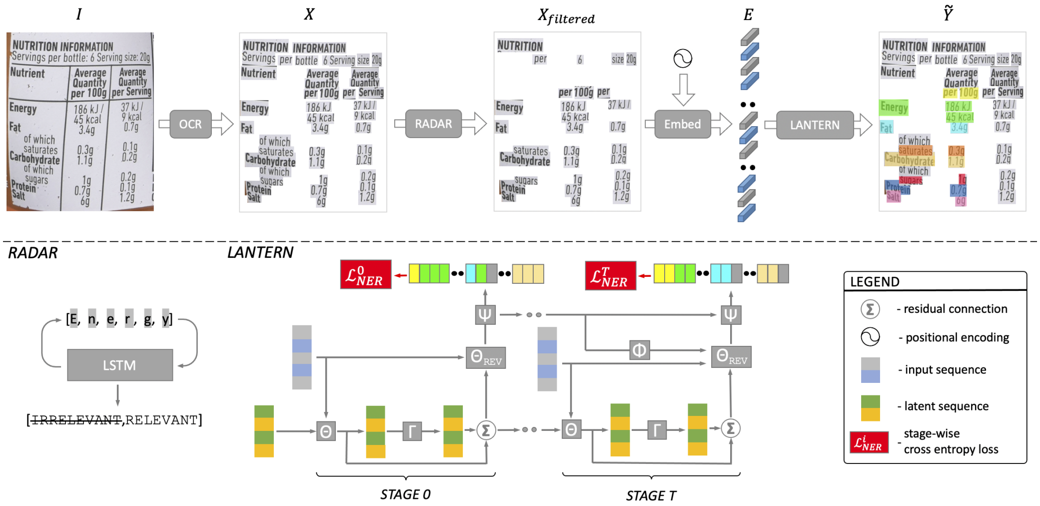

Figure 2: **Detailed overview of our proposed large-sequence parsing model.** Given an image $\mathbf{I}$, we first extract all the words using a state-of-the-art algorithm, obtaining the complete list of words $\mathbf{X}$. Next, $\mathbf{X}$ is filtered using **RADAR** thus obtaining the most relevant words with respect to our entity recognition task, $\mathbf{X}_{filtered}$. The filtered list is embedded into a language and positional representation, $\mathbf{E}$, of the remaining words. Lastly, $\mathbf{E}$ goes through **LANTERN** to generate predictions $\tilde{\mathbf{Y}}$.

				Carbohydrates		Energy		Proteins		Salt		Sugar		Total Fat		Saturated Fat	
Model	**All**	**Other**	**Per 100**	**Key**	**Value**	**Key**	**Value**	**Key**	**Value**	**Key**	**Value**	**Key**	**Value**	**Key**	**Value**	**Key**	**Value**
(Beltagy et al., 2020)	0.74	0.96	0.78	0.78	0.71	0.80	0.70	0.78	0.71	0.74	0.68	0.78	0.73	0.75	0.72	0.67	0.52
(Zaheer et al., 2020)	0.76	0.96	0.79	0.77	0.74	0.83	0.69	0.78	**0.78**	0.79	0.77	0.80	0.77	0.79	0.74	0.64	0.53
(Xu et al., 2020b)	0.61	0.96	**0.86**	0.59	0.55	0.69	0.67	0.45	0.41	0.41	0.4	0.55	0.53	0.51	0.50	0.48	0.44
Ours	**0.86**	**0.97**	0.83	**0.94**	**0.82**	**0.94**	**0.87**	**0.93**	0.77	**0.91**	**0.79**	**0.94**	**0.76**	**0.89**	**0.76**	**0.87**	**0.78**

Table 2: **TREAT Test Set - Word Level Results**. The LayoutLM baseline failed to provide quality results as its context window was limited to 512 elements. Approaches of (Beltagy et al., 2020) and (Zaheer et al., 2020) obtain superior results as they are able to handle larger sequences (*i.e.* 2048). The highest F1-score is obtained with **LANTERN** as it is able to grasp the entire textual context of the product content.

such as titles, individual statements or website references. As illustrated in figure 1 (b), the downloaded images contain a high variability in terms of vocabulary corpus, mainly related to ingredient or nutritional aspects. Check figure 1 (a) for a labelling sample and table 1 for dataset statistics. Additionally, we perform an analysis from a linguistic point of view and notice the following distribution of dominant languages: 40% english, 11% spanish, 11% italian, 10% german and 8% french. The rest of 20% are data samples with mixed or other languages, unable to determine the dominant one.

