Combining Visual and Textual Features for Semantic Segmentation of Historical Newspapers

Barman, Raphaël ; Ehrmann, Maud ; Clematide, Simon ; Oliveira, Sofia Ares ; Kaplan, Frédéric

Abstract: The massive amounts of digitized historical documents acquired over the last decades naturally lend themselves to automatic processing and exploration. Research work seeking to automatically process facsimiles and extract information thereby are multiplying with, as a first essential step, document layout analysis. If the identification and categorization of segments of interest in document images have seen significant progress over the last years thanks to deep learning techniques, many challenges remain with, among others, the use of finer-grained segmentation typologies and the consideration of complex, heterogeneous documents such as historical newspapers. Besides, most approaches consider visual features only, ignoring textual signal. In this context, we introduce a multimodal approach for the semantic segmentation of historical newspapers that combines visual and textual features. Based on a series of experiments on diachronic Swiss and Luxembourgish newspapers, we investigate, among others, the predictive power of visual and textual features and their capacity to generalize across time and sources. Results show consistent improvement of multimodal models in comparison to a strong visual baseline, as well as better robustness to high material variance.

DOI: https://doi.org/10.46298/jdmdh.6107

Journal of Data Mining and Digital Humanities

Combining Visual and Textual Features for Semantic Segmentation of Historical Newspapers

Raphaël Barman[1], **Maud Ehrmann**[1], **Simon Clematide**[2], **Sofia Ares Oliveira**[1], **Frédéric Kaplan**[1]

[1]École polytechnique fédérale de Lausanne, Switzerland
[2]Universität Zürich, Switzerland

Corresponding author: Raphaël Barman , raphael.barman@epfl.ch

Abstract

The massive amounts of digitized historical documents acquired over the last decades naturally lend themselves to automatic processing and exploration. Research work seeking to automatically process facsimiles and extract information thereby are multiplying with, as a first essential step, document layout analysis. Although the identification and categorization of segments of interest in document images have seen significant progress over the last years thanks to deep learning techniques, many challenges remain with, among others, the use of more fine-grained segmentation typologies and the consideration of complex, heterogeneous documents such as historical newspapers. Besides, most approaches consider visual features only, ignoring textual signal. We introduce a multimodal neural model for the semantic segmentation of historical newspapers that directly combines visual features at pixel level with text embedding maps derived from, potentially noisy, OCR output. Based on a series of experiments on diachronic Swiss and Luxembourgish newspapers, we investigate the predictive power of visual and textual features and their capacity to generalize across time and sources. Results show consistent improvement of multimodal models in comparison to a strong visual baseline, as well as better robustness to the wide variety of our material.

Keywords

historical newspapers; image segmentation; multimodal learning; deep learning; digital humanitites

INTRODUCTION

For several decades now, digitization efforts are slowly but steadily contributing to an increasing amount of facsimiles of cultural heritage documents. As a result, it is nowadays commonplace for many memory institutions to create and maintain digital repositories that offer rapid, time- and location-independent access to documents, allow to virtually bring together disperse collections, and ensure the preservation of fragile documents thanks to on-line consultation [Terras, 2011]. Beyond this great achievement in terms of preservation and accessibility, the next fundamental challenge –and real promise of digitization– is to exploit the *contents* of these digital assets, and therefore to adapt and develop appropriate document and language processing technologies to search and retrieve information from this 'Big Data of the Past' [Kaplan and di Lenardo, 2017].

Context. Efforts are, in this regard, well under way and the libraries, digital humanities (DH), natural language processing (NLP), and computer vision (CV) communities are pooling forces and expertise to push forward the processing of facsimiles, as well as the extraction and linking

of the information contained therein.[1] This momentum is particularly vivid in the domain of digitized newspaper archives for which there has been a notable increase of research initiatives over the last years. Those range from individual works dedicated to the development of tools [Yang et al., 2011, Dinarelli and Rosset, 2012, Moreux, 2016, Wevers, 2019] or the usage of those tools [Kestemont et al., 2014, Lansdall-Welfare et al., 2017], to evaluation campaigns [Rigaud et al., 2019, Clausner et al., 2019], including the emergence of large consortia projects seeking to apply computational methods to historical newspapers at scale, such as *ViralTexts*[2], *Oceanic Exchanges*[3], *impresso*[4], *NewsEye*[5], and *Living with Machines*[6] [Ridge et al., 2019].

Overall, this research contributes a pioneering set of text and image analysis tools, system architectures, and interfaces covering several aspects of historical newspaper processing. They usually focus on all or part of the typical digitized newspaper pipeline which consists, essentially, of three main steps: facsimile processing, in order to derive the structure and the text from the document image (via, respectively, optical layout recognition and optical character recognition processes); content enrichment, in order to extract and link relevant information from both textual and visual part of the contents; and, finally, exploration support, in order to search and visualize the enriched resources via e.g. application programming or graphical user interfaces.

Motivation. While encouraging, these efforts are still at an early stage and many challenges have yet to be addressed, especially with respect to document layout analysis (first processing phase). Document layout analysis aims at segmenting a document image into meaningful segments and at classifying those segments according to their contents [Eskenazi et al., 2017]. Two types of classification are traditionally distinguished: physical layout analysis, with a focus on the nature of the content (is this segment a textual block, a diagram, a picture, a decoration, a graphic, etc.), and logical layout analysis, with a focus on the function of the content (is this textual block a title, a footer, an article, etc.). Those segments are then fed into optical character recognition (OCR) programs that recognize their textual content.

With newspapers, these image segmentation and classification processes are particularly difficult because of the complexity and diversity of the object. A newspaper page consists of multiple, heterogeneous elements which feature different layout characteristics (text, map, table, illustration), different contents (regular articles, serial, advertisements) and which, additionally, evolve through time, differ according to newspapers, and are in different languages. Besides, facsimiles can be of variable quality due to the conservation state of the originals and this can also affect layout analysis performances.

Although difficult, layout analysis is however essential for historical newspaper understanding and exploitation, and their quality has a direct impact on downstream processes [Binmakhashen and Mahmoud, 2019]. From an information retrieval and user viewpoint, being able to query at the level of meaningful segments such as articles –instead of whole pages–, and to facet over different types of segments are undeniable advantages. From an NLP viewpoint, most

[1] These interdisciplinary efforts were recently streamlined within the far-reaching project 'Europe Time Machine': https://www.timemachine.eu

[2] A project aiming at mapping networks of reprinting in 19th-century newspapers and magazines (US, 2012-2016): https://viraltexts.org

[3] A project tracing global information networks in historical newspaper repositories from 1840 to 1914 (US/EU, 2017-2019): https://oceanicexchanges.org

[4] https://impresso-project.ch

[5] A digital investigator for historical newspapers (EU, 2018-2021): https://www.newseye.eu

[6] A project which aims at harnessing digitised newspaper archives (UK, 2018-.): https://www.turing.ac.uk/research/research-projects/living-machines

analysis of semantic nature such as entity linking, topic modelling or text classification requires and/or performs far better on semantically self-sufficient, autonomous content items. For some processes, it can also be useful to filter out unwanted elements, either because of too noisy in terms of OCR or less relevant in terms of contents (e.g. transport schedule, cross-words, weather reports, TV programs, etc.). Finally, from a media history viewpoint, the automatic classification of content items can enable a better understanding of the evolution of newspaper sections through time and across collections.

