A Named Entity Extraction System for Historical Financial Data

Wassim Swaileh, Thierry Paquet, Sébastien Adam, Andres Rojas Camacho

A Named Entity Extraction System for Historical Financial Data

Chapter · August 2020

DOI: 10.1007/978-3-030-57058-3_23

Wassim Swaileh
Université de Cergy-Pontoise

Thierry Paquet
Université de Rouen

A Named Entity Extraction System for Historical Financial Data

Wassim Swaileh[2], Thierry Paquet[1], Sébastien Adam[1], and Andres Rojas Camacho[1]

[1] LITIS EA4108, University of Rouen Normandie, France
first.lastname@univ-rouen.fr
[2] ETIS, UMR 8051, CY Cergy Paris Université, ENSEA, CNRS , France
fisrt.lastname@cyu.fr

Abstract. Access to long-run historical data in the field of social sciences, economics and political sciences has been identified as one necessary condition to understand the dynamics of the past and the way those dynamics structure our present and future. Financial yearbooks are historical records reporting on information about the companies of stock exchanges. This paper concentrates on the description of the key components that implement a financial information extraction system from financial yearbooks. The proposed system consists in three steps: OCR, linked named entities extraction, active learning. The core of the system is related to linked named entities extraction (LNE). LNE are coherent n-tuple of named entities describing high level semantic information. In this respect we developed, tested and compared a CRF and a hybrid RNN/CRF based system. Active learning allows to cope with the lack of annotated data for training the system. Promising performance results are reported on two yearbooks (the French Desfossé yearbook (1962) and the German Handbuch (1914-15)) and for two LNE extraction tasks: capital information of companies and constitution information of companies.

Keywords: Linked named entities extraction · Active learning · CRF · String embedding.

1 Introduction

Access to long-run historical data in the field of social sciences, economics and political sciences has been identified as one necessary condition to understand the dynamics of the past and the way those dynamics structure our present and future. In this context, the EURHISFIRM project has been founded by the EU to develop a Research Infrastructure with a focal point on the integration of financial and corporate governance historical information of firms. During its design phase, one of the objectives of EURHISFIRM is to design and develop an intelligent and collaborative system for the extraction and enrichment of data from historical paper sources such as yearbooks and price lists that were published over years in the many European stock exchanges.

In this paper, we focus on the processing of yearbooks. Figure 1 shows two examples of such documents, one German (GR) and one French (FR). These yearly publications were intended to provide updated information about the companies of the stock exchanges, including their name, date of creation, financial status, governing board members, headquarters address, branch address, financial information such as capital amount, date and amount of capital increase, balance sheet of the year including assets and liabilities, etc... As shown on these two examples, Yearbooks have mostly textual contents organised in specific sections, on the contrary of prices lists that contain tabular data. As a consequence, extracting information from yearbooks requires the design of a general named entity extraction system from OCRed yearbooks. This represents a real complex challenge related to document image segmentation, optical character recognition (OCR) and linked named entity recognition (NER).

This paper concentrates on the description of a key component of the EURHISFIRM platform that implements a financial information extraction system from Yearbooks. The rest of this paper is organised as follows: Section 2 gives a brief overview of related works. The system architecture is then described in section 3. Section 4 emphasises on the key component of the system: the Linked Financial Named Entities (LFNE) extraction. In this purpose we study a Conditional Random Field (CRF) based approach and a hybrid recurrent Neural Networks / CRF (RNN-CRF) based approach. Then, in section 5, we report on the system performance on two specific sections of the two Yearbooks under study. We also analyse the system's performance when introducing an active learning strategy in order to cope with the lack of annotated training data.

2 Related work

Information Extraction (IE) from born digital texts has been extensively studied in the field of Natural Language Processing (NLP), notably through the well known Message Understanding Conferences (MUC) that were organised from 1987 to 1997. In 1995, the first competition on Named Entity Recognition (NER) was introduced during MUC-6[3]. Since 1999, the yearly conference on Natural Language Learning (CoNLL) covers a large framework of topics about NLP, mostly through machine learning approaches. Information extraction from scanned documents has been by fare much less studied. The contributions in this respect have concentrated on analysing the performance degradation caused by OCR errors [9]. While some studies have been carried out on synthetic data by introducing character errors generated randomly, most recent studies have been motivated by digitisation projects of historical documents so as to enhance search performance and offering high level semantic indexing and search. In general, OCR quality is erratic on historical documents and there is a tendency for unusual non-alphabetic characters to appear. In addition, the OCR system has problems with layout and is frequently unable to distinguish marginal notes from the main body of the text, giving rise to discontinuous sequences of words

[3] https://cs.nyu.edu/faculty/grishman/muc6.html

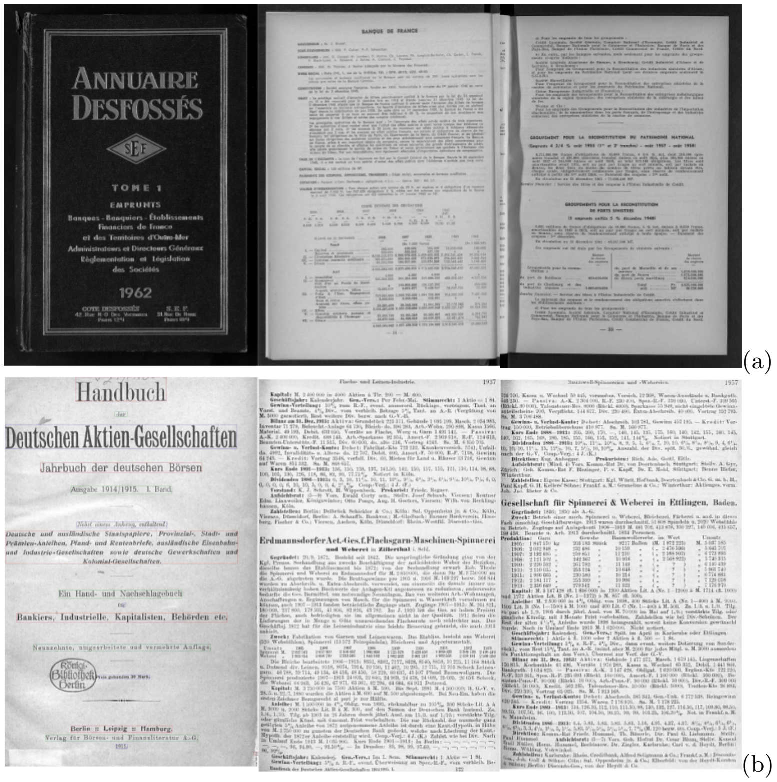

ANNUAIRE DESFOSSÉS
TOME 1
EMPRUNTS
1962
(a)
Handbuch
Deutschen Aktien-Gesellschaften
Jahrbuch der deutschen Börsen
(b)

