Impact of OCR Quality on Named Entity Linking

Elvys Linhares Pontes, Ahmed Hamdi, Nicolas Sidère, Antoine Doucet

Impact of OCR Quality on Named Entity Linking

Elvys Linhares Pontes[1,2], Ahmed Hamdi[2], Nicolas Sidere[2], and Antoine Doucet[2]

[1] University of Avignon, Avignon, France
elvys.linhares-pontes@univ-avignon.fr
[2] University of La Rochelle, La Rochelle, France
firstname.lastname@univ-lr.fr

Abstract. Digital libraries are online collections of digital objects that can include text, images, audio, or videos. It has long been observed that named entities (NEs) are key to the access to digital library portals as they are contained in most user queries. Combined or subsequent to the recognition of NEs, named entity linking (NEL) connects NEs to external knowledge bases. This allows to differentiate ambiguous geographical locations or names (John Smith), and implies that the descriptions from the knowledge bases can be used for semantic enrichment. However, the NEL task is especially challenging for large quantities of documents as the diversity of NEs is increasing with the size of the collections. Additionally digitized documents are indexed through their OCRed version which may contains numerous OCR errors. This paper aims to evaluate the performance of named entity linking over digitized documents with different levels of OCR quality. It is the first investigation that we know of to analyze and correlate the impact of document degradation on the performance of NEL. We tested state-of-the-art NEL techniques over several evaluation benchmarks, and experimented with various types of OCR noise. We present the resulting study and subsequent recommendations on the adequate documents and OCR quality levels required to perform reliable named entity linking.

Keywords: Named Entity Linking · Deep Learning · Digital Library · Indexing.

1 Introduction

A named entity is a real-world object, such as persons, locations, organizations, etc. Named entities have been shown to be key to digital library access as they are contained in a majority of the search queries submitted to digital library portals. They were notably found in 80% of queries submitted to Gallica, the portal of the national library of France [2].

Collecting data from different sources leads to reveal the problem of duplicate and ambiguous information about named entities. Therefore they are often not

distinctive since one single name may correspond to several entities. A disambiguation process is thus essential to distinguish named entities to be indexed in digital libraries.

Named Entity Linking (NEL) is the task of recognizing and disambiguating named entities by linking them to entries of a Knowledge Base (KB). Knowledge bases (e.g. Wikipedia[3], DBpedia[15], YAGO[23], and Freebase[1]) contain rich information about the worlds entities, their semantic classes, and their mutual relationships. NEL is a challenging task because a named entity may have multiple surface forms, such as its full name, partial names, aliases, abbreviations, and alternate spellings [22]. Besides digital libraries, this task is important to several NLP applications, e.g. information extraction, information retrieval (for the adequate retrieval of ambiguous information), content analysis (for the analysis of the general content of a text in terms of its topics, ideas or categorizations), question answering and knowledge base population.

In a nutshell, NEL aims to locate mentions of an NE, and to accurately link them to the right entry of a knowledge base, a process that often requires disambiguation (see Figure 1). A NEL system typically performs two tasks: named entity recognition (NER) and entity disambiguation (ED). NER extracts entities in a document, and ED links these entities to their corresponding entities in a KB. Until recently, the common approach of popular systems was to solve these two sub-problems independently. However, the significant dependency between these two tasks is ignored and errors caused by NER will propagate to the ED without the possibility of recovery. Therefore, recent approaches [13] propose the joint analysis of these sub-tasks in order to reduce the amount of errors.

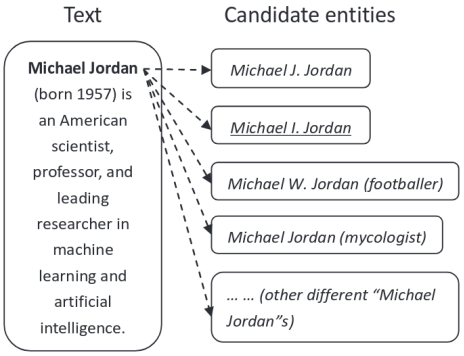

Fig. 1. An example of the named entity linking task. The named entity mention detected from the text is in bold face and the correct mapping entity is underlined (Shen *et al.* [22]).

NEL in digital libraries is especially challenging due to the fact that most digitized documents are indexed through their OCRed version. This causes nu-

[3] http://www.wikipedia.org/

merous errors due to the state of documents, following aging, bad storage conditions and/or the poor quality of initial printing materials. For instance, Chiron et al. [2] analyzed a collection of OCRed documents with 12M characters from 9 sources written in 2 languages. This collection is composed of documents from 1654-2000 and contains error rates that vary from 1% to 4%.