Finally, the task of newspaper segmentation is also to be seen within the current context of large-scale newspaper projects. Facing both the digitized newspaper material reality and user needs, these initiatives help, on the one hand, reveal the defects of legacy layout and text acquisition outputs from libraries and, on the other, emphasize the needs of finer-grained qualification of newspaper sections for scholarship purposes, as well as of efficient large-scale, trans-collection and diachronic processing of newspaper facsimiles. In this regard, the '*impresso* - Media Monitoring of Past'[7] project –in the context of which the present work was carried out– is a case in point. Led by an interdisciplinary team, *impresso* aims at semantically indexing a multilingual corpus of digitized newspapers and integrating the resulting data into historical research workflows by means of a newly developed user interface[8]. By doing so, it appeared desirable to compensate for the deficiencies of old layout analysis.

Proposition. In this context, this paper presents an innovative approach for the semantic segmentation of historical newspapers. 'Semantic' in that the targeted image segment typology goes beyond physical and/or logical characteristics and considers fine-grained semantic content item types (e.g. a segment is not only an article, but also e.g. a serial or death notice, or not only a table, but also e.g. election results or stock exchange information). 'Innovative' in that the approach makes joint use of visual and textual features, in an attempt to replicate human comprehension which uses both modalities simultaneously when confronted with document images. Already tested in very few recent studies [Yang et al., 2017, Katti et al., 2018, Dang and Nguyen Thanh, 2019, Denk and Reisswig, 2019], we believe it is the first time a multi-modal document image segmentation approach is applied on newspapers, what is more of historical nature.

Objective. Our objective is twofold. First, we wish to assess whether the combination of visual with textual features can efficiently segment newspapers images. In this regard, the recent advances of deep learning approaches for semantic image segmentation and text processing suggest that positive results can be achieved: visual-based neural architectures trained for natural images have shown good adaptation to document images, and single architectures have demonstrated their capacities to adapt to different tasks [Ares Oliveira et al., 2018]. As for text, language models based on embeddings have shown their capacity to support a variety of tasks, from named entity recognition to question answering [Collobert et al., 2011]. Second, we wish to investigate whether this multi-modal representation can better support generalization across time and newspapers. The same newspaper section can indeed change drastically in terms of layout through time and across titles while enjoying a certain stability in terms of textual contents.

Contributions. We present a series of experiments for the segmentation of several newspapers covering different time periods according to four semantic classes. These experiments are based on a modified version of *dhSegment*, a generic deep-learning approach that operates

[7] https://impresso-project.ch
[8] https://impresso-project.ch/app

pixel-wise document segmentation [Ares Oliveira et al., 2018]. Architecture's code, ground-truth data sets as well as models are publicly released.

Section I presents prior works and specifies where the present approach sits with respect to them. Section II introduces the approach, and Section III details the experimental setup. Section IV reports and discusses three series of experiments and Section 4.4.2 considers the limits, but also future application scenarios of the approach and concludes.

I RELATED WORK

The survey of Eskenazi et al. [2017] gives an overview of the approaches for the segmentation of textual document images. Approaches are usually divided into three categories: top-down, when starting from the whole page in order to partition it, bottom up, when starting from small components in order to aggregate them, and hybrid. Classical algorithms heavily rely on specific document priors, e.g. having a "Manhattan" layout, and/or require large amounts of hand-crafted features. More recent approaches make use of deep neural networks, trading prior, hand-crafted features for the learning capacities of machine learning, especially deep neural networks. Those include the usage of convolutional neural networks [Chen et al., 2017], as well as several variants of the fully convolutional network (FCN) introduced by Long et al. [2015] [He et al., 2017a, Xu et al., 2017, Wick and Puppe, 2018, Ares Oliveira et al., 2018].

Considering newspapers images, several works have been proposed for their segmentation. Hebert et al. [2014] proposed an approach that performs physical and logical segmentation, and detects reading order on historical French newspapers. It is based on conditional random fields and a set of heuristics, targets high-level types such as titles, line and articles and achieves state of the art results with ca 85% of accuracy. A similar coarse-grained classification (line, image, illustation, text blocks) is done by Gatos et al. [1999] on Greek newspapers using an hybrid approach, and by Hadjar and Ingold [2003] and Bouressace and Csirik [2018] on contemporary Arabic newspapers using Run Length Smoothing Algorithm (RLSA). Lorang et al. [2015] focuses on a more specific type, that is poetic content items, and make use of manually crafted features to classify crops of newspaper images.

On the other side of the spectrum, another line of research performs newspaper content segmentation using text only (usually when images are not available) via the detection of homogeneous passages based on sentence or paragraph textual similarity [Riedl et al., 2019]. Those approaches can detect and classify segments of textual nature exclusively, but cannot identify their image boundaries, nor take into account more visual items.

Only a few recent work attempt to make use of image and/or localized, two-dimension text information. Meier et al. [2017] use a FCN based on image and OCR output information in order to detect articles in newspaper images (no further segment types). In this case text is reduced to a binary feature information (a pixel has text or not) and the lexical and semantic dimensions are not taken into account. Katti et al. [2018] introduced the concept of *chargrid*, a two-dimension representation of text where characters are localized on the image (thanks to the box coordinates) and encoded as a one-hot vector. This information is passed through an architecture that uses two encoders, one for the image information, the other for the character one and two decoders, one that produces semantic segmentation and the other that produces bounding boxes. Different model variants (image only, text only, both) are applied on images of administrative documents (invoices), and experiments show that the models based on both signals achieve better results. This is however opposed to a high-computing cost, as emphasized by the authors. Dang and Nguyen Thanh [2019] builds on this work and present an approach

based on a multi-stage attentional U-Net using a one-hot encoded character feature. Segmentation of template like administrative documents yield state of the art results in the order of 87% mIoU (see Section 3.4). Denk and Reisswig [2019] also extends Katti et al. [2018], considering not only characters, but words and their corresponding embeddings, with *BERTgrid* for the automatic extraction of key-value information from invoice images (amount, number, date, etc.). With the same architecture as Katti et al. [2018], they obtain best results with document representation based on one-hot character embeddings and word-level BERT embeddings [Devlin et al., 2019], with no image information. Performances differ quite a lot between classes (of key-value types). Finally, Yang et al. [2017] jointly uses visual and textual features in a network, via *text embedding maps* where the two-dimension text representation is mapped to the pixel information (cf. Section II). Textual features correspond here to sentence embeddings (average of words vectors obtained with word2vec [Mikolov et al., 2013]), and models are trained on several variants of an end-to-end, multi-modal fully convolutional network for the segmentation and coarse classification of image regions (figure, table, section heading, caption, list, paragraph). Models are tested on various datasets and results show significant, although variable across classes, performance improvements with the model using both visual and textual features.

The method we present builds on the work of Yang et al. [2017] in the sense that it also makes use of text embedding maps. It however differs in that we work with historical newspapers –therefore integrating the diachronic dimension–, target a more fine-grained segment typology and experiment with different embeddings.

II METHOD

Our objective is to segment newspaper images and to classify detected segments according to a fine-grained newspaper section typology. To this end, we introduce a method which performs supervised, pixel-wise multiclass classification using both visual and textual features. The method builds on *dhSegment*'s architecture.

2.1 Primary Architecture: *dhSegment*

dhSegment is an open-source, generic image document segmentation framework[9] [Ares Oliveira et al., 2018]. It consists of a CNN-based pixel-wise predictor coupled with task dependent post-processing blocks. Its network is based on a U-Net architecture [Ronneberger et al., 2015], where the encoder follows a deep residual network ResNet-50 [He et al., 2016] pre-trained on ImageNet [Deng et al., 2009]. *dhSegment* has demonstrated competitive results on multiple tasks, e.g. page extraction, baseline extraction, and layout analysis, thereby paving the way for efficient and generic document image segmentation. Its architecture is here modified in order to incorporate textual features.