Fig. 1. Samples of (a) French yearbook "Annuaire Desfossés, 1962", (b) German yearbook "handbuch, 1914/1915".

which can confound a NER system. In [2] the authors investigate person and place names recognition in digitised records of proceedings of the British parliament. The experiments show better recognition for person names (F=0.75) than for location names (F=0.66). This may be explained by the fact that person name recognition is more dependent on finding patterns in the text while location name recognition is more dependent on gazetteer resources. In [10] the authors focus on full name extraction on a corpus composed of 12 titles spanning a diverse range of printed historical documents with relevance to genealogy and family history research.The results show a certain correlation between WER and F-measure of the NER systems. From the analysis of errors the authors conclude that word order errors play a bigger role in extraction errors than do character recognition errors, which seems reasonable as extraction systems intensively exploit contextual word neighbours information. [12] evaluate the efficacy of some available tools for accurately extracting semantic entities that could be used for automatically tagging and classifying documents by using uncorrected OCR outputs from OCRopus and tesseract. The test data came from the Wiener Library, London, and King's College London's Serving Soldier archive. Performance of

NER were on overall lower than those reported in the literature due to the difference of the data sets generally used to train and test NER systems. The authors suggest that automatically extracted entities should be validated using controlled vocabularies or other kinds of domain knowledge in order to circumvent the variability in spelling these entities in general. In [4] the authors analyse the performance of the stanford NER [7] on a snapshot of the Trove[4] newspaper collection at the National Library of Australia (NLA). The results show that the pre-trained Stanford NER performs slightly better (F1-score = 0.71) that the trained Stanford NER (F1-score = 0.67) considering location, person and organisation names. All of these studies have considered Named Entity recognition as a post-processing stage to OCR and they show that NER performance is affected by the OCR errors, mostly word errors. Toledo et al. [14] proposed a standalone architecture, that integrates image analysis through a CNN and a BLSTM neural network for extracting family names and other entities from Spanish medieval handwritten birth records. This architecture is not affected by the accumulation of errors of the traditional methods. This approach shows excellent performance on named entity extraction but requires a preliminary word segmentation stage to operate, which remains a limitation.

We have seen in this literature review that most NER tasks applied on document images have considered OCR prior to NER. We now need to give a brief but specific overview of the NER literature. Named entity recognition approaches can be categorised into three main groups of approaches; 1) rule based approaches[15] 2) statistical approaches[1, 6] 3) mixed rule-based / statistical approaches [14]. The rule-based approaches were used in early NER systems where entity features were specified in advance by domain experts. These approaches are not flexible towards inter-entity and intra-entity ambiguities (confusions). They are known to provide a good precision but generally with a low recall. Besides they require a lot of human efforts to be developed as each rule is designed manually [15].

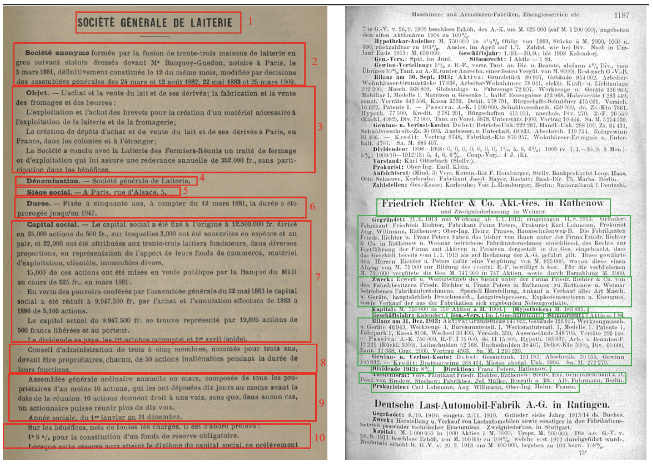
Several statistical models have been applied for NER. We can classify them in two main groups; 1) traditional statistical models 2) neural networks models. Among the various statistical models proposed in the literature such as Support Vector Machines or Hidden Markov Models, CRF[8] have emerged as the state of the art approaches. For any of these models, one needs to define handcrafted features that describe the syntactic and semantic nature of the entities to be extracted in the text, before training the model on a sufficient amount of annotated data. More recently, Recurrent Neural Networks (RNN) coupled with CRF have emerged as the new state of art architecture for NER [6]. Indeed, Recurrent Neural Networks allow training word embeddings with unannotated data, which then serve as efficient features for the NER task. Word embeddings however have some limitations as they are average representations of word's context in the training dataset, and they do not capture the specific context in which a word occur in a specific sentence. Besides, out of vocabulary words cannot be associated to relevant embedding as they do not pertain to the training dataset. To circumvent these limitations Akbik et al.[1] introduced the concept of string

[4] http://trove.nla.gov.au/

embedding which capture the left and right context of characters in texts, making the system free of any dictionary. Moreover, string embedding allow the design of dynamic word embeddings, where a word receives an embedding depending on its specific context within the sentence under study. A word may gets different embeddings when occurring in different contexts. In 2018, this approach achieved new state of art results on the CoNLL2003 NER data set.

3 System architecture

Financial Yearbooks contain structured information about companies which are organised in sections. sections are paragraphs or groups of paragraphs that report on some specific financial information related to a company. A section sometimes brings the actual composition of the governing board through a list of persons, while sometimes it brings some values of some financial indicators such as the capital of the company with the value of shares and the number of shares that compose the capital. Moreover, some sections also report on the evolution of these indicators over years. These historical records on these financial indicators are of particular interest for experts as they would bring access to long run financial time series to be analysed. Figure 2 gives one example page of the French Desfossé 1962 yearbook (left) and of the German Handbuch 1914-15 (right).



SOCIÉTÉ GÉNÉRALE DE LAITERIE

Société anonyme formée par la fusion de trente-trois maisons de laiterie en gros suivant statuts dressés devant Me Bacquoy-Guedon, notaire à Paris, le 3 mars 1881, définitivement constituée le 12 du même mois, modifiée par décisions des assemblées générales des 24 mars et 12 août 1882, 22 mai 1883 et 25 mars 1909.

Objet. — L'achat et la vente du lait et de ses dérivés; la fabrication et la vente des fromages et des beurres:

L'exploitation et l'achat des brevets pour la création d'un matériel nécessaire à l'exploitation de la laiterie et de la fromagerie;

La création de dépôts d'achat et de vente du lait et de ses dérivés à Paris, en France, dans les colonies et à l'étranger;

La Société a conclu avec la Laiterie des Fermiers-Réunis un traité de fermage et d'exploitation qui lui assure une redevance annuelle de 352.000 fr., sans participation dans les bénéfices.

Dénomination. — Société générale de Laiterie.

Siège social. — A Paris, rue d'Alsace, 5.

Durée. — Fixée à cinquante ans, à compter du 12 mars 1881, la durée a été prorogée jusqu'en 1942.