This paper aims to analyze the impact of OCR quality on the NEL task[4]. Unfortunately, available corpora for NEL do not contain the noise presented in digital libraries. In order to overcome this problem, we simulated various OCRed versions of available data sets for NEL. Then, we tested state-of-the-art systems to investigate the impact of different OCR noises in NEL performance. The impact of OCR noise on NEL performance was as we expected: the greater the word and character errors, the greater the degradation of linking named entities to a knowledge base. However, we are now able to provide the DL community with recommendations on the OCR quality that is required for a given level of expected NEL performance.

Interestingly, the analyzed NEL approaches generated similar results for the character and the combination of all OCR degradation. This indicates that these systems have been able to analyze different OCR degradation without significantly reducing their performance.

The remainder of the paper is organized as follows : Section 2 makes a brief overview of the most recent and available NEL approaches in the state of the art. Then, available resources and our experimental evaluation are respectively described in Sections 3 and 4. The paper is concluded in Section 5.

2 An overview of named entity linking

Given a knowledge base containing a set of named entities and a set of documents, the goal of named entity linking is to map each named entity in these documents to its corresponding named entity in a knowledge base [22]. NEL approaches can be divided into two classes:

- **Disambiguation approaches** only analyze gold standard named entities in a document and disambiguates them to the correct entry in a given KB [6, 14, 20].
- **End-to-end approaches** extract candidate entities from documents and then disambiguate them to the correct entries in a given KB [13].

Despite these two different classes, most works in the state of the art are based on three modules: candidate entity generation, candidate entity ranking, and unlinkable mention prediction [22]. More precisely, the first module aims to retrieve related entity mentions in KB that refer to mention in a document. Several works are based on name dictionary-based techniques [7], surface form

[4] To the best of our knowledge, no studies have been conducted on NEL using OCRed documents.

expansion from the local document [25], and methods based on search engine [9] to identify candidate entities.

After selecting candidate entities, the second module attempts to rank the most likely link in KB for a mention. The main approaches are mainly based on supervised and unsupervised methods. These methods consider various techniques to analyze and rank entities, e.g. name string comparison [26], entity popularity [7], entity type [4], textual context [16], and coherence between mapping entities [3]. Finally, the last module validates whether the top-ranked entity identified in the candidate entity ranking module is the target entity for a mention.

Recent neural network methods [6, 14] have established state-of-the-art results, out-performing engineered features based models. These methods use deep learning[5] to analyze features, relationships, and complex interactions among features which allow a better analysis of documents and improve their performance. These methods combine context-aware word, span and entity embeddings with neural similarity functions to analyze the context of mentions and disambiguate correctly them to a KB.

Next subsections describe relevant and available NEL systems. Section 2.1 makes a brief description of disambiguation approaches and Section 2.2 focuses on the end-to-end approaches.

2.1 Disambiguation approaches

Disambiguation approaches consider having already identified the named entities in the documents. In this case, these approaches aim to analyze the context of these entities to disambiguate them in a KB. In this context, Ganea and Hofmann [6] proposed a deep learning model for joint document-level entity disambiguation[6]. In a nutshell, they embed entities and words in a common vector space and use a neural attention mechanism over local context windows to select words that are informative for the disambiguation decision. Their model contains a conditional random field that collectively disambiguates the mentions in a document (Figure 2).

Inspired on the Ganea and Hofmann's approach [6], Le and Titov [14] treated relations between mentions as latent variables in their neural NEL model[7]. They rely on representation learning and learn embeddings of mentions, contexts, and relations in order to reduce the amount of human expertise required to construct the system and make the analysis more portable across languages and domains.