We build a vocabulary with $15,000$ unique tokens using (Kudo, 2018). Up to 24% of the resulted tokens are made of 5 characters, while tokens with different sizes occupy a relatively uniform space (*e.g.* 13% tokens are 7 characters long, 10% tokens are 6 characters long and tokens with 2 or 3 characters take 8% of the vocabulary space).

IV. Methodology. Let $\mathbf{I} \in \mathbb{R}^{h \times w}$ denote an image containing a list $\mathbf{X} = (x_1 \ldots x_{M_\mathbf{I}})$ of $M_\mathbf{I}$ words, with $x_i = (w_i, c_i)$ where w_i is the OCR (AWS-Textract, 2019) extracted word and $c_i \in [0,1]^4$ represents the bounding box coordinates relative to dimensions of $\mathbf{I}$. Additionally, let $\mathbf{Y} = (y_1 \ldots y_{M_I})$ denote the class labels for each word w_i in $\mathbf{X}$.

Given the high inflation of `[OTHER]` inside **TREAT** (see table 1) we constructed **RADAR** (**R**elev**A**nt wor**D** cl**A**ssifie**R**) using a bidirectional LSTM model (Hochreiter and Schmidhuber, 1997) at character level to predict if the word is relevant or irrelevant (`[OTHER]`). Thus, we have $\mathbf{X}_{filtered} = \{x_i \mid \mathbf{RADAR}(x_i) > \alpha, i = 1 \ldots M_\mathbf{I}\}$ where α is validated to obtain a recall of 1 for relevant words. To simplify notation, we identify $\mathbf{X}_{filtered}$ with $\mathbf{X}$.

The language based embeddings for w_i are supplemented with positional encodings (Vaswani et al., 2017). Thus, we obtain the embedding representation $\mathbf{E} = (e_1 \ldots e_{M_\mathbf{I}})$ of $\mathbf{X}$ where $e_i = (e^w_{i1} \ldots e^w_{id_w}, e^x_{i1} \ldots e^x_{id_x}, e^y_{i1} \ldots e^y_{id_y})$.

This leads to an input sequence $\mathbf{E} \in \mathbb{R}^{M\times d}$, where $d = d_w + d_x + d_y$. The dimensions d_w, d_x and d_y denote language embedding sizes while x and y denote positional variables with respect to image width and height, respectively. Additionally, a latent block representation denoted with $\mathbf{L}^{\texttt{INIT}} \in \mathbb{R}^{N\times d}$ ($N \ll M$) is learned. The intuition behind the latent block is to learn a projection of the most relevant information with respect to the NER task.

The **LANTERN** module is applied in a stage-wise manner. Each stage receives as input the embedded input sequence, $\mathbf{E}$, and a latent block, $\mathbf{L}^t \in \mathbb{R}^{N\times d}$, and it predicts an array, $\tilde{\mathbf{Y}}^t$. In the following, we will describe the computational stages $t \in \{0 \dots T\}$. For $t = 0$, we have $\mathbf{L}^{\texttt{INIT}} \in \mathbb{R}^{N\times d}$.

STAGE t=0: we start by applying a cross-attention mechanism $\Theta : \mathbb{R}^{M\times d} \rightarrow \mathbb{R}^{N\times d}$, over $\mathbf{E}$ and $\mathbf{L}^{\texttt{INIT}}$. Functions $k(\cdot)$, $q(\cdot)$ and $v(\cdot)$ represent the keys, queries and values, respectively.

$$\Theta(\mathbf{E}, \mathbf{L}) = softmax(\frac{q(\mathbf{L})k(\mathbf{E})^\top}{\sqrt{d}})v(\mathbf{E}) \quad (1)$$