Capital social. — Le capital social a été fixé à l'origine à 12.500.000 fr. divisé en 25,000 actions de 500 fr., sur lesquelles 3,000 ont été souscrites en espèces et au pair, et 22,000 ont été attribuées aux trente-trois laitiers fondateurs, dans diverses proportions, en représentation de l'apport de leurs fonds de commerce, matériel d'exploitation, clientèle, immeubles divers.

15,000 de ces actions ont été mises en vente publique par la Banque du Midi au cours de 525 fr., en mars 1881.

En vertu des pouvoirs conférés par l'assemblée générale du 22 mai 1883 le capital social a été réduit à 9.947.500 fr. par l'achat et l'annulation effectués de 1883 à 1886 de 5,105 actions.

Le capital actuel de 9.947.500 fr. se trouve représenté par 19,895 actions de 500 francs libérées et au porteur.

Le dividende se paye les 1er octobre (acompte) et 1er avril (solde).

Conseil d'administration de trois à cinq membres, nommés pour trois ans, devant être propriétaires, chacun, de 50 actions inaliénables pendant la durée de leurs fonctions.

Assemblée générale ordinaire annuelle en mars, composée de tous les propriétaires d'au moins 10 actions, qui les ont déposées dix jours au moins avant la date de la réunion. 10 actions donnent droit à une voix, sans que, dans aucun cas, un actionnaire puisse réunir plus de dix voix.

Année sociale, du 1er janvier au 31 décembre.

Sur les bénéfices, nets de toutes les charges, il est d'abord prélevé:

1° 5 °/o pour la constitution d'un fonds de réserve obligatoire.

Lorsque cette réserve aura atteint le dixième du capital social, ce prélèvement

Maschinen- und Armaturen-Fabriken, Eisengiessereien etc. 1187

Die G.-V. v. 26./9. 1905 beschloss Erhöh. des A.-K. um M. 625 000 (auf M. 1 200 000), angeboten den alten Aktionären 1906 zu 100%.

Hypothekar-Anleihe: M. 750 000 in 4½% Oblig. von 1899, Stücke à M. 2000, 1000 u. 500, rückzahlbar zu 103%. Auslos. im April auf 1./7. Zahlst. wie bei Div. Noch in Umlauf Ende 1913: M. 659 000. **Geschäftsjahr:** 1./10.—30./9.; bis 1908 Kalenderj.

Gen.-Vers.: Spät. im Juni. **Stimmrecht:** 1 Aktie = 1 St.

Gewinn-Verteilung: 5% z. R.-F., vertr. Tant. an Dir. u. Beamte, alsdann 4% Div., vom Übrigen 10% Tant. an A.-R. (unter Anrechn. einer festen Vergüt. von M. 9000), Rest nach G.-V.-B.

Bilanz am 30. Sept. 1913: Aktiva: Grundstück 80 907, Gebäude 954 002, Arbeiter-Wohnhäuser-Grundstücke 17 038, Arbeiter-Wohnhäuser 78 031, elektr. Kraft- u. Lichtanlage 202 200, Masch. 368 809, Gleisanlage u. Fahrzeuge 72 807, Werkzeuge u. Geräte 116 969, Mobiliar 1, Modelle 1, Matrizen u. Gesenke 1, halbf. Erzeugnisse 423 968, Holzvorräte 1 283 449, sonst. Vorräte 642 556, Kassa 3339, Debit. 578 791, Bürgschafts-Schuldner 415 031, Versich. 16 632, Patente 1. — Passiva: A.-K. 1 200 000, Schuldverschreib. 659 000, do. Zs.-Kto 7661, Hypoth. 77 500, Kredit. 2 781 201, Bürgschaften 415 031, unerhob. Div. 270, R.-F. 26 550 (Rückl. 4082), Div. 72 000, Tant. an Vorst. 3878, Unterstütz. 1000, Vortrag 10 444. Sa. M. 5 254 539.

Gewinn- u. Verlust-Konto: Debet: Betriebs-Unk. 222 287, Handl.-Unk. 288 100, Zs. 84 431, Schuldverschreib.-Zs. 30 093, Ausbesser. u. Unterhalt. 49 833, Abschreib. 119 254, Reingewinn 91 406. — Kredit: Vortrag 9748, Fabrikat.-Kto 870 957, Wohnhäuser-Erträgnis u. Unterhalt. 4701. Sa. M. 885 407.

Dividenden: 1898—1908: 0, 0, 0, 0, 0, 0, 0, 1½, 5, 5, 6%; 1909 (v. 1./1.—30./9. 9 Mon.): 5%; 1909/10—1912/13: 5, 4, 6, 6%. Coup.-Verj.: 4 J. (K).

Vorstand: Karl Otterbach (Stellv.).

Prokurist: Ober-Ing. Emil Klein.

Aufsichtsrat: (Mind. 3) Vors. Komm.-Rat F. Homburger, Stellv. Bankprokurist Leop. Haas, Otto Schaerer, Karlsruhe; Fabrikant Jacob Mayer, Rastatt; Bank-Dir. Th. Marba, Berlin.

Zahlstellen: Ges.-Kasse; Karlsruhe: Veit L. Homburger; Berlin: Nationalbank f. Deutschl.

Friedrich Richter & Co. Akt.-Ges. in Rathenow

und Zweigniederlassung in Weimar.

Gegründet: 21./6. 1913 mit Wirkung ab 1./1. 1913; eingetragen 14./8. 1913. Gründer: Fabrikant Friedrich Richter, Fabrikant Franz Peters, Prokurist Karl Lohmann, Prokurist Aug. Willmann, Rathenow; Ober-Ing. Heinr. Prause, Baumschulenweg-B. Die Fabrikanten Friedr. Richter u. Franz Peters haben das bisher von ihnen unter der Firma Friedr. Richter & Co. in Rathenow u. Weimar betriebene Fabrikunternehmen einschliessl. des Rechts zur Fortführung der Firma mit Aktiven u. Passiven dergestalt in die Ges. eingebracht, dass das Geschäft bereits vom 1./1. 1913 als auf Rechnung der A.-G. geführt gilt. Diese gewährte den Herren Richter u. Peters dafür eine Vergütung von M. 825 000, wovon diese einen Abzug von M. 75 000 zur Bildung des gesetzl. R.-F. bewilligt haben. Für die verbliebenen M. 750 000 vergütete die Ges. M. 747 000 in 747 Aktien, sowie durch Barzahlung M. 3000.

Zweck: Erwerb u. Weiterbetrieb des bisher unter der Firma Friedr. Richter & Co. von den Fabrikbesitzern Friedr. Richter u. Franz Peters in Rathenow zu Rathenow u. Weimar betriebenen Fabrikunternehmens. Speziell Herstellung, Ankauf u. Verkauf aller Art Masch. u. Geräte, hauptsächlich Dreschmasch., Langstrohpressen, Explosionsmotoren u. Eisenguss, sowie Verkauf der aus der Fabrikation sich ergebenden Nebenprodukte.