Raiman and Raiman [20] proposed a system for integrating symbolic knowledge into the reasoning process of a neural network through a type system[8]. They constrained the behavior to respect the desired symbolic structure, and

[5] Including all types of deep learning techniques, such as transfer, reinforcement and multi-task learning.
[6] The code is publicly available: https://github.com/dalab/deep-ed
[7] The code is publicly available: https://github.com/lephong/mulrel-nel
[8] The code is publicly available: https://github.com/openai/deeptype

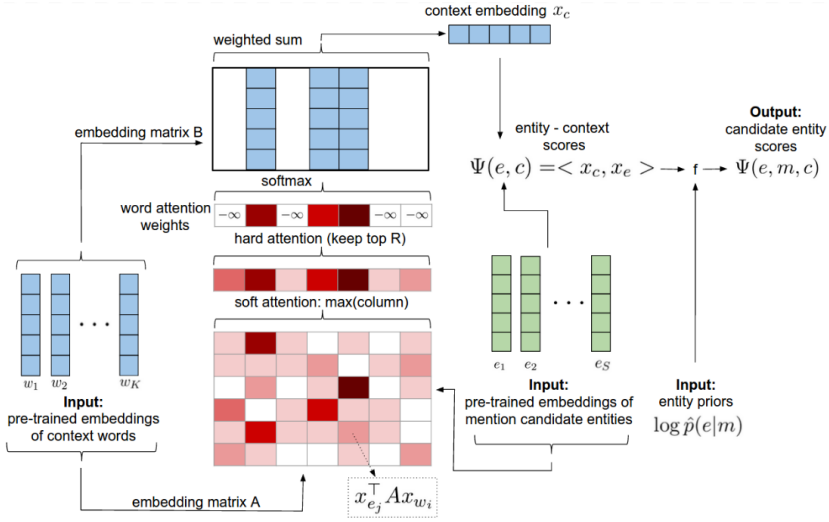

Fig. 2. Architecture of the Ganea and Hofmann's approach. Their method uses a local model with neural attention to process context word vectors, candidate entity priors, and embeddings to generates the candidate entity scores (Ganea and Hofmann [6]).

automatically design the type system without human effort. Their model first uses heuristic search or stochastic optimization over discrete variables that define a type system informed by an Oracle and a learnability heuristic. Then, classifier parameters are fitted using gradient descent.

2.2 End-to-end approach

Following the idea of jointly analyzing the NER and ED tasks, Kolitsas *et al.* [13] proposed a neural end-to-end NEL system that jointly discovers and links entities in a document[9]. Their model replaces engineered features by neural embeddings. They first generate all possible spans (mentions) that have at least one possible entity candidate. Then, each mention-candidate pair receives a context-aware compatibility score based on word and entity embeddings [6] coupled with neural attention and a global voting mechanism.

3 Resources

To the best of our knowledge, there is no publicly available corpora in the literature that are addressed to both named entity linking and post-OCR correction. There are either corpora where named entities are well recognized and linked, but the text is not noisy or conversely, there are corpora where the text generated by an OCR process is aligned with the original text, but named entities

[9] The code is publicly available: https://github.com/dalab/end2end_neural_el

are not annotated. Therefore we started from existing NEL corpora to build their versions in noisy context. We present in the following two subsections the available NEL corpora used in this work and a method to synthesize and inject OCR degradation into these corpora.

3.1 Original data sets

Available corpora for entity linking are mainly divided into six data sets:

- **AIDA-CoNLL** data set [10] is based on CoNLL 2003 data that was used for NER task. This data set is divided into AIDA-train for training, AIDA-A for validation, and AIDA-B for testing. This data set contains 1393 Reuters news articles and 27817 linkable mentions.
- **AQUAINT** data set [18, 8] is composed of 50 short news documents (250-300 words) from the Xinhua News Service, the New York Times, and the Associated Press. This data set contains 727 mentions.
- **ACE2004** data set [21, 8] is a subset of the ACE2004 coreference documents with 57 articles and 306 mentions, annotated through crowdsourcing.
- **MSNBC** data set [3, 8] is composed of 20 news articles from 10 different topics (two articles per topic: Business, U.S. Politics, Entertainment, Health, Sports, Tech & Science, Travel, TV News, U.S. News, and World News), having 656 linkable mentions in total.
- Finally, **CWEB** [5] and **WIKI** [21] data sets are composed of 320 documents each.

3.2 Simulated data sets

As we mentioned above, our work faces the lack of appropriate resources for named entity linking over OCRed documents. In order to overcome this problem, we started with corpora described in section 3.1 and we then injected to each of them many text degradation caused by an OCR engine. We have built such ground truth as follow:

- clean texts have been converted into images
- images are contaminated using common degradation related to storage conditions and use of scanners [12]
- an OCR system is used to extract the text (we have used tesseract open source OCR engine v-4.0.0[10]).

We have used the DocCreator tool [12] to add four common types of OCR degradation that may be present on digital libraries material:

- **bleed-through** is typical degradation of double-sided pages. It simulates the ink from the back side that seeps through the front side.