This will basically result in a projection of the input sequence $\mathbf{E}$ to the latent space $\mathbb{R}^{N\times d}$ through $\mathbf{L}^{\texttt{INIT}}$. We will denote the resulted projected latent block with $\mathbf{L}^{\texttt{INIT}}_{\Theta} = \Theta(\mathbf{E}, \mathbf{L}^{\texttt{INIT}})$.
Next, $\mathbf{L}^{\texttt{INIT}}_{\Theta}$ is passed through a transformer encoder, $\Gamma : \mathbb{R}^{N\times d} \rightarrow \mathbb{R}^{N\times d}$, to learn an implicit statistic between the elements of the latent block elements. The initial latent block information is added as a residual information on the resulting encoded information from Γ, thus having

$$\mathbf{L}^{\texttt{INIT}}_{\Gamma} = \Gamma(\mathbf{L}^{\texttt{INIT}}_{\Theta}) + \mathbf{L}^{\texttt{INIT}} \quad (2)$$

Lastly, we apply a reversed cross-attention mechanism, $\Theta_{\texttt{REV}} : \mathbb{R}^{N\times d} \rightarrow \mathbb{R}^{M\times d}$ over the input sequence $\mathbf{E}$ and the latent block $\mathbf{L}^{\texttt{INIT}}_{\Gamma}$ processed through transformer Γ.

$$\Theta_{\texttt{REV}}(\mathbf{E}, \mathbf{L}) = softmax(\frac{q(\mathbf{E})k(\mathbf{L})^\top}{\sqrt{d}})v(\mathbf{L}) \quad (3)$$

The reversed cross-attention operation provides a reprojection of the transformed latent block information $\mathbf{L}^{\texttt{INIT}}_{\Gamma}$ to the input space, $\mathbb{R}^{M\times d}$. Thus, the resulted information from the reversed cross-attention will be, $\mathbf{L}^{\texttt{INIT}}_{\Theta_{\texttt{REV}}} = \Theta_{\texttt{REV}}(\mathbf{E}, \mathbf{L}^{\texttt{INIT}}_{\Gamma})$.

$\mathbf{L}^{\texttt{INIT}}_{\Theta_{\texttt{REV}}}$ is next passed through a feed-forward network, $\Psi(\cdot)$, which provides the class predictions for the sequence's entities, $\tilde{\mathbf{Y}}^{t=0}$ (*i.e.* $t = 0$ denotes the first stage), and it is being optimized using a cross-entropy loss function.

STAGE t>0: This process is repeated for several iterations, in a multi-stage setup. For the next stages (*i.e.* stage $0 < t \leq T$), the latent sequence $\mathbf{L}^t$ is initialized with the latent block information obtained from equation 2 of stage $t-1$.

$$\mathbf{L}^t = \begin{cases} \mathbf{L}^{\texttt{INIT}}_{\Gamma}, & \text{if } t = 1 \\ \mathbf{L}^{t-1}_{\Gamma}, & \text{otherwise} \end{cases}$$

The prediction of stage $t-1$ of the NER task, $\tilde{\mathbf{Y}}^{t-1}$, is fed to the feed-forward module of the current stage, t, as a weight factor. Also, $\tilde{\mathbf{Y}}^{t-1}$, is used as a weight to the attention-matrix computed at the current stage, t, obtained via a projection function $\Phi : \mathbb{R}^{M\times d_{target}} \rightarrow \mathbb{R}^{M\times N}$ where d_{target} is the total number of entities to be predicted

$$\Theta_{\texttt{REV}}(\mathbf{E}, \mathbf{L}, \mathbf{A}) = softmax(\frac{q(\mathbf{E})k(\mathbf{L})^\top}{\sqrt{d}} \odot \mathbf{A})v(\mathbf{L})$$

where $\odot(\cdot) : \mathbb{R}^{M\times N} \rightarrow \mathbb{R}^{M\times N}$ represent the Hadamard product and $\mathbf{A} \in \mathbb{R}^{M\times N}$ is a weight matrix. Thus, the reprojected latent block information from stage t becomes, $\mathbf{L}^t_{\Theta_{\texttt{REV}}} = \Theta_{\texttt{REV}}(\mathbf{E}, \mathbf{L}^t_{\Gamma}, \Phi(\tilde{\mathbf{Y}}^{t-1})))$.