Kapital: M. 750 000 in 750 Aktien à M. 1000. **Hypotheken:** M. 183 835.

Geschäftsjahr: Kalenderj. **Gen.-Vers.:** Im I. Geschäftshalbj. **Stimmrecht:** 1 Aktie = 1 St.

Bilanz am 31. Dez. 1913: Aktiva: Grundstücke 147 652, Gebäude 326 927, Werkzeugmasch. u. Geräte 46 943, Werkzeuge 1, Bureauuntensil. 1, Werkstattutensil. 1, Modelle 1, Patente 1, Fuhrpark 1, Kassa 8106, Wechsel 36 400, Versich. 350, Aussenstände 348 705, Vorräte 295 146. — Passiva: A.-K. 750 000, R.-F. I 75 000, do. II 15 000, Hypoth. 183 835, Arb.- u. Beamten-F. 12 235 (Rückl. 3000), Leihschulden 52 768, Buchschulden 38 467, Delkr.-Kto 5000, Div. 60 000, Tant. 11 368, Grat. 2000, Vortrag 4563. Sa. M. 1 210 239.

Gewinn- u. Verlust-Konto: Debet: Gesamtunk. 151 182, Abschreib. 20 155, Gewinn 100 932. — Kredit: Bruttogewinn 268 464, Mieten abzügl. Unk. 3806. Sa. M. 272 270.

Dividende 1913: 8%. **Direktion:** Franz Peters, Rathenow.

Aufsichtsrat: Vors. Fabrikant Friedr. Richter, Rathenow; Stellv. Exz. Generalleutnant z. D. Paul von Bredow, Stechow; Fabrikbes. Jul. Müller, Benrath a. Rh.; Alb. Fuhrmann, Berlin.

Prokuristen: Carl Lohmann, Aug. Willmann, Ober-Ing. Heinr. Prause.

Deutsche Last-Automobil-Fabrik A.-G. in Ratingen.

Gegründet: 8./10. 1910; eingetr. 2./11. 1910. Gründer siehe Jahrg. 1913/14 ds. Buches.

Zweck: Herstellung u. Verkauf von Lastautomobilen sowie sonstiger in den Fabrikationsbetrieb passender technischer Erzeugnisse. Zweigniederlass. in Stuttgart.

Kapital: M. 1 000 000 in 1000 Aktien à M. 1000. Urspr. M. 200 000. Die a.o. G.-V. v. 26./8. 1911 beschloss Erhöh. um M. 200 000 zu 108%, welche erst 1912 durchgeführt wurde. Nochmals erhöht lt. G.-V. v. 20./3. 1913 um M. 600 000, begeben zu 106 bezw. 108%.

75*

Fig. 2. Section structure of the two yearbooks, the French Desfossé 1962 (left) and the German handbuch 1914-15 (right).

We can see that for any of these sources, sections (highlighted in color rectangles) start with a bold title that identifies the type of the section, then the text of the section is running over one single or multiple paragraphs. In these examples we can notice that some sections of the German yearbook may run one after the other within a single paragraph, whereas sections of the French yearbook always start with a new paragraph. In the same way, a description of a company always start with a new page of the French Yearbooks, whereas multiple companies can occur within a single page of the German yearbook.It is to be noticed that every sections do not contain information of similar interest for scientific financial research, and that each yearbook is reporting financial information in different ways. However there are some stable information which any yearbook is reporting. Among the most important ones we can mention the following section types : *Founded, Purpose, Governing Board, Capital, Balance Sheets, Fiscal Year*. Through these examples we can also notice that sections contain different types of information. Some of them are lists of persons, others contain addresses, dates, amounts etc... There is a need to develop a general named entity extraction system that can be adapted specifically to each section type.

From the presentation of these examples, we can now introduce the general architecture of the extraction system that was designed for multilingual financial information extraction in yearbooks which is depicted on figure 3. Considering the generic system that we aim to develop, we choose to organise the processing chain in a sequential manner by first running the OCR and then the extraction system. Both sources have been processed with Omnipage [5]. We observed a very good OCR quality on the French yearbook while performance on the German yearbook appear to be slightly lower. In both cases the quality was considered sufficiently good to carry out the extraction process. No OCR correction was applied prior to the extraction process.

The OCR produces a set of text blocks which are ordered from top to bottom. Then a standard but language specific pre-processing stage allows to detect each textual component as a token. Depending on the layout of the document, section detection is performed using a rule-based approach by exploiting on a set of predefined keywords. For example, the keyword CAPITAL (French) or Kapital (German) identifies the sections that contain essential information about the company financial status.

Then information extraction is performed on each identified section and using a specific extraction algorithm. Following the literature review we choose to develop a machine learning approach for the following reasons. It offers a generic framework that can be optimised efficiently for every section to deal with. The approach is language agnostic provided human experts can provide training data. Tagging conventions of texts (words and numerical sequences) can be defined with experts (Historians) so that both the historians and the IT specialists arrive to a better understanding of the real final needs. Notice that there is no unique approach here for adopting tagging conventions, and we may follow an iterative process through trials and errors before arriving to the final tag set for a specific section. More details will be given about tagging conventions in section 4.1.

Once the tagging conventions have been defined the system follows the steps depicted on figure 3. Human tagging of a training data set is required before training the system. Depending on the difficulty of the extraction task of the section considered, the size of the data set has to be adjusted conveniently but

[5] https://www.kofax.com/Products/omnipage

we have no reliable estimate of it. This process may be iterated to get acceptable results on the test data set. Another strategy to circumvent from the lack of annotated data is to iterate training and annotation of new examples through active learning [13]. In this manner we may ask the user to annotate only the most relevant examples for which the system is less confident. We implemented an active learning scheme that is presented in section 4.4.