[10] https://github.com/tesseract-ocr

- **phantom degradation** simulates degradation in worn documents. Following successive uses, some characters can be progressively eroded. The digitization process generates phantom ink around characters.
- **character degradation** simulates degradation due to the age of the document or the use of a scanner incorrectly set. It consists in adding small ink spots on characters and can induce the erasure of characters.
- **blurring** simulates a blurring effect, as can be encountered during a typical digitization process with focus issue.

These degradation allowed building four versions for each NEL corpus. We additionally define two versions that we call respectively LEV-0 and LEV-MIX. The LEV-0 version corresponds to re-OCRred version of original images with no degradation added. It aims to evaluate the OCR engine through sharp images. Whereas the LEV-MIX version is the result of simultaneously applying the four types of degradation to the original texts[11]. Figure 3 shows the impact of the different types of degradation on the sharpness of images comparing to the original one.

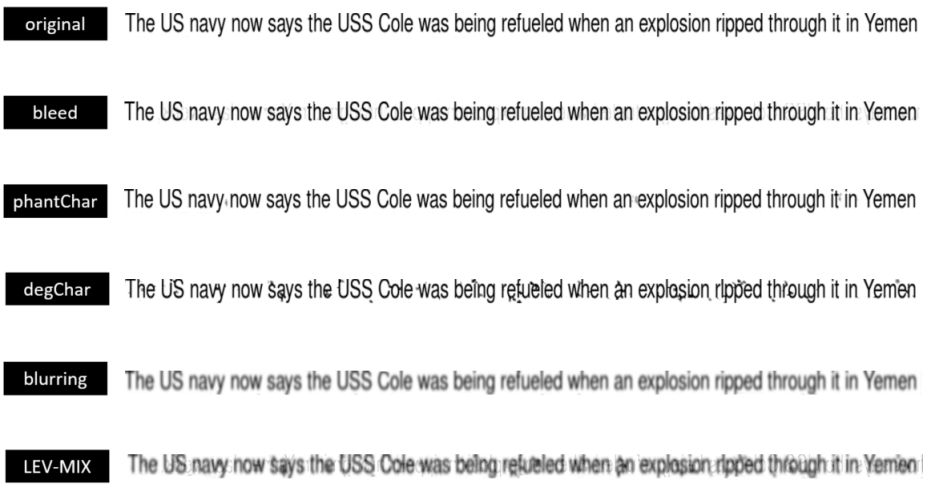

Fig. 3. Original image versus degraded images.

Following the text extraction by the OCR, noisy text have been aligned to its original version using the tool RETAS [24]. This tool allows to make this correspondence at character and word levels. Figure 4 shows an example of such alignment made between the ground truth and its OCRed version. The alignment reflects the various errors made by the OCR engine. The difference

[11] Real-world documents contain simultaneously several OCR degradations. Therefore, the LEV-MIX version represents better the degraded documents on digital libraries.

between the two texts are denoted by the presence of the character @. Each @ in the ground truth indicates the insertion of one character by the OCR while @ in the noisy text indicates that one character has been deleted from the original text. All data used in this work are available for public by following this link : "http://bit.ly/2Icag5E". This link provides for each test corpus the degraded images, the noisy texts extracted by the OCR and their aligned version with clean data at the word and the character levels.

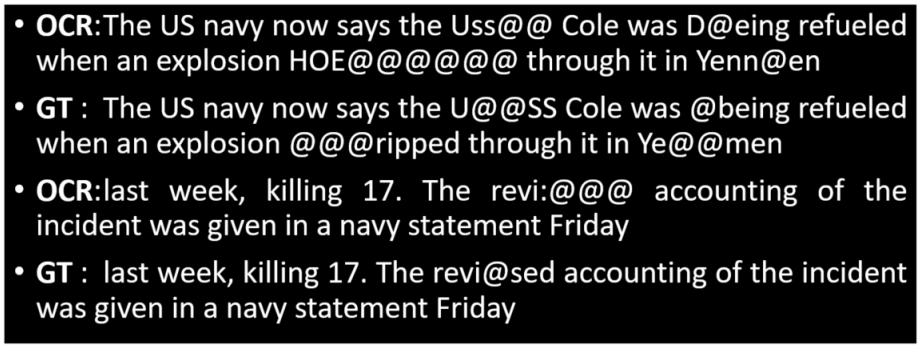

Fig. 4. Text alignment: ground truth and OCR output.