At each stage, a cross-entropy loss is applied over the output of $\Psi : \mathbb{R}^{M\times d} \rightarrow \mathbb{R}^{M\times d_{target}}$, which are cumulated until the final stage T, $\mathcal{L}_{\texttt{NER}} = \sum_{t=0}^{T} \mathcal{L}^t_{\texttt{NER}}(\tilde{\mathbf{Y}}^t, \mathbf{Y})$.

Given an input image $\mathbf{I}$ and a learned latent block $\mathbf{L}$, the framework outputs an array $\tilde{\mathbf{Y}}$ with class predictions for all the relevant words identified within image $\mathbf{I}$. A step-by-step breakdown of our pipeline is illustrated in figure 2 and in algorithm 1.

V. Experiments. We conduct experiments on our proposed dataset **TREAT**. Prior to applying **LANTERN**, we filtered a part of the irelevant word corpus (*i.e.* `[OTHER]`) using **RADAR**. The accuracy of **RADAR** is 85% and we filtered 48% of the irrelevant words. The reason behind filtering such a low quantity of `[OTHER]` with respect to the obtained accuracy is because the majority of the word corpus from **TREAT** represents numerical values (*e.g.* *100g*, *30ml*) which we marked with `[NUM]`. These words (`[NUM]`) cannot be filtered without leveraging their positional context as they can be linked with nutrient keys. Thus, we decided to introduce them as such inside **LANTERN** without

(Stage 0) (Stage 1) (Stage 2) (Stage 3) (Stage 4) (Stage 5) (Ground Truth)

Figure 3: **Multi-stage Qualitative Results obtained with LANTERN on TREAT Test Set.** We show the progressive improvement of the generated entities achieved via the multi-stage component of LANTERN. The refinement is clearly visible from stage to stage, as the model turns its attention towards the words describing the nutrients of interest.

RADAR filtration. In table 2 we report the word-level precision, recall and F1-score. We use a total of 6 stages with a total embedding size of $d = 64$, where d_w, d_x and d_y are set to the values of 32, 16 and 16, respectively. The dimension N of latent block $\mathbf{L}$ is 256. Our method is able to achieve superior results compated to similar NER frameworks (Xu et al., 2020b), (Beltagy et al., 2020) and (Zaheer et al., 2020), some of them being designed for long sequences. For a fair evaluation, we compare against other competitive NER baselines using text and position embeddings only.

Algorithm 1 LANTERN

Input: $\mathbf{I} \in \mathbb{R}^{h \times w}$, $\mathbf{L}^{\mathtt{INIT}} \in \mathbb{R}^{N \times d}$

Output: $\tilde{Y} = (\tilde{y_1} \dots \tilde{y_M})$

$\mathbf{X} \leftarrow \text{OCR}(\mathbf{I})$

$\mathbf{X}_{filtered} \leftarrow \mathbf{RADAR}(\mathbf{X})$

$\mathbf{E} \leftarrow Embed(\mathbf{X}_{filtered})$

$\mathbf{L}_\Theta^{\mathtt{INIT}} \leftarrow \Theta(\mathbf{E}, \mathbf{L}^{\mathtt{INIT}})$

$\mathbf{L}_\Gamma^{\mathtt{INIT}} \leftarrow \Gamma(\mathbf{L}_\Theta^{\mathtt{INIT}}) + \mathbf{L}^{\mathtt{INIT}}$

$\tilde{\mathbf{Y}}^0 \leftarrow \Psi(\Theta_{\mathtt{REV}}(\mathbf{E}, \mathbf{L}_\Gamma^{\mathtt{INIT}}))$

for $t = 1 \dots T$ **do**

  if $t = 1$ **then**

    $\mathbf{L}^t \leftarrow \mathbf{L}_\Gamma^{\mathtt{INIT}}$

  else

    $\mathbf{L}^t \leftarrow \mathbf{L}_\Gamma^{t-1}$

  $\mathbf{L}_\Theta^t \leftarrow \Theta(\mathbf{E}, \mathbf{L}^t)$

  $\mathbf{L}_\Gamma^t \leftarrow \Gamma(\mathbf{L}_\Theta^t) + \mathbf{L}^t$

  $\tilde{\mathbf{Y}}^t \leftarrow \Psi(\Theta_{\mathtt{REV}}(\mathbf{E}, \mathbf{L}_\Gamma^t, \Phi(\tilde{\mathbf{Y}}^{t-1})))$

$\tilde{\mathbf{Y}} \leftarrow \tilde{\mathbf{Y}}^T$

In figure 3 we showcase multi-stage predictions, thus highlighting the refinement aspect of our framework on a practical example.