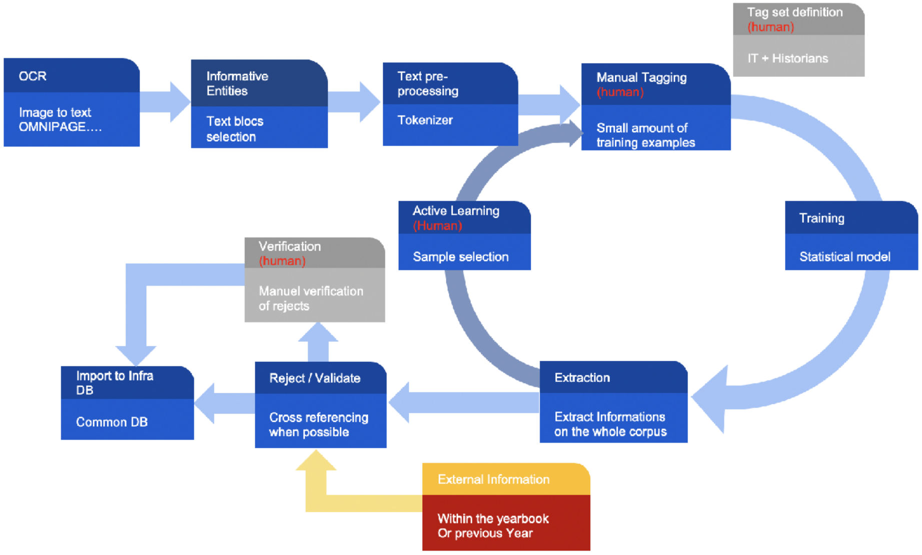


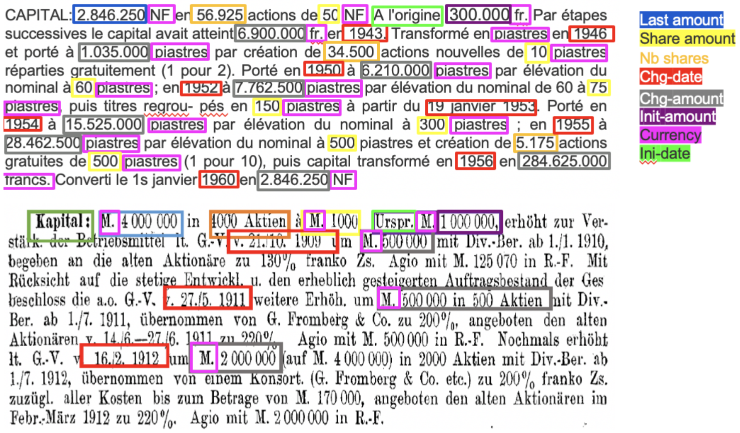
Fig. 3. The flow of processes of the system.

4 Linked Named Entity extraction

As highlighted in section 3, financial yearbooks contain many information organised in sections. Each section is reporting about specific financial information and requires a dedicated extraction module. Some sections are simply lists of items such as lists of persons for which simple rules encoded with regular expressions may suffice to get the information extracted. Some other sections are much more difficult to analyse as shown on figure 4 which gives two examples of the section ”Capital” for both French and German yearbooks.

4.1 Tagging conventions

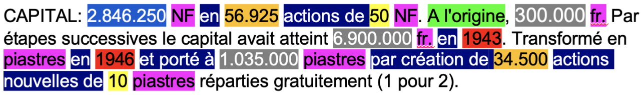
The two sections on figure 4 highlight the typical content of financial yearbooks whatever the stock exchange and the language used. The section ”Capital” reports the current capital amount of the company, its currency, the number and amount of shares that build the total capital amount. Moreover, the section ”capital” is also reporting the history of the capital by showing the dates when the capital changed from the creation date of the company. Each capital change should occur with the date of change, the new capital value, the currency, the new number of shares with their amount. As these named entities are reported



CAPITAL: 2.846.250 NF en 56.925 actions de 50 NF. A l'origine 300.000 fr. Par étapes successives le capital avait atteint 6.900.000 fr. en 1943. Transformé en piastres en 1946 et porté à 1.035.000 piastres par création de 34.500 actions nouvelles de 10 piastres réparties gratuitement (1 pour 2). Porté en 1950 à 6.210.000 piastres par élévation du nominal à 60 piastres ; en 1952 à 7.762.500 piastres par élévation du nominal de 60 à 75 piastres, puis titres regrou- pés en 150 piastres à partir du 19 janvier 1953. Porté en 1954 à 15.525.000 piastres par élévation du nominal à 300 piastres ; en 1955 à 28.462.500 piastres par élévation du nominal à 500 piastres et création de 5.175 actions gratuites de 500 piastres (1 pour 10), puis capital transformé en 1956 en 284.625.000 francs. Converti le 1s janvier 1960 en 2.846.250 NF

Kapital: M. 4 000 000 in 4000 Aktien à M. 1000. Urspr. M. 1 000 000, erhöht zur Verstärk. der Betriebsmittel lt. G.-V. v. 21./10. 1909 um M. 500 000 mit Div.-Ber. ab 1./1. 1910, begeben an die alten Aktionäre zu 130% franko Zs. Agio mit M. 125 070 in R.-F. Mit Rücksicht auf die stetige Entwickl. u. den erheblich gesteigerten Auftragsbestand der Ges beschloss die a.o. G.-V. v. 27./5. 1911 weitere Erhöh. um M. 500 000 in 500 Aktien mit Div.-Ber. ab 1./7. 1911, übernommen von G. Fromberg & Co. zu 200%, angeboten den alten Aktionären v. 14./6.—27./6. 1911 zu 220%. Agio mit M. 500 000 in R.-F. Nochmals erhöht lt. G.-V. v. 16./2. 1912 um M. 2 000 000 (auf M. 4 000 000) in 2000 Aktien mit Div.-Ber. ab 1./7. 1912, übernommen von einem Konsort. (G. Fromberg & Co. etc.) zu 200% franko Zs. zuzügl. aller Kosten bis zum Betrage von M. 170 000, angeboten den alten Aktionären im Febr.-März 1912 zu 220%. Agio mit M. 2 000 000 in R.-F.

Fig. 4. Information to be extracted from the section Capital for both the French and German yearbooks with the tagging conventions.



CAPITAL: 2.846.250 NF en 56.925 actions de 50 NF. A l'origine, 300.000 fr. Par étapes successives le capital avait atteint 6.900.000 fr. en 1943. Transformé en piastres en 1946 et porté à 1.035.000 piastres par création de 34.500 actions nouvelles de 10 piastres réparties gratuitement (1 pour 2).

Fig. 5. Tagging the linked named entities, with tag "Link" in blue.

through a textual description and not placed into a table, a certain variability was introduced in phrasing the text at the time of publication. In addition, some information are sometimes partially missing. Figure 4 shows a case where the same tag set can be used for French and German. Notice that irrelevant words in the text are labelled with the tag "Other", as is the standard convention adopted for information extraction. One other important aspect is related to how these various information should be linked together to provide timely coherent n-tuples of information in a tabular form as follows : date - capital amount - currency - number of shares - amount of share. Such a 5-tuple is made of linked named entities and we wish the extraction process to extract not only each individual entities but also their linking attributes with the other entities they relate to. In this purpose we have introduced a specific "Link" tag that serves for tagging every non informative words within a single n-tuple, so that a n-tuple is any sequence of tags between two "Other" tags, see figure 5.