In order to evaluate the quality of the text generated by an OCR engine, one of the most popular measures is Character Error Rate (CER) [11]. The CER is the proportion of errors made on characters compared to the original text. There are three types of errors:

- **Substitution** where one character has been replaced by another.
- **Deletion** where a character has simply not been recognized by the OCR.
- **Insertion** where an additional character has been wrongly added.

Table 1 details the OCR error rates at the character levels in the different data sets.

OCR deg.	AIDA	ACE2004	AQUAINT	CWEB	MSNBC	WIKI	Average
LEV-0	1.0	0.8	0.4	1.3	1.1	0.6	0.9
Bleed	1.0	0.7	0.4	1.2	0.3	0.6	0.7
Blur	1.8	1.3	0.6	2.6	0.9	1.1	1.4
CharDeg	3.4	2.5	2.1	3.0	2.3	2.5	2.6
PhantChar	1.1	0.8	0.5	1.3	0.5	0.7	0.8
LEV-MIX	4.8	4.8	2.5	5.4	3.1	3.2	4.0

Table 1. CER on the OCRed versions of data sets.

The CER varies between 0.4% and 5.4% according to the documents and the type of degradation added. However, the noise distributed on the documents is homogeneous. It leads to a higher rate of error on words. The Word Error Rate (WER) is another measure used to translate the quality of a corpus [17]. Unlike the CER, this measure is the rate of correctly recognized words. Two words are different if they differ in at least one character. Table 2 describes the WER error rates in the different corpora.

OCR deg.	AIDA	ACE2004	AQUAINT	CWEB	MSNBC	WIKI	Average
LEV-0	3.9	3.0	2.0	3.1	2.3	3.0	2.9
Bleed	3.9	3.0	1.9	2.7	1.6	3.0	2.7
Blur	5.4	3.8	2.2	5.8	3.2	3.5	4.0
CharDeg	16.9	14.4	13.4	13.6	14.3	14.1	14.4
PhantChar	5.6	4.6	4.0	4.3	3.4	4.5	4.4
LEV-MIX	18.2	15.4	13.7	16.6	15.4	14.4	15.6

Table 2. WER on the OCRed versions of data sets.

As shown in Tables 1 and 2, among OCR degradation, the character degradation showed to be the most critical noise by generating the highest error rate at the character and word levels. Interesting, the bleed version of data sets generated similar or smaller CER and WER values than their LEV-0 version. Finally, the combination of OCR degradation (LEV-MIX) achieved the worst CER and WER.

4 Experimental evaluation

In this section, we introduce the results of our experiments with different types of injected noise.

4.1 Automatic evaluation

The main evaluation measures for entity linking systems are precision, recall and F1-measure. Precision is the fraction of correctly linked entity mentions that are generated by a system. Recall takes into account all entity mentions that should be linked and determines how correct linked entity mentions are with regard to total entity mentions that should be linked. Finally, the F1-measure is defined as the harmonic mean of precision and recall.

These measures can be calculated on a full corpus (micro-averaging) or averaged by document (macro-averaging)[12]. Since knowledge bases contain millions of entities, for simplicity, we focused on mention-entity pairs for which the ground-truth related to known entity.

[12] In this paper, we relied on the micro F1 scores.

4.2 Training settings

In order to analyze the effects of OCR degradation, we tested the systems of Ganea and Hofmann 2017's [6] and Le and Titov's [14]. We used the pre-trained models of each system. More precisely, Ganea and Hofmann trained their entity embeddings on the Wikipedia (Feb 2014) corpus and they used the pre-trained Word2Vec word vectors with 300 dimensions[13]. Le and Titov used GloVe [19] word embeddings trained on 840B tokens and they selected the best 7 relations on the development scores. Both of them trained their systems on the AIDA-train data set.

4.3 Entity linking over clean and noisy data

Tables 3 and 4 respectively show the results of the systems of Ganea and Hofmann's [6] and Le and Titov's [14] on the OCRed data sets. As we expected, the CER and WER generated by OCR degradation are roughly correlated to the performance of NEL. LEV-0 and bleed data sets have low CER and WER levels which produced less impact on the F1 results. Character degradation has larger CER and WER values which caused a bigger drop in the NEL performance. Among all data sets, ACE2004 was the most affected by OCR degradation (a drop of almost 20 percentage points for the character degradation).