2.2 Text embedding map

Considering textual and visual information at the same time supposes to jointly encode their signals. To this end, and as briefly introduced in Section I, it is possible to map the one-dimensional representation of textual information (e.g. a word vector) into a three-dimensional one by 'positioning' the embedding representation into a two-dimensional space (e.g. a word has a certain width and height when written or printed on a page). This new textual embedding (corresponding to a word or a character) is therefore equivalent to the original vector, augmented with the positioning information (width and height). We refer to this three-dimensional representation of textual information, as introduced in Yang et al. [2017], as a 'text embedding map'.

[9] https://github.com/dhlab-epfl/dhSegment

The three-dimensional encoding of textual information is generated by using the results of an OCR process which outputs text tokens along with their coordinates on the image. Considering for example the left image of Figure 1, an OCR engine produces the token "TEMPS" located in the bounding box $[(10, 195), (10, 300), (40, 300), (40, 195)]$. Looking the token up in an embedding space returns its (textual) vector, which can then be associated with the bounding box information, thereby creating a three-dimension map.

This process can be formally defined as follows. Given an image of size $H \times W$ and a list of tokens T where each token t is associated with a bounding box $\mathbf{b}_t$ on the image, a text embedding map G of size $H \times W \times N$ is produced, where N is the dimension of the embeddings. Specifically, all pixels contained in the bounding box of a token t are defined as the set $\mathbf{b}_t \in \mathbb{R}^2$ and each pixel $g_{i,j} \in G$ of the text embedding map is computed with

$$g_{ij} = \begin{cases} E(t) & \text{if } (i,j) \in \mathbf{b}_t \\ 0^N & \text{otherwise} \end{cases}$$

where $E(t)$ is a mapping of $t \to \mathbb{R}^N$ corresponding for example to a word embedding, and 0^N is a null vector in case there is no text in the corresponding pixel. Each pixel overlapping with a bounding box of a token is therefore mapped to its corresponding embedding. A pixel spanning two bounding boxes is attached to the one that has the closest center.

The final result is a text embedding map, *i.e.* a three-dimension matrix where the first two dimensions correspond to the image-localized representation of the text, and the third to the embedding. Given that it has the same shape as the results of 2D convolutional layers, this construction offers the advantage that it can be processed directly by classical image processing neural network.

Figure 1: Visualization of a Flair-based, PCA-reduced text embedding map projected on three dimensions (red, green, blue).

One way of 'visualizing' this embedding map is to project each word vector, using principal component analysis (PCA), to a new one of dimension three where each dimension corresponds to a color (red, green, blue). This produces a colored text embedding map where the third dimension (the textual one) is transformed into a color value. The idea here is to 'see' the textual information, based on the fact that if two words have the same color, they also share the projection of their embeddings, and therefore textual features. Figure 1 shows such a colouring of textual information with segments of a weather forecast item (left) and a stock exchange table (right). Notwithstanding their similar layouts (*i.e.* a table with same number of columns, with a title on top and some text below) and the drastic dimension reduction (2048 from the original vector to 3), it is possible to observe information about the text, with differences that could not be easily caught by visual features only. For example, numbers are grey, punctuation

is green, stop-words have a yellowish tint, and the weather forecast segment contains a column with letters only.

2.3 Model

Our model architecture is a modified version of *dhSegment*,[10] where the only modification is the addition of the text embedding map. It takes as input an image of a newspaper and its corresponding text embedding map, and outputs a pixel probability map. Figure 2 displays the architecture, with the **T** marker indicating where the text embedding maps are concatenated (on the channel axis) to the visual feature maps. The size relative to the original image size I is indicated at each step of the network, and the depth of the feature maps is indicated below the blocks for each step, considering that an embedding feature map of size 300 is input at **T**.

Variants of this model were tested during a pilot phase, in particular different input levels of the text embedding map. Two options were experimented in this regard: at the beginning of the network, in which case the text embedding map passes through most of the network and the textual signal is treated like the visual one, and at the end, in which case it adds further contextual information to the image feature map and support the final decision of the pixel class. Our preliminary experiments showed that inputting the textual features early in the network is the best option, so do all architectures used in SectionIV.

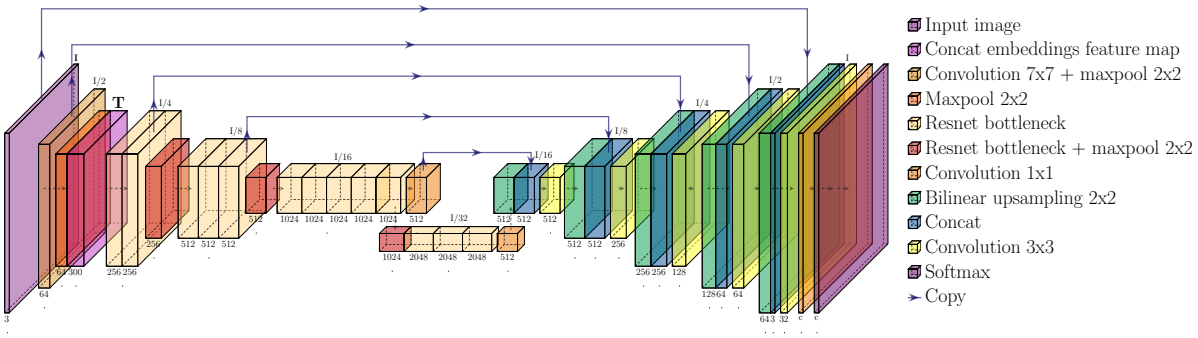


Figure 2: The model architecture used.

III EXPERIMENTAL SETUP

We apply this semantic segmentation method on historical newspapers of the *impresso* project collection, considering four semantic classes. This section introduces the corpora and the typology used for classifying image segments (Section 3.1), presents the embeddings used for the experiments (Section 3.2), specify the training setup (Section 3.3) and details the evaluation framework (Section 3.4).

3.1 Datasets

3.1.1 Corpora

Since the only freely available historical newspaper image dataset annotated with content item types considers broad categories alone (e.g. article, caption, header, etc.) [Clausner et al., 2015], we created two new datasets.

Swiss newspapers. The first one originates from the Swiss National Library and the still-existing journal *Le Temps*.[11] It is composed of three titles in French language from the Romandy region with long publication history, namely: the *Journal de Genève* (JDG, 1826-1994),

[10] More details on *dhSegment* architecture can be found in the original paper.

[11] Both are partners of the *impresso* project: https://www.nb.admin.ch/snl/en/home.html and https://www.letemps.ch/

Class/Newspaper	JDG	GDL	IMP	LUXWORT
Serial	137	108	103	-
Weather	156	68	41	-
Death notice	153	69	102	1765
Stocks	275	135	79	-
Pages w/o annotations	1393	697	1326	17188
Total	1982	1008	1634	18953

Table 1: Dataset statistics. Note that page numbers do not add up to the total number of annotated pages because a single page can contain more than one class.

the *Gazette de Lausanne* (GDL, 1804-1991), and the *Impartial* (IMP, 1881-2017). JDG and GDL, issued in neighboring cities, can be considered as siblings and were merged in 1991. In order to have a long term, diachronic ground truth, newspaper issues were sampled across the whole publication spans with three issues every three years for JDG (used for training and evaluation) and every five years for GDL and IMP (used for evaluation only). Because of misalignment problems between facsimiles and token coordinates of the original OCR, all images of the selected issues were re-OCRed with Abbyy FineReader application[12].

This material was manually annotated according to the four semantic classes of our typology (see Section 3.1.2), using the VGG Image annotator [Dutta and Zisserman, 2019]. Annotation was done at the pixel level and not at the content item instance level, meaning that each pixel of the image has a label indicating that it belongs to a specific class (or not), but not that it belongs to a specific instance of a class. Several reasons motivate this annotation at pixel level: in most cases, there is only one instance per page, and in case of multiple instances, it might be non-contiguous regions of the same instance. Besides, instance separation and merging can also be done in a post-processing step.

