ABSTRACT

Knowledge base population (KBP) from texts involves the extraction and organization of information from unstructured textual data to enhance or create a structured knowledge base. This process is crucial for various applications, such as natural language understanding, question-answering systems, and knowledge-driven decision-making. However the difficulty lies in the complexity of natural language, which is nuanced, ambiguous, and context-dependent. Extracting accurate and reliable information requires overcoming challenges such as entity disambiguation and relation extraction which are time-consuming tasks for users.Shadowfax is an interactive platform designed to support users by streamlining the process of knowledge base population (KPB) from text documents. Unlike other existing tools, it relies on a unified machine learning model to extract relevant information from unstructured text, enabling operational agents to gain a quick overview. The proposed system supports a variety of natural language processing (NLP) tasks using a single architecture, while presenting information in the most comprehensive way possible to the end user.

CCS CONCEPTS

• **Computing methodologies** → **Information extraction**.

KEYWORDS

Information Extraction, Knowledge Base Population, Deep-Learning, Data Mining, User in the Loop, End-to-End

1 INTRODUCTION

The exponential growth of online unstructured information has made the task of extracting valuable knowledge very challenging for many fields such as journalism, business intelligence [16] or medicine [18]. Fully manual extraction is a solution which tends to become costly in terms of time and resources while being subject to a risk of error insertion due to the meticulousness needed. Navigating through such abundant data requires sophisticated systems capable of processing large datasets to extract and update already known information, often stored in a Knowledge Base (KB). Over the past few years multiple Knowledge Base Population (KBP) systems trying to extract knowledge from raw text and to link it to entities stored into a KB have been proposed [29] . However, these solutions are often composed of multiple isolated processing modules [9, 12], weighing down the system, and reducing possible interaction between the components, resulting in a loss of information and an accumulation of errors.

In this paper we propose Shadowfax, a cost-effective and practical solution which differs from existing solutions in that it is relies on a *unified Information Extraction (IE) model which merges multiple KBP components into a single one* to output knowledge graphs from unstructured documents. To the best of our knowledge, this is the first platform based on a unified model, enabling the KB to be enriched in a single pass, instead of using four models for entity extraction, co-reference resolution, relationship extraction and entity linking respectively, at a much higher cost in terms of time and resources. There is a risk of error amplification in using successively four models, which is avoided in our approach. In synthesis, our contributions are as follows:

- we propose a model classifying different kinds of interactions between pairs of candidate entities in a text and thus producing the whole textual graph *in a single pass.*
- we measure the performances of the proposed model on two different datasets, showing that results are not affected by the removal of specialized models.
- we integrate the latter model into Shadowfax, a KBP system featuring an HMI (Human Machine Interface) to assist users in the task of Knowledge Base Population.

2 RELATED WORK

Textual KBP encompasses several NLP tasks [14], comprising Named Entity Recognition (NER), Coreference Resolution (CR), Relation Extraction (RE), and Entity Linking (EL). Those tasks are generally considered separately. The aim of NER is to detect entity mentions within the text and to assign pertinent types to these mentions. This recognition is commonly solved by using a Transformer [21] and a trained classification layer to assign a IOB (Inside, Outside, Beginning) format tag as well as a type to each token in the text [17]. Co-reference resolution is used to group textual mentions referring

to the same entity [6, 13]. Using the text, it is then possible to identify the relations between the previously created textual entities [4, 20]. Finally, EL identifies the database entities that appear in the text, and creates new ones for entities that lack prior recognition. This last step is commonly solved by creating entity embeddings, wherein the contextual similarity between entities is represented by the distance in a vectorial space [3, 23]. A traditional approach to tackling the KBP is built following a workflow that integrates state-of-the-art models for each sub-task [9, 15]. However, optimizing components individually does not guarantee overall improvement. In addition, this method faces other challenges, including extensive training times and resource-intensive inference.