OCR deg.	AIDA	ACE2004	AQUAINT	CWEB	MSNBC	WIKI	Average
Clean	0.9144	0.8893	0.9021	0.7774	0.9350	0.7799	0.8663
LEV-0	0.9039	0.7906	0.8867	0.7372	0.9168	0.7491	0.8307
Bleed	0.9039	0.7906	0.8906	0.7348	0.9171	0.7490	0.8310
Blur	0.9044	0.7934	0.8856	0.7237	0.9106	0.7379	0.8259
CharDeg	0.9016	0.6873	0.7902	0.6938	0.8498	0.6597	0.7637
PhantChar	0.9039	0.7867	0.8802	0.7344	0.9101	0.7462	0.8269
LEV-MIX	0.9036	0.6856	0.8043	0.6798	0.8456	0.6560	0.7625

Table 3. NEL results (micro F1 scores) using the Ganea and Hofmann 2017's system [6] on the OCRed versions of data sets.

The LEV-MIX version of the data sets generated the highest CER and WER values with all kinds of character and word errors presented in the four OCR degradations. Despite these high error rates, the combination of these OCR degradations does not appear to have a significant change in the performance of two NEL systems. Both systems achieved satisfactory results (average micro F1-score greater than 0.75) for all OCRed versions.

The analysis of the relations of mentions as latent variables improved the analysis of candidates mentions and increased F1 scores for almost all clean

[13] http://bit.ly/1R9Wsqr

OCR deg.	AIDA	ACE2004	AQUAINT	CWEB	MSNBC	WIKI	Average
Clean	0.9306	0.9014	0.8923	0.7773	0.9411	0.7783	0.8702
LEV-0	0.9071	0.8000	0.8683	0.7434	0.9314	0.7446	0.8325
Bleed	0.9091	0.8000	0.8723	0.7415	0.9318	0.7443	0.8332
Blur	0.9037	0.7934	0.8671	0.7282	0.9253	0.7344	0.8253
CharDeg	0.8984	0.7028	0.7795	0.6931	0.8610	0.6585	0.7655
PhantChar	0.9071	0.7915	0.8602	0.7401	0.9270	0.7393	0.8275
LEV-MIX	0.8985	0.6959	0.7936	0.6838	0.8613	0.6549	0.7647

Table 4. NEL results (micro F1 scores) using the Le and Titov's system [14] on the OCRed versions of data sets.

data sets (except AQUAINT and WIKI data sets). However, errors produced by the OCR degradation were more critical for this analysis. Indeed, the mentions affected by OCR errors degraded the analysis of the relationships between other mentions that led to a decrease in the performance of this approach.

In order to better evaluate the impact of OCR quality on NEL results, we calculated for both systems used in this work the δ measure which represents the average decrease rate between the F1-score given in clean data and the F1-scores given in noisy data for each degradation. For both systems, δ do not exceed 11% in noisy data. The maximum decrease of NEL F1-score is given by the LEV-MIX degradation.

Figure 5 shows the evolution of the δ measure according to degradation. Types of degradation have been sorted according to OCR rates. CER and WER curves are also given for comparative reasons. We refer by δ_1 and δ_2 the decrease measures of NEL F1-scores given by Ganea and Hofmann system and Le and Titov's system respectively. Both systems achieved good NEL performances (a drop of up to 0.05 in F-score values) when the WER is below 5%. For higher WER values, the loss of NEL performance doubled.

Despite the complexity of the NEL task and the occurrence of several types of errors in the documents of digital libraries, the systems achieved interesting results. This proves that they can be used to distinguish ambiguous named entities in degraded documents. Some word correction strategies, such as auto-encoders, language models, and so on, could be used to decrease the impact of OCR degradation on NEL.

5 Conclusion

This paper is the first attempt that we know of that ever investigated the impact of document degradation on the task of named entity linking.

The broad collection of topics that may be covered in digital libraries implies that simply recognizing named entities is not sufficient to identify them. Named entity linking can contribute to solving this problem by analyzing the context of these entities and disambiguating them thanks to a knowledge base. We have