This annotation process yielded a total of 1,982 annotated pages for JDG, 1,008 for GDL and 1,634 for IMP. Table 1 shows the class distribution for the three titles. Pages without annotations do not contain any classified content items.

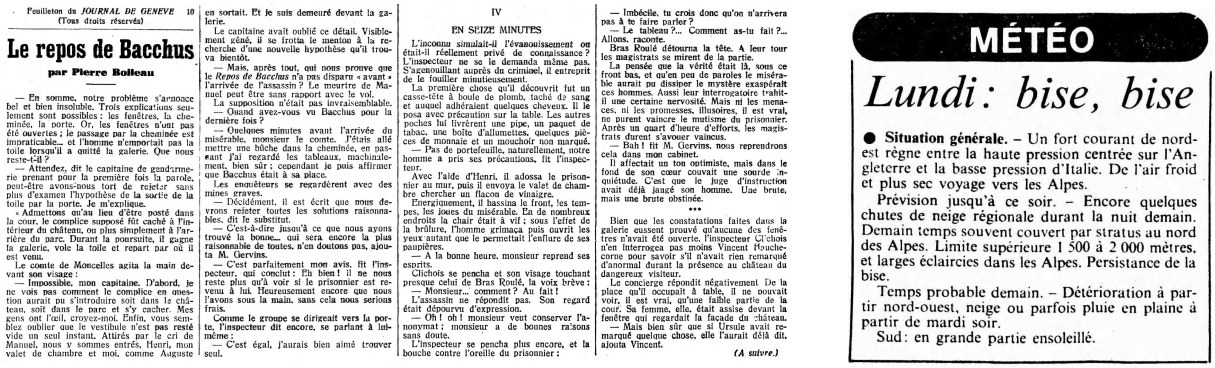
Luxembourgish newspaper. The second dataset consists of a single title, the *Luxemburger Wort* (LUXWORT), a Luxembourgish newspaper from the Bibliothèque Nationale du Luxembourg[13] published since 1848 with contents in German, French and Luxembourgish. The library, who outsourced OCR and layout recognition for its newspaper collection, performed a manual check of the recognized segments. Having this at hand, we chose the work with the death notices which, in the LUXWORT, amount to ca. 90,000 segments in 17,000 page images. For our corpus, we sampled 34 issues per year between 1848 and 1950. This resulted in 18,953 images with 1,765 death notices.

3.1.2 Classes

As mentioned earlier, newspapers feature a wide variety of contents which change over time and across newspapers. Given our objectives, we selected four classes of content items likely to, on the one hand, be of historical or practical interest (to use them as search facet or to filter them out before processing) and, on the other, present a mix of visual and textual variation:

[12] Version 11, https://www.abbyy.com

[13] A partner of the *impresso* project: https://bnl.public.lu/fr.html



Feuilleton du *JOURNAL DE GENEVE* 10
(Tous droits réservés)

Le repos de Bacchus

par Pierre Boileau

— En somme, notre problème s'annonce bel et bien insoluble. Trois explications seulement sont possibles : les fenêtres, la cheminée, la porte. Or, les fenêtres n'ont pas été ouvertes ; le passage par la cheminée est impraticable... et l'homme n'emportait pas la toile lorsqu'il a quitté la galerie. Que nous reste-t-il ?

— Attendez, dit le capitaine de gendarmerie prenant pour la première fois la parole, peut-être avons-nous tort de rejeter sans plus d'examen l'hypothèse de la sortie de la toile par la porte. Je m'explique.

« Admettons qu'au lieu d'être posté dans la cour, le complice supposé fût caché à l'intérieur du château, ou plus simplement à l'arrière du parc. Durant la poursuite, il gagne la galerie, vole la toile et repart par où il est venu.

Le comte de Moncelles agita la main devant son visage :

— Impossible, mon capitaine. D'abord, je ne vois pas comment le complice en question aurait pu s'introduire soit dans le château, soit dans le parc et s'y cacher. Mes gens ont l'œil, croyez-moi. Enfin, vous semblez oublier que le vestibule n'est pas resté vide un seul instant. Attirés par le cri de Manuel, nous y sommes entrés, Henri, mon valet de chambre et moi, comme Auguste en sortait. Et je suis demeuré devant la galerie.

Le capitaine avait oublié ce détail. Visiblement gêné, il se frotta le menton à la recherche d'une nouvelle hypothèse qu'il trouva bientôt.

— Mais, après tout, qui nous prouve que le *Repos de Bacchus* n'a pas disparu « avant » l'arrivée de l'assassin ? Le meurtre de Manuel peut être sans rapport avec le vol.

La supposition n'était pas invraisemblable.

— Quand avez-vous vu Bacchus pour la dernière fois ?

— Quelques minutes avant l'arrivée du misérable, monsieur le comte. J'étais allé mettre une bûche dans la cheminée, en passant j'ai regardé les tableaux, machinalement, bien sûr ; cependant je puis affirmer que Bacchus était à sa place.

Les enquêteurs se regardèrent avec des mines graves.

— Décidément, il est écrit que nous devrons rejeter toutes les solutions raisonnables, dit le substitut.

— C'est-à-dire jusqu'à ce que nous ayons trouvé la bonne... qui sera encore la plus raisonnable de toutes, n'en doutons pas, ajouta M. Gervins.

— C'est parfaitement mon avis, fit l'inspecteur, qui conclut : Eh bien ! il ne nous reste plus qu'à voir si le prisonnier est revenu à lui. Heureusement encore que nous l'avons sous la main, sans cela nous serions frais.

Comme le groupe se dirigeait vers la porte, l'inspecteur dit encore, se parlant à lui-même :

— C'est égal, j'aurais bien aimé trouver seul.

IV

EN SEIZE MINUTES

L'inconnu simulait-il l'évanouissement ou était-il réellement privé de connaissance ? L'inspecteur ne se le demanda même pas. S'agenouillant auprès du criminel, il entreprit de le fouiller minutieusement.

La première chose qu'il découvrit fut un casse-tête à boule de plomb, taché de sang et auquel adhéraient quelques cheveux. Il le posa avec précaution sur la table. Les autres poches lui livrèrent une pipe, un paquet de tabac, une boîte d'allumettes, quelques pièces de monnaie et un mouchoir non marqué.

— Pas de portefeuille, naturellement, notre homme a pris ses précautions, fit l'inspecteur.

Avec l'aide d'Henri, il adossa le prisonnier au mur, puis il envoya le valet de chambre chercher un flacon de vinaigre.

Énergiquement, il bassina le front, les tempes, les joues du misérable. En de nombreux endroits la chair était à vif ; sous l'effet de la brûlure, l'homme grimaça puis ouvrit les yeux autant que le permettait l'enflure de ses paupières.

— A la bonne heure, monsieur reprend ses esprits.

Clichois se pencha et son visage touchant presque celui de Bras Roulé, la voix brève :

— Monsieur... comment ? Au fait !

L'assassin ne répondit pas. Son regard était dépourvu d'expression.

— Oh ! oh ! monsieur veut conserver l'anonymat ; monsieur a de bonnes raisons sans doute.

L'inspecteur se pencha plus encore, et la bouche contre l'oreille du prisonnier :

— Imbécile, tu crois donc qu'on n'arrivera pas à te faire parler ?

— Le tableau ?... Comment as-tu fait ?... Allons, raconte.

Bras Roulé détourna la tête. A leur tour les magistrats se mirent de la partie.

La pensée que la vérité était là, sous ce front bas, et qu'en peu de paroles le misérable aurait pu dissiper le mystère exaspérait ces hommes. Aussi leur interrogatoire trahit-il une certaine nervosité. Mais ni les menaces, ni les promesses, illusoires, il est vrai, ne purent vaincre le mutisme du prisonnier. Après un quart d'heure d'efforts, les magistrats durent s'avouer vaincus.