Method	Precision	Recall	F1-Score
LANTERN w/o RADAR	0.83	0.80	0.81
LANTERN w RADAR	**0.84**	**0.87**	**0.86**
LANTERN w/o weight sharing	0.83	0.80	0.81
LANTERN w weight sharing	**0.82**	**0.82**	**0.82**
LANTERN w Language	0.39	0.42	0.40
LANTERN w Language + Positional	**0.83**	**0.80**	**0.81**

Table 3: **Ablation study on various components of LANTERN over TREAT dataset.** The highest impact in terms of performance is obtained with the **RADAR** based filtration of `[OTHER]`. Parameter weight sharing across all the stages also impacts positively the performance and it helps the model to convergence faster.

In table 3 we analyse the impact of the various model components and embedding types used by our model, thus validating the architectural innovative aspects of our method. In figure 1 (c) we show the impact of the multi-stage component. The refinement aspect of the method is visible between consecutive stages leading to significant improvement between the first and last stages.

VI. Conclusions. We propose a novel method for NER, specifically designed for indefinitely large sequences. It leverages a multi-stage transformer-based pipeline which breaks the quadratic computational constraints of the attention mechanism by projecting the input sequence to a latent space representation and a stage-wise setup which acts as a context refinement of the entity predictions. The methodology was evaluated on a novel and difficult NER use case, nutritional information extraction, proving superior results over other strong baselines specifically designed for long sequence parsing.

Acknowledgements. We thank George Leotescu and Adrian-Mihai Alecu for helpful insights.

References

Eneko Agirre, Oier López de Lacalle, and Aitor Soroa. 2018. The risk of sub-optimal use of open source nlp software: Ukb is inadvertently state-of-the-art in knowledge-based wsd. *arXiv preprint arXiv:1805.04277*.

Joshua Ainslie, Santiago Ontanon, Chris Alberti, Vaclav Cvicek, Zachary Fisher, Philip Pham, Anirudh Ravula, Sumit Sanghai, Qifan Wang, and Li Yang. 2020. Etc: Encoding long and structured inputs in transformers. In *EMNLP*, Online. Association for Computational Linguistics.

AWS-Textract. 2019. Textract-aws ocr engine. `https://aws.amazon.com/textract`.

Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *NeurIPS*, volume 33, pages 1877–1901. Curran Associates, Inc.

Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wen-tau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. 2018. Quac: Question answering in context. *arXiv preprint arXiv:1808.07036*.

Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc V Le, and Ruslan Salakhutdinov. 2019. Transformer-xl: Attentive language models beyond a fixed-length context. *arXiv preprint arXiv:1901.02860*.

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*.

Oi-Mean Foong, Suet-Peng Yong, and Farha-Am Jaid. 2015. Text summarization using latent semantic analysis model in mobile android platform. In *AMS*, pages 35–39. IEEE.

Siddhant Garg, Thuy Vu, and Alessandro Moschitti. 2020. Tanda: Transfer and adapt pre-trained transformer models for answer sentence selection. In *AAAI*.

Łukasz Garncarek, Rafał Powalski, Tomasz Stanisławek, Bartosz Topolski, Piotr Halama, Michał Turski, and Filip Graliński. 2021. Lambert: layout-aware language modeling for information extraction. In *International Conference on Document Analysis and Recognition*, pages 532–547. Springer.

Anirudh Goyal, Aniket Didolkar, Alex Lamb, Kartikeya Badola, Nan Rosemary Ke, Nasim Rahaman, Jonathan Binas, Charles Blundell, Michael Mozer, and Yoshua Bengio. 2021. Coordination among neural modules through a shared global workspace. *arXiv preprint arXiv:2103.01197*.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

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*.