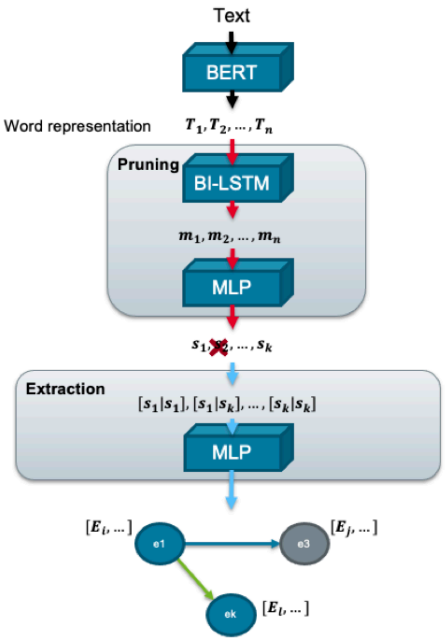

Figure 1: Shadowfax Information Extraction model architecture

3 SYSTEM OVERVIEW

Shadowfax is a web application designed to assist companies, organizations or independent users to populate a Knowledge Base (KB). The solution can be used to correct extracted results after an automatic population, or as support during a manual process. The platform breaks down into a technical part for Information Extraction (shown in Figure 1) and Entity Linking (EL), and a visualization, manipulation and validation part on the user side (Figure 2).

3.1 Information Extraction

The unified model is responsible for producing the textual graph, summarizing entities of interest and relations between them from an input document. As detailed in Figure 1, the unified model is composed of two parts: a first step in charge of pruning mentions and a second one of predicting interactions between those mentions.

Pruning. Recent NER solutions [25, 27, 31] address the task as the prediction of spans. In addition to proving its superiority, this approach is also more flexible than Part Of Speech (POS) tagging, as it allows the handling of nested mentions ("[Bank of [China]]"). Span prediction, on the other hand requires a strategy for correctly handling the large number of candidates, which can result in significant overhead. In our model, we utilize a pruning module with an architecture inspired from Yu et al. [27], although originally applied to predict the textual mentions and their types directly. A fine-tuned BERT [5] Pretrained Language Model (PLM) is used to represent each word within the input text. The model concatenates output token representations from the two last layers. If a word consists of several tokens, the average of their representations is used to form that of the word. Then, a 3-layer Bi-LSTM [8] creates a new representation of words by reinforcing contextual awareness. Subsequently, candidate spans are formed by concatenating the head and tail word representations of the text fragment. Finally, a Multi-Layer Perceptron (MLP) uses these concatenations to predict whether the spans actually refer to entities.

Entity interactions. Once the entity mentions have been identified, an extraction module classifies the interactions between all possible pairs of mentions. Since the PLM serves as a common base for both modules, the word representations produced by the encoder are reused for the interaction of the entities. Using the concatenation of the representations of two spans, a second Multi-Layer Perceptron predicts the interactions existing between all the possible pairings of the previously selected spans. The output of this MLP is a vector of dimensions equal to the number of interactions of interest (entity or relation types and co-reference). A mask is applied when the pair contains two identical spans in order to predict only the types of the entity. On the contrary, only the dimensions related to relation types and co-reference are considered for different spans. From the output the graph is constructed by grouping together the pairs of mentions for which a co-reference has been predicted. The types associated with an entity are the union of those assigned to each of its mentions. The same goes for the relations, a kind of relation is considered between two entities if it is predicted between at least one pair of mentions belonging to these entities.

3.2 Entity Linking

EL involves a range of resolution strategies that depend on the data structure within the target KB [22]. Shadowfax tackles linking through contextual similarity. Similarly to the two-steps solution proposed by Li et al. [10] and Prieur et al. [14], a filtering step uses a fine-tuned PLM to retrieve a list of candidates appearing in texts considered as similar. This encoder takes as input the text for which mentions of a target entity are enclosed within special tokens ("[ENT]", "[/ENT]") to improve the quality of the representations as stated in [2]. The output representation of the "[SEP]" token, inserted at the beginning of the text is then used as the vectorial representation of the in-context entity. Similar contextual representations are then retrieved by the average of the cosine similarity. A matching phase subsequently reorders the list of candidates obtained above. This phase maintains the precedent input format but employs a more refined encoder for ambiguous candidates. During