— Bah ! fit M. Gervins, nous reprendrons cela dans mon cabinet.

Il affectait un ton optimiste, mais dans le fond de son cœur couvait une sourde inquiétude. C'est que le juge d'instruction avait déjà jaugé son homme. Une brute, mais une brute obstinée.

***

Bien que les constatations faites dans la galerie eussent prouvé qu'aucune des fenêtres n'avait été ouverte, l'inspecteur Clichois n'en interrogea pas moins Vincent Pouchecorne pour savoir s'il n'avait rien remarqué d'anormal durant la présence au château du dangereux visiteur.

Le concierge répondit négativement De la place qu'il occupait à table, il ne pouvait voir, il est vrai, qu'une faible partie de la cour. Sa femme, elle, était assise devant la fenêtre qui regardait la façade du château.

— Mais bien sûr que si Ursule avait remarqué quelque chose, elle l'aurait dit dit, ajouta Vincent.

(A suivre.)

(a) *Serial*

MÉTÉO

Lundi : bise, bise

● **Situation générale.** – Un fort courant de nord-est règne entre la haute pression centrée sur l'Angleterre et la basse pression d'Italie. De l'air froid et plus sec voyage vers les Alpes.

Prévision jusqu'à ce soir. – Encore quelques chutes de neige régionale durant la nuit demain. Demain temps souvent couvert par stratus au nord des Alpes. Limite supérieure 1 500 à 2 000 mètres, et larges éclaircies dans les Alpes. Persistance de la bise.

Temps probable demain. – Détérioration à partir nord-ouest, neige ou parfois pluie en plaine à partir de mardi soir.

Sud : en grande partie ensoleillé.

(b) *Weather forecast*

MM. les membres de la **Société suisse des Fabricants de Ressorts**, sont avisés du décès de leur cher et regretté collègue,

Monsieur Charles Bühler

membre du Comité central.

19184 **Le Comité central.**

(c) *Death notice*

Titres les plus traités

151 000	American Smelting	25,75 (— 0,75)
127 200	International Paper	33,62 (— 0,50)
120 300	McDonalds Corp	29,62 (— 0,88)
100 400	Republic Corp	8,12 (— 0,62)
68 200	Budd Co	6,12 (— 0,75)
67 000	Lanvin Charles of Ritz Inc	26,12 (inch.)
63 400	Telex Corp	10,38 (— 0,38)
54 100	General Cable	18,— (inc.)
52 000	Armstrong Cork	25,38 (— 0,38)
50 000	National Cash Register	30,38 (inch.)
49 200	Computer Sciences	6,62 (+ 0,12)
49 000	Southern Company	21,— (— 0,38)
46 100	Penney J.C.	39,25 (+ 0,25)
45 900	Leasco Data Preffered	17,75 (+ 0,25)
45 900	North American Car Corp	22,12 (inch).

(d) *Stock exchange table*

Figure 3: Example images for each of the selected classes. All images are from the *Journal de Genève*.

- *Serial*, i.e. an excerpt of a bigger work published over time in several issues of a newspaper, corresponding to the French *roman-feuilleton*. Serials often span several columns in a horizontal layout and can span several pages.
- *Weather Forecast*, i.e. a text or illustration with the prediction of weather, or even a report of past weather measurements.
- *Death Notice*, i.e. a small notice published by relatives of a deceased person.
- *Stock Exchange Table*, i.e. a table reporting the values of different national stocks.

The degree of confusability of document image segments depends on several dimensions. First, the level of refinement of the typology naturally impacts what is confusable with what: distinguishing generic articles from advertisements is less difficult than distinguishing job adverts from purely commercial ones. The typology we consider here is already finer-grained compared to usual newspaper segmentation with e.g. a specific type of table among the tables (*Stock Exchange*) and a specific type of article among the articles (*Weather Forecast*). Next, within a given typology, visual and textual facets are the two main dimensions determining the degree of confusability of segments. Naturally, the more distinct on both dimensions the better. Finally, these dimensions are complemented by the time and source factors, since considering segments from different newspapers, and/or in synchrony or in diachrony also greatly impacts their confusability, not only with other types but also with themselves.

Let's examine our classes, shown in Figure 3, in this light. Regarding the visual confusability of *Serials* with respect to other items and themselves, in both synchrony and diachrony, they can be considered as rather distinct and stable: during most of their publication through time and across different newspapers, they are located at the bottom of the front page, topped with a thick black line. This makes them visually distinct compared to other items and should ensure good recognition performances using visual features only. As per their textual contents, these can vary and are, to some extent, confusable with regular journalistic contents. In contrast,

Weather Forecast segments feature a great visual variability over time (only text, then map and text, then only maps), but a clear textual stability. Here, the consideration of textual features should help improve recall, *i.e.* removing false negatives. *Death notices* are visually very similar to advertisements (small textual segments surrounded by a thick, black frame) but have different textual contents. For this class, one can therefore expect a high confusability with advertisements, and taking into account textual features should help to remove false positives. A similar situation holds for *Stock Exchange Tables*: despite their distinctive layout compared to other content items, they are still visually confusable with other tables (e.g. transport, voting). They, however, enjoy a certain visual and textual stability and their recognition should not drastically suffer across time and sources.

Overall, these four classes contain various combinations of visual and textual confusability, which makes them suitable for exploring the benefit of adding textual features for semantic segmentation.

3.2 Embeddings

In order to investigate the effectiveness of different types of embeddings used to build the text embedding maps, we experimented with embeddings having different embedding levels (word or character), contextualized word representations or not (contextual or non-contextual), different languages (mono- or multilingual), and different training data (in- and out-domain). To this end, we use fastText word embeddings, which make use of characters n-grams to learn subword embeddings [Bojanowski et al., 2017]; Byte-Pair encoded subword embeddings (BPEmb), which learn subwords rather than using fixed n-grams [Sennrich et al., 2015]; and character-based Flair embeddings [Akbik et al., 2018], a character-level variant of the contextual string embeddings introduced in [Peters et al., 2018]. In total, six flavors of these embeddings are considered, with three different stacks. Table 2 summarizes the main characteristics of the used embeddings.

First, four pre-trained embeddings of the Flair library[14] are used with their default implementation settings, as follows:

- *fastText-fr*, *i.e* the French fastText embeddings of size 300 pre-trained on Common Crawl and Wikipedia;
- *flair-fr*, *i.e* the French Flair embeddings of size 4096 pre-trained on Wikipedia;
- *flair-multi*, *i.e* the multilingual Flair embeddings of size 4096 pre-trained on the JW300 corpus [Agić and Vulić, 2019] with more than 300 languages;
- *BPEmb-multi*, *i.e* the multilingual Byte-pair encoding embeddings of size 300 trained on the 275 most common Wikipedia languages [Heinzerling and Strube, 2018].

Next, in order to test the effect of in-domain embeddings, two models were trained on a corpus of 2GB of text of the *Luxemburger Wort* for the period 1848-1950. The first is a FastText model trained on lowercase space-separated input, with at least 3 occurrences per token, a context windows of 8 tokens, a sub-word max character n-gram length of 6, resulting in embeddings of size 300 (*fastText-luxwort*). The second model is a Flair one trained on the raw OCR output for 96 hours on a NVIDIA Tesla V100, with a batch size of 600, a context length of 250 characters and a hidden size of 2048, resulting in embeddings of size 4096 (*flair-luxwort*).