4.2 CRF based model

A Conditional Random Field models (CRF) [5] allows to compute the conditional probability of a sequence of labels $Y = \{y_1, y_2, \dots, y_T\}$ given a sequence of input features $X = \{x_1, x_2, \dots, x_T\}$ with the following equation :

$$P(Y|X) = \frac{1}{Z_0} exp(\sum_{t=1}^{T} \sum_{k} w_k \times \phi_k(y_{t-1}, y_t, x, t)) \quad (1)$$

where $\phi(y_{t-1}, y_t, x, t)$ is a feature function that maps the entire input sequence of features X paired with the entire output sequence of labels Y to some d-dimensional feature vector. Weight parameters w_k are optimised during training. The normalisation factor Z_0 is introduced to make sure that the sum of all conditional probabilities is equal to 1. Once the optimal weights $\hat{w}$ are estimated, the most likely sequence of output labels $\hat{Y}$ for a given sequence of input features X is estimated as follows:

$$\hat{Y} = arg \max_{Y} P(Y|X) \tag{2}$$

CRF have been introduced for Natural Language Processing by considering binary features. $\phi(y_{t-1}, y_t, x, t)$ is a binary feature function that is set to 1 when labels and input tokens match a certain property. In this study, we use a 5 tokens width sliding window and a set of 33 X 5 template features which produce a total of thousands of binary features depending on the section considered.

4.3 String embedding (BLSTM-CRF) model

In the literature, Recurrent Neural Network (RNN) architectures have been introduced so as to learn tokens embeddings. These embeddings are then used in place of the handcrafted features in a CRF model. Most of the state of the art NER systems use pre-trained word embeddings with a standard BLSTM-CRF setup $[1, 11, 6, 3]$. In addition to pre-trained word embeddings Lample et al. .[6] have introduced character-level word embeddings so as to circumvent from possible out of Vocabulary words. Similarly Peter et al. [11] introduced contextual word embeddings extracted from a multi-layer bidirectional language model of tokens (biLM). Recently, Akbik et al. [1] have used the internal states of two LSTM character language models to build contextual word embeddings, namely contextual string embeddings. Compared to other state of the art systems, this model is able to provide embeddings to any word and not only the known vocabulary words of the training set. Each language model consists of a single layer of 2048 Long Short Term Memory (LSTM) cells. A language model estimates the probability $P(x_{0:T})$ of a sequence of characters $(x_0, \ldots, x_T \Leftrightarrow x_{0:T})$ with the following equation.

$$P(x_{0:T}) = \prod_{t=0}^{T} P(x_t|x_{0:t-1}) \tag{3}$$

where $P(x_t|x_{0:t-1})$ is the probability of observing a character given its past. A forward language model $(\overrightarrow{LM})$ computes the conditional probability using the LSTM hidden states as follows:

$$P(x_t|x_{0:t-1}) \approx \prod_{t=0}^{T} P(x_t|\overrightarrow{h_t}; \theta) \tag{4}$$

where $\overrightarrow{h_t}$ represents a view of the LSTM of the past sequence of characters of character x_t while θ represents the model parameters. Similarly, a backward language model $(\overleftarrow{LM})$ computes the probability in the reverse direction as follows:

$$P(x_t|x_{t+1:T}) \approx \prod_{t=0}^{T} P(x_t|\overleftarrow{h_t}; \theta) \tag{5}$$

The word embedding w_i of word i that starts at character x_b and ends at character x_e in the sentence is obtained by the concatenation of the hidden states of the forward and the backward LM as follows:

$$w_i = \left[\overleftarrow{h}_{b-1}, \overrightarrow{h}_{e+1}\right] \tag{6}$$

Notice that the two character language models can be trained on un-annotated large corpora as they are trained to predict the next/previous character. Then, following the architecture proposed in [1], we use a hybrid BILSTM/CRF model for named entity recognition. A word level BLSTM captures word context in the sentence and its internal state feeds a CRF in place of handcrafted features. The word BLSTM is fed by the string embedding representation. This BILSTM/CRF architecture is trained on for each specific Named Entity Recognition task, while it is fed by the string embedding representation that is pre-trained on a large corpus of the language choosen. In the following experiments we used pre-trained string embeddings proposed by the authors for French and German.

4.4 Active learning scheme

Due to the lack of annotated data, we have introduced an active learning scheme. First, we start by training the extraction model with a few annotated examples. The trained model is then used to predict the annotation of all the unseen examples of the test data set. These automatically annotated examples are sorted according to their labelling score. The examples with a labelling score higher that 0.9 are used as additional training examples to the first training data set for a new training iteration. The examples with labelling score less than 0.5 are considered as bad examples. Thus a small set of those examples are annotated manually for enhancing the capacity of the extraction model towards this kind of bad examples.

5 Experiments

We report the evaluation results for the extraction of the linked named entities on two sections of two yearbooks: the CAPITAL and CONSTITUTION sections of the French Desfossé 1962, and the Kapital section of the German Handbuch 1914-1915. The French Desfossé 1962 Yearbook consists of 2376 pages, among which 1544 CAPITAL sections and 1547 CONSTITUTION sections have been detected. The German Handbuch 1914-1915 yearbook consists of 5174 pages in which we detected 3971 Kapital sections. Among the detected sections, 181 CAPITAL section, 91 CONSTITUTION section and 195 Kapital sections have been manually annotated to form the training and test data sets.

CRF-model configuration: We used the crf++ toolkit[6] for training the CRF-based extraction models. We used the default training parameters, the cut-off threshold for features selection is f = 3, and the C hyper-parameter that prevents over fitting has been set to C=1.5.

[6] https://taku910.github.io/crfpp/

BLSTM-CRF model configuration: We followed the training scheme introduced in [1] with a slight modification. In our configuration, we decreased the number of LSTM units from 256×2 to 64×2; we also set the learning rate to 0.2 with a mini-batch size of 8.

5.1 Extraction tasks

CAPITAL section named entity extraction : The information to be extracted from this section are every capital amounts, currencies and change dates. The tag set was derived from the examples in figure 5.

Kapital section named entity extraction : Kapital section contains the same set of named entities to be extracted as for the CAPITAL section. In addition, two new named entities have been considered; ***Cap-decr*** and ***Cap-incr***. These two labels refer to a increase or a decrease of the capital. Table 1 shows the extraction results on the CAPITAL and Kapital sections and using the CRF and the BLSTM-CRF extraction models. We observe very good performance on the CAPITAL section with both the CRF and the BLSTM-CRF model with small differences. For every entities we obtain precision and recall higher than 95% while the average F1-score is higher than 96%. We also observe similar excellent performance on the Kapital section. However, the BLSTM-CRF model performs better than the CRF model. Both the CRF and the BLSTM models can not deal with the Capital decrease entity due to lack of example in the training set.