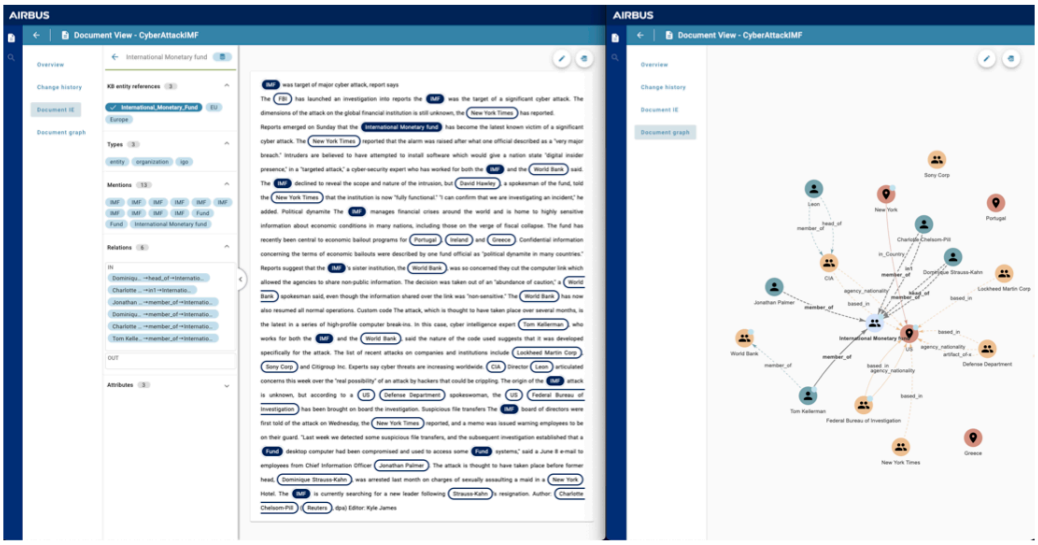

Figure 2: Visualization of textual information extracted on Shadowfax

KBP inference, newly processed texts augment the contextual backdrop of the entities in the KB to enhance and extend the linking capability.

3.3 Human in the loop

The risk of KB pollution caused by the automatic insertion of processing errors is a challenge that tends to constrain the application of EL as a support. This explains the articulation of the model, which places the user as the final validator of the IE. In this sense, the user has the ability to modify all the displayed extracted information (relations, attributes, entities, mentions, matching entities) and to add new elements. For a more synthetic, understandable and easy-to-analyse view, the extracted information is displayed as a graph. Parent types are illustrated by a given color and icon on the nodes. Previously seen elements are showcased with a highlight on the node tag or a solid arrow tail for the relations. Once the user considers the extracted information to be correct and exhaustive, they can save the document state. Once saved, the information is added to the KB and used for future extraction.

4 EXPERIMENTS

5 CONCLUSIONS

This paper introduced Shadowfax, a comprehensive solution that combines a unified Information Extraction model designed to provide a knowledge graph by maximizing interaction between NLP tasks and an Entity Linking module, seamlessly integrated into a Human-Machine-Interface to assist the users in the time-consuming task of Knowledge Base Population. Contrary to the expectation that sharing model weights across multiple tasks might adversely impact performance due to reduced specialization, our unified model shows that the proximity of the tasks and their reciprocal contribution offset the loss of specialization while reducing time and resources requirements.

While we plan to incorporate a modification history, future improvements of the solution will explore the integration of active learning approaches [7, 24] to refine the model through user corrections and behavior, given the favorable conditions of the use case. The exploration of solutions to integrate the entity linking task into the unified model is also planned for future work.

REFERENCES

[1] Alan Akbik, Tanja Bergmann, Duncan Blythe, Kashif Rasul, Stefan Schweter, and Roland Vollgraf. 2019. FLAIR: An Easy-to-Use Framework for State-of-the-Art NLP. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, Waleed Ammar, Annie Louis, and Nasrin Mostafazadeh (Eds.). Association for Computational Linguistics, Minneapolis, Minnesota, 54–59. https://doi.org/10.18653/v1/N19-4010

[2] Livio Baldini Soares, Nicholas FitzGerald, Jeffrey Ling, and Tom Kwiatkowski. 2019. Matching the Blanks: Distributional Similarity for Relation Learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Anna Korhonen, David Traum, and Lluís Màrquez (Eds.). Association for Computational Linguistics, Florence, Italy, 2895–2905. https://doi.org/10.18653/v1/P19-1279