Finally, different embedding stacks are considered, combining non-contextual (fastText or Byte-pair) with contextual embeddings (Flair). For experiments related to JDG, GDL and IMP, a stack

[14] https://github.com/flairNLP/flair

Name	Dim.	Level	Contextual	Lang	Training data
fastText-fr	300	word		fr	CC & Wikipedia
flair-fr	4096	char	✓	fr	Wikipedia
flair-multi	4096	char	✓	300 lang	JW300 corpus
BPEmb-multi	300	sub-word		275 lang	Wikipedia
fastText-luxwort	300	sub-word		3 lang	LUXWORT
flair-luxwort	4096	char	✓	3 lang	LUXWORT
fastText-flair-fr	4396	word+char	-	fr	CC+Wiki (stack)
BPEmb-flair-multi	4396	sub-word+char	-	multi	JW300+Wiki (stack)
fastText-flair-luxwort	4396	sub-word+char	-	3 lang	LUX+Wiki (stack)

Table 2: Overview of embeddings. The '-' sign means both contextual and non-contextual embeddings (stacks).

of *fastText-fr* and *flair-fr* is used: *fastText-flair-fr*. For experiments related to LUXWORT, two different configurations are used: a) a combination of pre-trained embeddings *BPEmb-multi* and *Flair-multi*, referred to as *BPEmb-flair-multi*, and b) a combination of in-domain embeddings with *fastText-luxwort* and *flair-luxwort*, referred to as *fastText-flair-luxwort*.

All embeddings and stacks were tested during a pilot phase, and the stacks appeared to be the best in our context. They combine contextual and non-contextual information, as well as word and sub-word information, and seem therefore more suitable to cope with old language and OCR output. Experiments presented in SectionIV are based on the stack embeddings exclusively.

3.3 Training and Post-processing

Text embedding maps are pre-computed for all images using the embedding stacks described above and the text associated with the images (original OCR for the Luxembourgish newspaper, ABBYY one for the Swiss). Before training, images are resized to fit in $5 \cdot 10^5$ pixels and are augmented by random scaling ($s \in [0.8, 1.2]$) and rotation ($r \in [-0.01, 0.01]rad$). All models are trained for 17,000 steps with a batch size of 4 and batch renormalization [Ioffe, 2017]. We use Adam optimizer [Kingma and Ba, 2014] with an exponentially decaying learning rate of .95 starting at 10^{-4}, and a weight regularization of 10^{-6}. In order to prevent overfitting a development set containing 10% of the training set is used. The final result for each model is reported on the weights where the loss on the development set was the lowest. Models are trained on a NVIDIA Tesla V100 GPU with 32GB of memory using the Tensorflow library[15] 1.13.1.

In terms of post-processing, the class mask is computed from the final output of the network, a pixel probability map. A pixel is considered as belonging to the background if it has probabilities smaller than 50% for all classes, otherwise to the class with the highest probability. In order to avoid small masks, connected components with an area smaller than 5% of the size of the image are discarded.

3.4 Evaluation setup

Given an image and a class, we wish to create a mask that contains pixels of the class. Figure 4 illustrates this procedure, with Figure 4a being the input and Figure 4b the ground truth, the

[15] https://www.tensorflow.org

latter with the mask coloring each pixel according to its class (here yellow for death notices pixels). On this base, several metrics are used to evaluate the models.

3.4.1 Metrics

(a) Original image (b) Ground truth (c) IoU = 0.23 (d) IoU = 0.86

Figure 4: Examples of the behaviour of the IoU metric.

Mean Intersection over Union. The Intersection over Union (IoU) is the standard metric for semantic image segmentation and measures how well two sets of pixels are aligned. It is computed as follows. Given an image i belonging to the set of images I, a class c belonging to the set of classes C, a set of predicted pixels P_{ic} of image i belonging to class c, and a set of ground-truth pixels G_{ic} of image i belonging to class c, the IoU for image i and class c is:

$$IoU_{ic} = \frac{|P_{ic} \cap G_{ic}|}{|P_{ic} \cup G_{ic}|}$$

Figure 4 shows a quantified example, where the prediction on images 4c and 4d are compared to the ground truth of image 4b. Image 4c has a too small and misaligned prediction, and therefore a low IoU, while 4d is better. It is important to note that the metric is computed at the image (or pixel) level, and not the content item instance level. Indeed, although there are four distinct death notice instances in Figure 4, the annotation makes no distinction and the model does not need to separate them to obtain a good score.

The mean Intersection over Union (mIoU) for a class c over the set of Image I corresponds to the average of the IoUs of all images where the union of the predicted and the ground-truth set

has at least a pixel of class c (the true negatives are thus not counted). This can be more formally defined as all images $J = \{i \in I \mid |P_{ic} \cup G_{ic}| > 0\}$. Then the mIoU is:

$$mIoU_c = \frac{1}{|J|} \sum_{j \in J} IoU_{jc}$$

Precision and Recall. The IoU does not qualify performances in terms of true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN). However, those values are of interest when considering whether a model can be used in concrete terms, *i.e.* if most of the segments are correctly recognized. The usual way to measure those values in segmentation is to consider an example as positive when above a certain threshold $\tau \in [0, 1]$ of IoU. In this case, the prediction is well enough aligned with the ground truth to be considered as correct. On this base, it is possible to consider a prediction with an IoU $\geq \tau$ as a TP, a prediction with no IoU (i.e. with a union of zero) as a TN, a prediction with an IoU of zero and no predicted pixels (i.e. intersection of zero and non-zero number of pixel in the ground truth) as a FN and, finally, a non-FN prediction with an IoU $< \tau$ as a FP. Given a threshold τ, it is therefore possible to compute precision and recall, as follows:

$$\text{Precision at } \tau = P@\tau = \frac{TP}{TP + FP}$$

$$\text{Recall at } \tau = R@\tau = \frac{TP}{TP + FN}$$

Finally, it is also possible to compute the average precision and recall over a range of thresholds. A range of threshold, is defined by a start τ_{start} an end τ_{end} and step size between two threshold τ_{step} using the following notation: $\tau_{start}{:}\tau_{step}{:}\tau_{end}$, for example a threshold between 50 and 95 with a step of 5 would be written as $50{:}5{:}95$. Given a threshold range the average metric M (which can be precision, recall or anything else) is then computed as follows:

$$M@\tau_{start}{:}\tau_{step}{:}\tau_{end} = \frac{1}{|\tau_{start}{:}\tau_{step}{:}\tau_{end}|} \sum_{\tau \in \tau_{start}{:}\tau_{step}{:}\tau_{end}} M@\tau$$

Let us emphasize once again that these metrics (IoU, mIoU, P, R) are computed at the page level and not at the instance level. If a page contains several instances of a class and the prediction matches some instances, but not enough to reach an IoU threshold larger than τ, the whole page is counted as negative.

3.4.2 Reported results

For each experiment, results are reported in terms of mIoU (as a percentage), and precision and recall with, respectively, a threshold of 60% and 80% and the average of threshold $50{:}5{:}95$ of the IoU.

Since models can have a high variance between runs, each model is trained ten times and the average and standard deviation of their performance are reported in the form of tables and boxplots. Even though most of our analyses are based on the mean, we indicate whether the difference of means between two models is significant by using Welch's t-test, following the recommendation of Reimers and Gurevych [2018]. The significance is indicated using stars (*), where their numbers corresponds to a certain p-value: one star (*) indicates that $p \leq 0.05$, two (**) that $p \leq 0.01$, three (***) that $p \leq 0.001$, and four (****) that $p \leq 0.0001$.

3.5 Material release

Annotated material is released in the VIA format as open data (under different right statements according to the source on Zenodo) under DOI 10.5281/zenodo.3706863.

The model architecture is thought as a plugin of *dhSegment*, named *dhSegment-text*. Two implementations are available. The first one, used for the present experiments and based on the TensorFlow implementation of *dhSegment*, is available on GitHub[16] and can be used for reproducibility purposes. The second one, based on the new pyTorch implementation of *dhSegment*,[17] is also available on GitHub[18] and can be used for training new models.