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*.

Andrew Jaegle, Sebastian Borgeaud, Jean-Baptiste Alayrac, Carl Doersch, Catalin Ionescu, David Ding, Skanda Koppula, Daniel Zoran, Andrew Brock, Evan Shelhamer, et al. 2021a. Perceiver io: A general architecture for structured inputs & outputs. *arXiv preprint arXiv:2107.14795*.

Andrew Jaegle, Felix Gimeno, Andrew Brock, Andrew Zisserman, Oriol Vinyals, and Joao Carreira. 2021b. Perceiver: General perception with iterative attention. *arXiv preprint arXiv:2103.03206*.

Taku Kudo. 2018. Subword regularization: Improving neural network translation models with multiple subword candidates. *arXiv preprint arXiv:1804.10959*.

Md Tahmid Rahman Laskar, Xiangji Huang, and Enamul Hoque. 2020. Contextualized embeddings based transformer encoder for sentence similarity modeling in answer selection task. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 5505–5514.

Chen-Yu Lee, Chun-Liang Li, Timothy Dozat, Vincent Perot, Guolong Su, Nan Hua, Joshua Ainslie, Renshen Wang, Yasuhisa Fujii, and Tomas Pfister. 2022. Formnet: Structural encoding beyond sequential modeling in form document information extraction. *arXiv preprint arXiv:2203.08411*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Peizhao Li, Jiuxiang Gu, Jason Kuen, Vlad I Morariu, Handong Zhao, Rajiv Jain, Varun Manjunatha, and Hongfu Liu. 2021. Selfdoc: Self-supervised document representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5652–5660.

Yuxiao Lin, Yuxian Meng, Xiaofei Sun, Qinghong Han, Kun Kuang, Jiwei Li, and Fei Wu. 2021. Bertgcn: Transductive text classification by combining gcn and bert. *arXiv preprint arXiv:2105.05727*.

N Moratanch and S Chitrakala. 2017. A survey on extractive text summarization. In *ICCCSP*, pages 1–6. IEEE.

Makbule Gulcin Ozsoy, Ferda Nur Alpaslan, and Ilyas Cicekli. 2011. Text summarization using latent semantic analysis. *Journal of Information Science*, 37(4):405–417.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *NeurIPS*.

Shuai Peng, Ke Yuan, Liangcai Gao, and Zhi Tang. 2021. Mathbert: A pre-trained model for mathematical formula understanding. *arXiv preprint arXiv:2105.00377*.

Alin-Ionut Popa and Cristian Sminchisescu. 2017. Parametric image segmentation of humans with structural shape priors. In *ACCV*.

Jack W Rae, Anna Potapenko, Siddhant M Jayakumar, and Timothy P Lillicrap. 2019. Compressive transformers for long-range sequence modelling. *arXiv preprint arXiv:1911.05507*.

Ionut-Catalin Sandu, Daniel Voinea, and Alin-Ionut Popa. 2022. Consent: Context sensitive transformer for bold words classification. *arXiv preprint arXiv:2205.07683*.

Ioana-Sabina Stoian, Ionut-Catalin Sandu, Daniel Voinea, and Alin-Ionut Popa. 2022. Unstructured object matching using co-salient region segmentation. In *CVPR IMW Workshop*, pages 5051–5060.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *NIPS*.

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*, page 1192–1200.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *NeurIPS*, 32.

Wenwen Yu, Ning Lu, Xianbiao Qi, Ping Gong, and Rong Xiao. 2021. Pick: processing key information extraction from documents using improved graph learning-convolutional networks. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 4363–4370. IEEE.

Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, et al. 2020. Big bird: Transformers for longer sequences. In *NeurIPS*.

Lei Zhang, Shuai Wang, and Bing Liu. 2018. Deep learning for sentiment analysis: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 8(4):e1253.

Peng Zhang, Yunlu Xu, Zhanzhan Cheng, Shiliang Pu, Jing Lu, Liang Qiao, Yi Niu, and Fei Wu. 2020. Trie: End-to-end text reading and information extraction for document understanding. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 1413–1422.