[3] Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. https://openreview.net/forum?id=r1xMH1BtvB

[4] Julien Delaunay, Hanh Thi Hong Tran, Carlos-Emiliano González-Gallardo, Georgeta Bordea, Nicolas Sidere, and Antoine Doucet. 2023. A Comprehensive Survey of Document-level Relation Extraction (2016-2023). arXiv:2309.16396 [cs.CL]

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*. 4171–4186. https://doi.org/10.18653/v1/n19-1423

[6] Vladimir Dobrovolskii. 2021. Word-Level Coreference Resolution. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*. 7670–7675. https://doi.org/10.18653/v1/2021.emnlp-main.605

[7] Paul Guélorget. 2022. *Active learning for the detection of objects of operational interest in open-source multimedia content. (Apprentissage actif pour la détection d'objets d'intérêt opérationnel dans les contenus multimédias)*. Ph. D. Dissertation. Polytechnic Institute of Paris, Palaiseau, France. https://tel.archives-ouvertes.fr/tel-03947344

[8] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Comput.* 9, 8 (1997), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

[9] Nikolaos Kolitsas, Octavian-Eugen Ganea, and Thomas Hofmann. 2018. End-to-End Neural Entity Linking. In *Proceedings of the 22nd Conference on Computational Natural Language Learning, CoNLL 2018, Brussels, Belgium, October 31 - November 1, 2018*. 519–529. https://doi.org/10.18653/v1/k18-1050

[10] Yuliang Li, Jinfeng Li, Yoshihiko Suhara, AnHai Doan, and Wang-Chiew Tan. 2020. Deep Entity Matching with Pre-Trained Language Models. *Proc. VLDB Endow.* 14, 1 (2020), 50–60. https://doi.org/10.14778/3421424.3421431

[11] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv:1907.11692 [cs.CL]

[12] Filipe Mesquita, Matteo Cannaviccio, Jordan Schmidek, Paramita Mirza, and Denilson Barbosa. 2019. KnowledgeNet: A Benchmark Dataset for Knowledge Base Population. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (Eds.). Association for Computational Linguistics, Hong Kong, China, 749–758. https://doi.org/10.18653/v1/D19-1069

[13] Nafise Sadat Moosavi and Michael Strube. 2016. Which Coreference Evaluation Metric Do You Trust? A Proposal for a Link-based Entity Aware Metric. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*. https://doi.org/10.18653/v1/p16-1060

[14] Maxime Prieur, Cédric du Mouza, Guillaume Gadek, and Bruno Grilhères. 2023. Evaluating and Improving End-to-End Systems for Knowledge Base Population. In *Proceedings of the 15th International Conference on Agents and Artificial Intelligence, ICAART 2023, Volume 3, Lisbon, Portugal, February 22-24, 2023*. 641–649. https://doi.org/10.5220/0011726000003393

[15] Maxime Prieur, Souhir Gahbiche, Guillaume Gadek, Sylvain Gatepaille, Kilian Vasnier, and Valerian Justine. 2023. K-pop and fake facts: from texts to smart alerting for maritime security. In *Proceedings of the The 61st Annual Meeting of the Association for Computational Linguistics: Industry Track, ACL 2023, Toronto, Canada, July 9-14, 2023*. 510–517. https://doi.org/10.18653/v1/2023.acl-industry.49

[16] Horacio Saggion, Adam Funk, Diana Maynard, and Kalina Bontcheva. 2007. Ontology-Based Information Extraction for Business Intelligence. In *The Semantic Web*, Karl Aberer, Key-Sun Choi, Natasha Noy, Dean Allemang, Kyung-Il Lee, Lyndon Nixon, Jennifer Golbeck, Peter Mika, Diana Maynard, Riichiro Mizoguchi, Guus Schreiber, and Philippe Cudré-Mauroux (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 843–856.