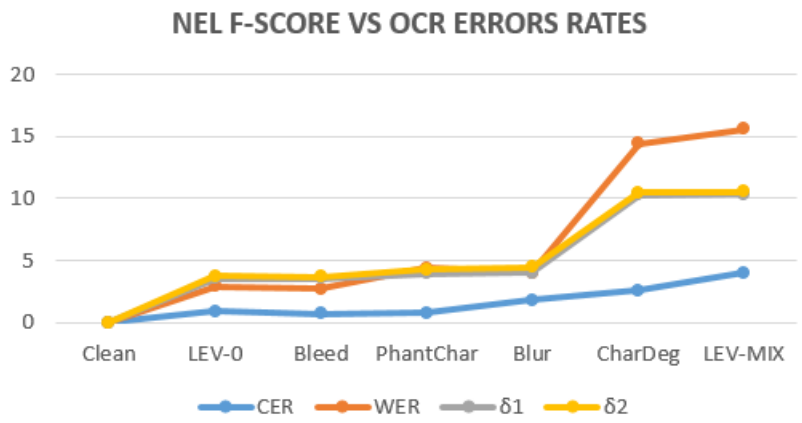

Fig. 5. NER F-score degradation according to OCR error rates

tested two available NEL systems on six available data sets with OCR degradation. Despite the character and word error rates of OCR degradation, these two systems achieved satisfying results for NEL. Some of these OCR degradations generated high error rates, which reduced system performance.

We were able to provide the DL community with recommendations on the OCR quality that is required for a given level of expected NEL performance.

Further work is under progress to analyze the performance of end-to-end NEL systems on OCRed data sets. More precisely, we want to investigate how different approaches can overcome the problem of OCR degradation and provide correct predictions. We also intend to analyze and to test the performance of these systems using real data on low-resources languages.

Acknowledgments

This work has been supported by the European Unions Horizon 2020 research and innovation program under grant 770299 (NewsEye).

References

1. Bollacker, K., Evans, C., Paritosh, P., Sturge, T., Taylor, J.: Freebase: A collaboratively created graph database for structuring human knowledge. In: Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data. pp. 1247–1250. SIGMOD '08, ACM, New York, NY, USA (2008). https://doi.org/10.1145/1376616.1376746, http://doi.acm.org/10.1145/1376616.1376746

2. Chiron, G., Doucet, A., Coustaty, M., Visani, M., Moreux, J.P.: Impact of ocr errors on the use of digital libraries: towards a better access to information. In: Proceedings of the 17th ACM/IEEE Joint Conference on Digital Libraries. pp. 249–252. IEEE Press (2017)
3. Cucerzan, S.: Large-scale named entity disambiguation based on Wikipedia data. In: Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL). pp. 708–716. Association for Computational Linguistics, Prague, Czech Republic (Jun 2007), https://www.aclweb.org/anthology/D07-1074
4. Dredze, M., McNamee, P., Rao, D., Gerber, A., Finin, T.: Entity disambiguation for knowledge base population. In: Proceedings of the 23rd International Conference on Computational Linguistics. pp. 277–285. COLING '10, Association for Computational Linguistics, Stroudsburg, PA, USA (2010), http://dl.acm.org/citation.cfm?id=1873781.1873813
5. Gabrilovich, E., Ringgaard, M., Subramanya, A.: FACC1: Freebase annotation of ClueWeb corpora, Version 1 (Release date 2013-06-26, Format version 1, Correction level 0) (Jun 2013)
6. Ganea, O.E., Hofmann, T.: Deep joint entity disambiguation with local neural attention. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. pp. 2619–2629. Association for Computational Linguistics (2017). https://doi.org/10.18653/v1/D17-1277, http://aclweb.org/anthology/D17-1277
7. Guo, S., Chang, M.W., Kiciman, E.: To link or not to link? a study on end-to-end tweet entity linking. In: Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 1020–1030. Association for Computational Linguistics, Atlanta, Georgia (Jun 2013), https://www.aclweb.org/anthology/N13-1122
8. Guo, Z., Barbosa, D.: Robust entity linking via random walks. In: Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management. pp. 499–508. CIKM '14, ACM, New York, NY, USA (2014). https://doi.org/10.1145/2661829.2661887, http://doi.acm.org/10.1145/2661829.2661887
9. Han, X., Zhao, J.: Nlpr_kbp in tac 2009 kbp track: A two-stage method to entity linking. In: In Proceedings of Test Analysis Conference 2009 (TAC 09). MIT Press (1999)
10. Hoffart, J., Yosef, M.A., Bordino, I., Fürstenau, H., Pinkal, M., Spaniol, M., Taneva, B., Thater, S., Weikum, G.: Robust disambiguation of named entities in text. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. pp. 782–792. EMNLP '11, Association for Computational Linguistics, Stroudsburg, PA, USA (2011), http://dl.acm.org/citation.cfm?id=2145432.2145521
11. Holley, R.: How good can it get? analysing and improving ocr accuracy in large scale historic newspaper digitisation programs. D-Lib Magazine **15**(3/4) (2009)
12. Journet, N., Visani, M., Mansencal, B., Van-Cuong, K., Billy, A.: Doccreator: A new software for creating synthetic ground-truthed document images. Journal of imaging **3**(4), 62 (2017)
13. Kolitsas, N., Ganea, O.E., Hofmann, T.: End-to-end neural entity linking. In: Proceedings of the 22nd Conference on Computational Natural Language Learning. pp. 519–529. Association for Computational Linguistics (2018), http://aclweb.org/anthology/K18-1050