	CAPITAL section						Kapital section					
	CRF model			BLSTM-CRF model			CRF model			BLSTM-CRF model		
Entity tags	Precision %	Recall %	F1-score	Precision %	Recall %	F1-score	Precision %	Recall %	F1-score	Precision %	Recall %	F1-score
ini-amount	95.97	95.94	95.94	96,78	95,96	96,35	89.81	77.97	83.37	97.15	89.68	93.16
ini-date	100	100	100	98,75	98,75	98,75	100	85.24	91.84	98.75	90.06	92.69
chg-amount	97.32	96.95	97.13	97,05	97,98	97,51	93.28	74.10	82.30	90.37	85.23	87.69
chg-date	97.72	94.55	96	98,04	96,05	97,01	90.20	81.73	85.73	91.64	89.01	90.29
last-amount	96.74	91.22	93.72	87,28	93,56	89,78	96.92	96.38	96.63	99.48	98.46	98.95
currency	97.25	94.12	95.63	96,09	96,46	96,26	96.91	89.79	93.19	97.78	95.85	96.79
link	98.07	95.12	96.54	97,68	95,63	96,63	91.71	84.89	88.07	90.94	90.24	90.49
Cap-decr	—	—	—	—	—	—	0	0	0	0	0	0
Cap-incr	—	—	—	—	—	—	91.26	89.30	90.23	91.48	94.42	92.88
Overall	97.57	95.27	96.39	96,69	96,43	96,55	93.65	85.84	89.55	94.16	92.17	93.13

Table 1. Performance obtained by the CRF and BLSTM models on the CAPITAL and Kapital sections.

CONSTITUTION section	CRF model			BLSTM-CRF model		
Entity tags	Precision %	Recall %	F1-score	Precision %	Recall %	F1-score
ini-status	91.30	95.45	93.33	94.51	96.59	95.50
ini-startdate	86.43	82.61	84.48	86.33	94.52	90.18
ini-enddate	50	33.25	39.94	59.17	42.92	44.03
ini-period	100	100	100	95.83	100	97.81
chg-status	50	50	50	50	33.33	40
chg-startdate	0	0	0	37.50	25	29
chg-enddate	0	0	0	0	0	0
chg-period	100	66.67	80	100	66.67	80
link	91.11	87.23	89.13	90.54	88.83	89.61
Overall	89.38	81.45	85.23	87.72	85.26	86.42

Table 2. Performance of the CRF and the BLSTM models for extracting the named entities of the CONSTITUTION section

CONSTITUTION section named entity extraction task From this section, we want to extract information about the company legal status, the date of creation, the period of activity and expiration date if applicable. We have introduced nine tags for this section, defined as follows: 1) ***ini-status***: initial legal status of the company once created. 2) ***ini-startdate***: the company creation date. 3) ***ini-enddate***: the company expiration date. 4) ***ini-period***: the company activity period. 5) ***chg-status***: the changed legal status of the company. 6) ***chg-startdate***: the start date of the changed legal status of the company. 7) ***chg-enddate***: the end date of the changed legal status of the company. 9) ***link***: the linking tag. In table 2, we report the results on the CONSTITUTION section using the CRF and BLSTM-CRF models. Due to the small size of the training data set, the results show lower performance compared to those reported on the the CAPITAL and Kapital sections.

5.2 Entity linking

Once the entities have been extracted, we link them into tuples called chunks. We consider three different chunks on the CAPTITAL and Kapital sections; 1) the ***ini-chunk*** consists of the ***ini-date***, ***ini-amount*** and ***currency*** labelled tokens. 2) the ***chg-chunk*** includes the ***chg-date***, ***chg-amount*** and ***currency*** labelled tokens. 3) the ***last-chunk*** enclose ***last-amount*** and ***currency*** labelled tokens. Notice that the date associated with the ***last-amount*** entity is the date of the yearbook (1962), and for this reason we don't consider extracting this information.

Linking method	Minimal distance			link tag		
chuncks	Precision %	Recall %	F1-score	Precision %	Recall %	F1-score
ini-chunk	96.80	94.46	95.62	95.21	95.94	95.57
chg-chunk	89.72	87.89	88.80	91.58	91.42	91.50
last-chunk	91.73	90.03	90.87	97.18	91.23	94.11
Overall	90.93	89.08	89.99	92.84	92.01	92.43

Table 3. Chunk extraction performance with the minimal distance entity linking method and using the link tag learning method

	CAPITAL section						Kapital section					
	CRF model			BLSTM-CRF model			CRF model			BLSTM-CRF model		
Linking chunks	Precision %	Recall %	F1-score	Precision %	Recall %	F1-score	Precision %	Recall %	F1-score	Precision %	Recall %	F1-score
ini-chunk	95.21	95.94	95.57	94.61	94.78	94.66	80.20	73.84	76.78	87.56	83.43	85.43
chg-chunk	91.58	91.42	91.50	93.96	94.74	94.35	70.18	63.64	66.68	75.66	71.74	73.58
last-chunk	97.18	91.23	94.11	89.98	94.76	92.26	95.82	96.82	96.30	95.42	98.42	98.41
Overall	92.84	92.01	92.43	93.39	94.75	94.06	78.94	73.91	76.31	83.72	80.50	82.05

Table 4. System performance results obtained by the CRF and BLSTM models for extracting and linking the named entities of the CAPITAL and Kapital sections

CONSTITUTION section	CRF model			BLSTM-CRF model		
Linking chuncks	Precision %	Recall %	F1-score	Precision %	Recall %	F1-score
ini-chunk	72.44	81.23	76.51	74.88	85.57	79.84
chg-chunk	37.50	17.71	22.42	30.38	27.98	28.26
Overall	67.55	65.63	66.47	64.57	71.49	67.79

Table 5. Results obtained by the CRF and the BLSTM-CRF models for extracting the linked named entities on the CONSTITUTION section

We experimented two methods for linking the entities into chunks. The minimum distance method regroups entities with their closest neighbour entity. Using the ***link*** tag introduced in section 4, we are able to learn how to link the entities. We then consider a sequence of linked entities, as entities of the same chunk. The two methods have been evaluated on the CAPITAL section using the CRF model, see table 3. The results show better performance when using the learned tag ***link***. We analyse the performance of the Linked Named Entities on the CAPITAL and Kapital sections given on table 4. The results show very good performance for extracting and linking the named entities of the CAPITAL section $\geq 90\%$, with the best results obtained by the BLSTM-CRF model. Performance on the German Kapital section is 10 points lower those for the French CAPITAL section. This may come from the effect of introducing two more tags (Cap-incr and Cap-decr) but this may also come from the lower regularity of the German style of writing, compared to the French yearbook. We consider two different chunks on the CONSTITUTION section. The ***ini-chunk*** is a group of the ***ini-status***, ***ini-startdate***, ***ini-enddate*** and ***ini-period*** labels. The ***chg-chunk*** consists of the ***chg-status***, ***chg-startdate***, ***chg-enddate*** and ***chg-period*** labels. The low performance in extracting the named entities on the CONSTITUTION section also affect the linking task as illustrated in table 5

5.3 Active learning

To show the effectiveness of the active learning scheme, we conducted three experiments on the CAPITAL section of the French Defossee 1926 Yearbook. During these experiments, we used 200 manually annotated examples for evaluating the performance of the trained models.