[17] 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, CoNLL 2003, Held in cooperation with HLT-NAACL 2003, Edmonton, Canada, May 31 - June 1, 2003*. 142–147. https://aclanthology.org/W03-0419/

[18] Mourad Sarrouti, Asma Ben Abacha, Yassine Mrabet, and Dina Demner-Fushman. 2021. Evidence-based Fact-Checking of Health-related Claims. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (Eds.). Association for Computational Linguistics, Punta Cana, Dominican Republic, 3499–3512. https://doi.org/10.18653/v1/2021.findings-emnlp.297

[19] Qingyu Tan, Lu Xu, Lidong Bing, Hwee Tou Ng, and Sharifah Mahani Aljunied. 2022. Revisiting DocRED - Addressing the False Negative Problem in Relation Extraction. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (Eds.). Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 8472–8487. https://doi.org/10.18653/v1/2022.emnlp-main.580

[20] Qingyu Tan, Lu Xu, Lidong Bing, Hwee Tou Ng, and Sharifah Mahani Aljunied. 2022. Revisiting DocRED-addressing the false negative problem in relation extraction. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. 8472–8487.

[21] 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 *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*. 5998–6008. https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

[22] Jin Wang, Yuliang Li, and Wataru Hirota. 2021. Machamp: A Generalized Entity Matching Benchmark. In *CIKM '21: The 30th ACM International Conference on Information and Knowledge Management, Virtual Event, Queensland, Australia, November 1 - 5, 2021*. 4633–4642. https://doi.org/10.1145/3459637.3482008

[23] Ledell Wu, Fabio Petroni, Martin Josifoski, Sebastian Riedel, and Luke Zettlemoyer. 2020. Scalable Zero-shot Entity Linking with Dense Entity Retrieval. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*. 6397–6407. https://doi.org/10.18653/v1/2020.emnlp-main.519

[24] Xingjiao Wu, Luwei Xiao, Yixuan Sun, Junhang Zhang, Tianlong Ma, and Liang He. 2022. A survey of human-in-the-loop for machine learning. *Future Gener. Comput. Syst.* 135 (2022), 364–381. https://doi.org/10.1016/j.future.2022.05.014

[25] Ikuya Yamada, Akari Asai, Hiroyuki Shindo, Hideaki Takeda, and Yuji Matsumoto. 2020. LUKE: Deep Contextualized Entity Representations with Entity-aware Self-attention. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (Eds.). Association for Computational Linguistics, Online, 6442–6454. https://doi.org/10.18653/v1/2020.emnlp-main.523

[26] Yuan Yao, Deming Ye, Peng Li, Xu Han, Yankai Lin, Zhenghao Liu, Zhiyuan Liu, Lixin Huang, Jie Zhou, and Maosong Sun. 2019. DocRED: A Large-Scale Document-Level Relation Extraction Dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Anna Korhonen, David Traum, and Lluís Màrquez (Eds.). Association for Computational Linguistics, Florence, Italy, 764–777. https://doi.org/10.18653/v1/P19-1074

[27] Juntao Yu, Bernd Bohnet, and Massimo Poesio. 2020. Named Entity Recognition as Dependency Parsing. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*. 6470–6476. https://doi.org/10.18653/v1/2020.acl-main.577

[28] Klim Zaporojets, Johannes Deleu, Chris Develder, and Thomas Demeester. 2021. DWIE: An entity-centric dataset for multi-task document-level information extraction. *Inf. Process. Manag.* 58, 4 (2021), 102563. https://doi.org/10.1016/j.ipm.2021.102563

[29] Ningyu Zhang, Xin Xu, Liankuan Tao, Haiyang Yu, Hongbin Ye, Shuofei Qiao, Xin Xie, Xiang Chen, Zhoubo Li, and Lei Li. 2022. DeepKE: A Deep Learning Based Knowledge Extraction Toolkit for Knowledge Base Population. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, Wanxiang Che and Ekaterina Shutova (Eds.). Association for Computational Linguistics, Abu Dhabi, UAE, 98–108. https://doi.org/10.18653/v1/2022.emnlp-demos.10

[30] Wenxuan Zhou, Kevin Huang, Tengyu Ma, and Jing Huang. 2021. Document-Level Relation Extraction with Adaptive Thresholding and Localized Context Pooling. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*. AAAI Press, 14612–14620. https://doi.org/10.1609/AAAI.V35I16.17717

[31] Enwei Zhu, Yiyang Liu, and Jinpeng Li. 2023. Deep Span Representations for Named Entity Recognition. arXiv:2210.04182 [cs.CL]