Finally, a selection of (best) trained models are released on the *dhSegment-text* repository, under a CC BY-SA 4.0 license.

IV EXPERIMENTS

In this section, we motivate and present four series of experiments that address important questions with regard to the automatic recognition of fine-grained semantic segments in historic newspapers. Since our material was published over a long period of time, we are specifically interested in the diachronic robustness of our models.

In Section 4.1, we examine the predictive power of visual and textual features on our four classes, representative of different difficulties. In Section 4.2, we test the generalization ability of our multimodal approach with respect to (a) the changes over time in newspaper layout and content, and (b) the transfer of models from one newspaper to a related one, which has not been part of the training material. In Section 4.3, we examine whether textual features allow to reduce the amount of training data given that they add another source of signal to the models. In Section 4.4, we focus on multilingual death notices from a single newspaper and examine (a) how the increase of training material improves the results, and (b) how valuable in-domain text embedding are, meaning character and word embeddings that were specifically trained on the multilingual and noisy OCR source material from the very newspaper.

4.1 Combining Visual and Textual Information

The first series of experiments addresses the following questions: (a) How well does fine-grained semantic segmentation perform on our four selected classes under the condition that training and test data are sampled representatively from the same newspaper? (b) How strong is the signal contained in the textual embedding maps? (c) What is the expected benefit of combining visual and textual features?

4.1.1 Experiment description

Here, models are trained on long-term diachronic JDG data only in order to reserve GDL and IMP datasets for generalization experiments (Section 4.2). The JDG dataset was randomly split to compose a training (1,387 images) and test (595 images) sets. Table 3 presents the class distribution, where it can be observed that class ratios are similar between the training and test sets. Given the homogeneity and representativity of the material, the results of this first series of experiments serve as an upper bound for our approach for diachronic, fine-grained image semantic segmentation.

In order to measure the effectiveness of using visual features only, textual features only, or a combination of both, we experimented with three modalities:

[16] https://github.com/dhlab-epfl/dhSegment-text
[17] https://github.com/dhlab-epfl/dhSegment-torch
[18] https://github.com/dhlab-epfl/dhSegment-text-torch

Class	Train size (ratio)	Test size (ratio)
Serial	101 (7.28%)	36 (6.05%)
Weather forecast	103 (7.43%)	53 (8.91%)
Death notice	107 (7.71%)	46 (7.73%)
Stock exchange table	189 (13.63%)	86 (14.45%)
Pages w/o annotations	982 (70.80%)	411 (69.08%)

Table 3: Distribution of the classes for the training and test sets.

1. Image: the model receives as input a newspaper image (pixels) only and relies solely on visual features. It is equivalent to the model described in [Ares Oliveira et al., 2018].
2. Text: the model receives as input a blank image and a text embedding map of a newspaper page and relies solely on textual features.
3. Image+Text: the model receives as input a newspaper image and its corresponding text embedding map and combines visual and textual features.

Each model with textual features uses the architecture described in Section 2.3, where text and image information are fused early in the network, as well as the *fastText-flair-fr* stack embeddings.

Metric	Modality	Serial	Weather	Death Notice	Stocks	Average
mIoU	Image	74.12±7.59	81.27±2.18	75.37±2.98	83.11±0.88	79.30±2.29
	Text	49.05±9.41	73.55±3.55	71.44±3.26	78.87±2.61	69.30±2.25
	Image+Text	76.73±5.90	81.38±3.34	****83.58±2.02	84.43±1.84	**82.16±1.72
P@60	Image	82.02±7.24	91.08±4.93	83.37±4.46	89.21±1.42	86.86±2.60
	Text	53.97±12.42	82.29±5.13	82.19±5.42	86.85±2.96	77.27±3.03
	Image+Text	83.24±7.17	91.81±4.67	***91.27±2.80	90.11±1.77	*89.43±1.79
P@80	Image	66.45±14.87	66.94±7.41	67.37±2.04	80.51±2.19	72.29±3.74
	Text	29.95±23.47	58.13±3.97	62.01±5.34	74.07±3.90	57.93±5.73
	Image+Text	71.54±15.05	71.37±7.28	***80.89±3.88	**83.49±2.11	**78.07±3.76
P@50:5:95	Image	65.37±10.22	69.10±2.84	66.77±2.43	78.36±1.31	71.53±2.82
	Text	36.97±14.21	60.02±3.46	59.78±3.70	70.99±3.32	58.46±2.96
	Image + Text	68.12±10.47	70.98±3.81	****76.18±2.10	79.38±2.55	*74.80±2.49
R@60	Image	97.91±1.95	78.74±1.98	93.02±2.88	93.78±0.73	90.64±1.03
	Text	**100.00±0.00	****90.57±3.64	88.07±2.91	91.71±1.23	91.81±1.82
	Image+Text	**100.00±0.00	****87.14±2.75	90.37±1.44	93.73±1.08	***92.39±0.95
R@80	Image	97.52±2.26	73.06±2.57	91.49±3.67	93.15±0.81	88.94±1.40
	Text	**100.00±0.00	****87.18±4.82	84.72±4.15	90.42±1.30	89.26±2.78
	Image+Text	**100.00±0.00	****84.00±3.32	89.27±1.48	93.27±1.14	***91.37±1.00
R@50:5:95	Image	95.38±3.72	67.87±2.38	88.30±4.73	92.53±0.77	87.13±1.40
	Text	85.00±10.8	****79.09±3.72	75.16±4.00	87.42±1.73	84.89±3.15
	Image + Text	96.00±5.16	****77.68±3.64	85.05±2.63	92.42±1.27	**89.35±1.24

Table 4: Results of the first series of experiments reported as mean values ± standard deviation computed from 10 runs. Stars indicate statistically significant improvements from Text and Image+Text relative to Image.

Results are shown in Table 4 and Figure 5. In general, the Image model outperforms the pure Text model by a large margin in terms of mIoU and precision. Except for recall-oriented setups, there is no advantage of restricting the models to textual features only. As expected, Image

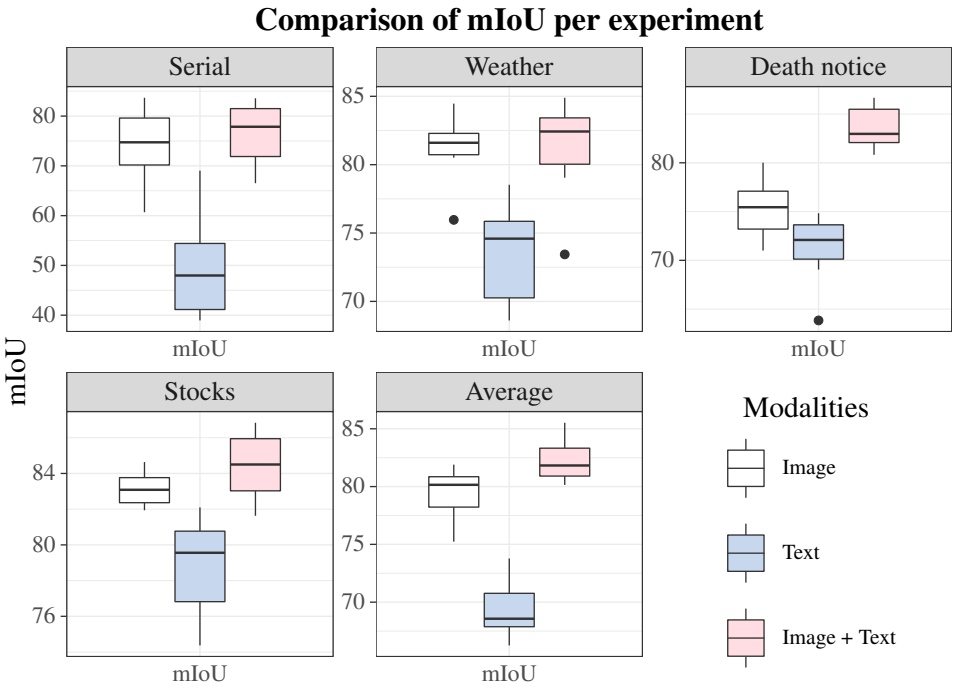


Figure 5: Box plots of the mIoU of the first series of experiments.

model is stronger on classes that are more visually distinct (*Weather* and *Stocks*) than on classes that are mainly text based (*Serial* and *Death notice*).