Size of training data set	Precision	Recall	F1-score	Nb of gen. examples (score$\geq$ 90%)
M10	92.74	60.04	72.89	118
M20	93.81	83.28	88.23	170
M30	93.14	84.36	88.53	216
M40	94.97	90.67	92.77	248
M50	95.51	90.74	93.06	258

Table 6. Performance with manually annotated examples with respect to the size of the training data set

Size of training data set	Precision	Recall	F1-score	Nb of gen. examples (score$\geq$ 90%)
M10	92.74	60.04	72.89	118
M10+A118	91.97	58.55	71.55	188
M10+A188	93.42	56.51	70.42	332
M10+A332	94.64	54.23	68.95	406
M10+406	94.6	50.04	65.45	438

Table 7. Performance with 10 annotated examples in addition to automatically generated examples whose labelling score $\geq$ 90%

Size of training data set	Precision	Recall	F1-score	Nb. of gen. examples (score $\geq$ 90%)
M10	92.74	60.04	72.89	118
M10+A118+C10	93.66	87.24	90.34	216
M20+A216+C10	93.04	92.95	92.99	300
M10+A300+C10	93.99	94.53	94.26	382
M10+A382+C10	94.36	95.15	94.76	476

Table 8. System performance when training the extraction model on 10 manually annotated examples augmented by the generated examples whose labelling score $\geq$ 90% in addition to 10 corrected examples

The first experiment shows the effect of increasing the size of the training data set on the performance. Training process starts with ten manually annotated examples (M10), then by adding 10 more examples at each iteration we observe an improvement of the extraction performance in term of precision, recall and

F1-score form 92.74, 60.04, 72.89 to 95.51, 90.74, 93.06 as illustrated in table 6. At the same time, we applied the trained models on the whole set of unlabelled examples found in the yearbook.

In the second experiment, we studied the effect of increasing the training data set with the examples labelled by the model it-self whose labelling score is higher than 0.9. At the beginning of the training process, we used only ten manually annotated examples for training the initial CRF model. Then, we apply the initial model on the whole set of unlabelled examples to obtain their labels with their labelling scores. For the next training iteration, the training data set consists of the ten manually annotated examples (M10) plus 118 automatically labelled examples that have received a labelling score ≥ 0.9. Repeating this process five times we can get 418 automatically annotated training examples. We observe a precision increase from 92.74% up to 95.51% but the overall recall and F1-score degraded at every training iteration, as illustrated in table 7.

From the second experiment, we can say that the model learns better the same examples by specialising to almost similar examples. To tackle this problem, the training data set must contain more heterogeneous examples. We introduced this notion in the third experiment during which we not only inject labelled data with high scores but also some poorly labelled examples with a labelling score < 0.5 (C10) which are manually corrected and then introduced in a new training data set for the next training iteration. After five active learning iterations, we observe a quick increase of recall and F1-score with a slight degradation of precision (see table 8.). In comparison with the results obtained from the first experiment, we observe that with only 30 automatically selected and manually annotated examples and three training iterations, the performance (precision: 93.04; recall: 92.95; F1-score: 92.99) reach the performance obtained during the first experiment (precision: 95.51; recall: 90.74; F1-score: 93.06) for which we used 50 training examples and five training iterations.

6 Conclusion

Financial yearbooks are historical records reporting on information about the companies of stock exchanges, including their name, date of creation, financial status, governing board members, headquarters and branches addresses, financial information such as capital amount, date and amount of capital increase, balance sheet of the year including assets and liabilities, etc... In this paper we have presented the key components that implement a financial information extraction system from financial yearbooks. The proposed system consists in three steps: OCR, linked named entities extraction, active learning. The core of the system is related to linked named entities extraction (LNE) with CRF and BLSTM-CRF. The experiments have been conducted on two yearbooks with two languages (French and German). Very promising results have been obtained on three different extraction tasks on three different sections showing very good performance with state of the art named entity extraction models that are specialised to each entities with a limited amount of annotated data through the introduction of active learning.

7 Acknowledgement

This work has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N 777489.

References

1. Akbik, A., Blythe, D., Vollgraf, R.: Contextual string embeddings for sequence labeling. In: Proceedings of the 27th International Conference on Computational Linguistics. pp. 1638–1649 (2018)
2. Grover, C., Givon, S., Tobin, R., Ball, J.: Named entity recognition for digitised historical texts. In: LREC (2008)
3. Huang, Z., Xu, W., Yu, K.: Bidirectional lstm-crf models for sequence tagging. arXiv preprint arXiv:1508.01991 (2015)
4. Kim, S.M., Cassidy, S.: Finding names in trove: Named entity recognition for australian historical newspapers. In: ALTA (2015)
5. Lafferty, J., McCallum, A., Pereira, F.C.: Conditional random fields: Probabilistic models for segmenting and labeling sequence data (2001)
6. Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., Dyer, C.: Neural architectures for named entity recognition. arXiv preprint arXiv:1603.01360 (2016)
7. Manning, C., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S., McClosky, D.: The Stanford CoreNLP natural language processing toolkit. In: Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations. pp. 55–60 (Jun 2014)
8. McCallum, A., Li, W.: Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons. In: Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003
9. Miller, D., Boisen, S., Schwartz, R., Stone, R., Weischedel, R.: Named entity extraction from noisy input: Speech and ocr. In: Proceedings of the Sixth Conference on Applied Natural Language Processing. pp. 316–324. ANLC '00, Association for Computational Linguistics (2000)
10. Packer, T., Lutes, J., Stewart, A., Embley, D., Ringger, E., Seppi, K., Jensen, L.: Extracting person names from diverse and noisy ocr text. In: Proceedings of the Fourth Workshop on Analytics for Noisy Unstructured Text Data (2010)
11. Peters, M.E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., Zettlemoyer, L.: Deep contextualized word representations. arXiv preprint arXiv:1802.05365 (2018)
12. Rodriquez, K.J., Bryant, M., Blanke, T., Luszczynska, M.: Comparison of named entity recognition tools for raw ocr text. In: KONVENS (2012)
13. Shen, Y., Yun, H., Lipton, Z.C., Kronrod, Y., Anandkumar, A.: Deep active learning for named entity recognition. arXiv preprint arXiv:1707.05928 (2017)
14. Toledo, J.I., Carbonell, M., Fornés, A., Lladós, J.: Information extraction from historical handwritten document images with a context-aware neural model. Pattern Recognition **86**, 27–36 (2019)
15. Wang, S., Xu, R., Liu, B., Gui, L., Zhou, Y.: Financial named entity recognition based on conditional random fields and information entropy. In: 2014 International Conference on Machine Learning and Cybernetics. vol. 2, pp. 838–843. IEEE (2014)