14. Le, P., Titov, I.: Improving entity linking by modeling latent relations between mentions. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 1595–1604. Association for Computational Linguistics (2018), http://aclweb.org/anthology/P18-1148
15. Lehmann, J., Isele, R., Jakob, M., Jentzsch, A., Kontokostas, D., Mendes, P.N., Hellmann, S., Morsey, M., van Kleef, P., Auer, S., Bizer, C.: DBpedia - a large-scale, multilingual knowledge base extracted from wikipedia. Semantic Web Journal **6**(2), 167–195 (2015), http://jens-lehmann.org/files/2015/swj_dbpedia.pdf
16. Li, Y., Wang, C., Han, F., Han, J., Roth, D., Yan, X.: Mining evidences for named entity disambiguation. In: Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 1070–1078. KDD '13, ACM, New York, NY, USA (2013). https://doi.org/10.1145/2487575.2487681, http://doi.acm.org/10.1145/2487575.2487681
17. Lund, W.B., Kennard, D.J., Ringger, E.K.: Combining multiple thresholding binarization values to improve ocr output. In: Document Recognition and Retrieval XX. vol. 8658, p. 86580R. International Society for Optics and Photonics (2013)
18. Milne, D., Witten, I.H.: Learning to link with wikipedia. In: Proceedings of the 17th ACM Conference on Information and Knowledge Management. pp. 509–518. CIKM '08, ACM, New York, NY, USA (2008). https://doi.org/10.1145/1458082.1458150, http://doi.acm.org/10.1145/1458082.1458150
19. Pennington, J., Socher, R., Manning, C.: Glove: Global vectors for word representation. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). pp. 1532–1543 (2014)
20. Raiman, J., Raiman, O.: Deeptype: Multilingual entity linking by neural type system evolution. In: Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018. pp. 5406–5413 (2018), https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/17148
21. Ratinov, L., Roth, D., Downey, D., Anderson, M.: Local and global algorithms for disambiguation to wikipedia. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1. pp. 1375–1384. HLT '11, Association for Computational Linguistics, Stroudsburg, PA, USA (2011), http://dl.acm.org/citation.cfm?id=2002472.2002642
22. Shen, W., Wang, J., Han, J.: Entity linking with a knowledge base: Issues, techniques, and solutions. IEEE Transactions on Knowledge and Data Engineering **27**(2), 443–460 (Feb 2015). https://doi.org/10.1109/TKDE.2014.2327028
23. Suchanek, F.M., Kasneci, G., Weikum, G.: Yago: A core of semantic knowledge. In: Proceedings of the 16th International Conference on World Wide Web. pp. 697–706. WWW '07, ACM, New York, NY, USA (2007). https://doi.org/10.1145/1242572.1242667, http://doi.acm.org/10.1145/1242572.1242667
24. Yalniz, I.Z., Manmatha, R.: A fast alignment scheme for automatic ocr evaluation of books. In: Document Analysis and Recognition (ICDAR), 2011 International Conference on. pp. 754–758. IEEE (2011)
25. Zhang, W., Sim, Y.C., Su, J., Tan, C.L.: Entity linking with effective acronym expansion, instance selection and topic modeling. In: Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence - Volume Volume Three. pp. 1909–1914. IJCAI'11,

AAAI Press (2011). https://doi.org/10.5591/978-1-57735-516-8/IJCAI11-319, http://dx.doi.org/10.5591/978-1-57735-516-8/IJCAI11-319
26. Zheng, Z., Li, F., Huang, M., Zhu, X.: Learning to link entities with knowledge base. In: Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics. pp. 483–491. HLT '10, Association for Computational Linguistics, Stroudsburg, PA, USA (2010), http://dl.acm.org/citation.cfm?id=1857999.1858071