With respect to the class average[19] results, Image+Text models perform significantly better than Image for every metric, attesting a real gain in using the combination of visual and textual features for the task. The better precision of Image and the good recall of Text play well together, leading also to less variance across models, as the smaller standard deviations indicate. For all modalities, there is a big drop in precision when augmenting the IoU threshold. This indicates that it is hard to be precise about the location of a segment and that the Image+Text model is more robust than the single modality models.

The recall of *Weather* is better for both models using textual features. This means that these features are essential for the retrieval of the class *Weather*. As illustrated in Figure 6, weather reports may contain images, maps, and text. While the first two types are visually distinct, vocabulary might be the only semantically distinct feature for purely textual weather reports.

The mIoU and precision of *Death notice* is significantly higher for the Image+Text model than any single modality model. In particular, the gains versus the Image model are important for the for the mIoU (+5.8%-10.% at 95% confidence), for the P@60 (+4.4%-11.4% at 95% confidence), for the P@80 (+10.5%-16.5% at 95% confidence) and for the P@50:5:95 (+7.3%-11.6% at 95% confidence). This strong increase in precision shows that the Image+Text model is much more robust against false positives than the Image one. As illustrated in Figure 7, advertisements can have similar layout, but very different textual content.

The absence of significant differences in terms of mIoU and precision between the Image and Image+Text approaches for *Serial* can be explained by the lack of strong characteristics in either of the modalities. Visually, serials look similar to the rest of the newspaper. Textually, their

[19] In all experiments, average corresponds to micro-average.

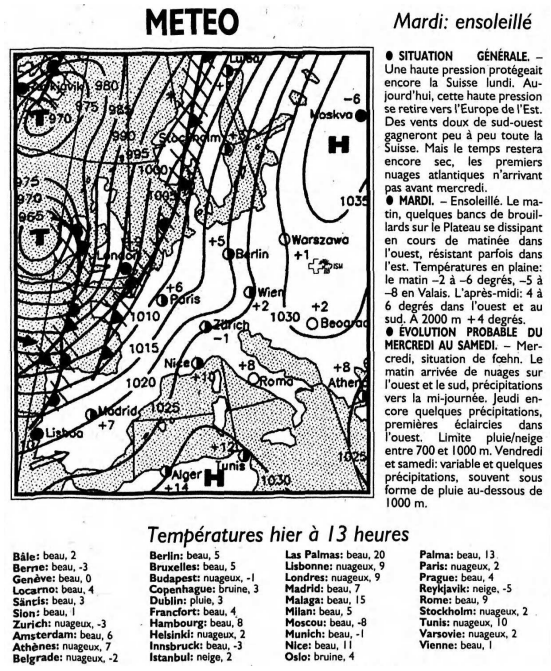


Figure 6: A Weather forecast with both visual and textual features. The Image + Text model finds both the text and the image, whereas the Image model only finds the map and the table.

Figure 7: A death notice (top) and an advertisement (bottom) with similar layouts, but very different textual features. The Text model correctly detects only the top example, whereas the Image model is misled by the advertisement.

vocabulary does not differ much from the rest of the newspaper either. However, the fact that it reaches a higher recall score for both models using textual features indicates that they are important for retrieval. The lower precision of the Text model shows that these features are also present in other articles.

Finally, the similar results between all approaches for *Stocks* show that the visual and textual signals are both strong enough to detect this class. The reason for the slightly lower score of the Text model could be crucially missing visual information about purely visual elements such as the lines of a table.

4.1.2 Summary

The first series of experiments assesses a consistent gain in performance by combining visual and textual features. The gain is particularly strong with content items as *Death Notices* that exhibit easily confusable visual features, but have distinct textual features. We also observe a better recall for *Weather Reports*, that consists of a mix of visual and textual elements. Even though the Image+Text model does not improve much on purely textual classes such as *Serial* or visually distinct classes such as *Stocks*, it still performed at least as well as a model using only the image.

4.2 Generalizing Through Time and Across Newspapers

The second series of experiments addresses the following questions: (a) Do models trained on textual and visual features perform better than purely visual models when applied to an unseen time period of the same newspaper? (b) Do models trained on textual and visual features perform better than purely visual models when applied to a related newspaper, where no issue

was part of the training material?

4.2.1 Experiment Description

The first experiment on generalization *through time* uses the JDG dataset, with material from the periods 1826-1968 and 1992-1998 as training data (1,394 pages), and 1969-1991 as test data (588 pages) Note that the test period has a different layout than the other periods [Buntinx et al., 2017]. The second experiment on generalization *across newspapers* trains on the same training set as the first series of experiments (Section 4.1), but tests on the data from GDL (1,008 pages) and IMP (1,634 pages). Both experiments compare the generalization ability of the Image and Image+Text models by testing them on layouts never seen before. This setting is therefore more challenging than the previous one where the training set was sampled uniformly over time and representative of the test set.

Class	JDG train	JDG Time train	JDG Time test	GDL test	IMP test
Serial	101 (7.28%)	134 (9.61%)	3 (0.51%)	108 (10.71%)	103 (6.30%)
Weather	103 (7.43%)	132 (9.47%)	24 (4.08%)	68 (6.75%)	41 (2.51%)
Death notice	107 (7.71%)	124 (8.90%)	29 (4.93%)	69 (6.85%)	102 (6.24%)
Stocks	189 (13.63%)	211 (15.14%)	64 (10.88%)	135 (13.39%)	79 (4.83%)
Pages w/o annotations	982 (70.80%)	923 (66.21%)	470 (79.93%)	697 (69.15%)	1326 (81.15%)

Table 5: Distribution of the classes for the different datasets

Distributions of classes for each dataset are shown in Table 5. In general, the distribution between the original JDG dataset and the other one changes. The closest dataset is GDL, which is not surprising since both newspapers come from neighbouring cities. The largest difference is between the two time periods, with a much lower ratio of content items of the four classes for the test period. The same embeddings as in Section 4.1 are used, that is to say *Fasttext-Flair-fr* (c.f. Section 3.2 and Table 2).

4.2.2 Results and Discussion

Exp.	Modalities	Serial	Weather	Death Notice	Stocks	Average
Time	Image	8.00±2.66	29.44± 6.28	51.29±12.88	****68.30**±3.31	54.65±5.47
Time	Image+Text	******25.08**±7.37	******60.65**±10.27	******77.52**± 4.41	60.17±7.42	****62.84**±6.37
GDL	Image	67.79±6.62	58.60± 3.00	63.06± 3.26	72.38±2.35	67.59±3.12
GDL	Image+Text	***73.81**±4.08	59.16± 2.22	******75.32**± 1.69	72.65±1.77	****71.54**±1.21
IMP	Image	42.45±7.86	7.04± 4.92	40.14± 3.81	42.45±2.08	40.71±2.82
IMP	Image+Text	*****56.70**±4.23	******17.53**± 4.11	******67.36**± 4.43	*****49.46**±3.84	******54.97**±3.25

Table 6: Results for the mIoU metric. Mean metric ± standard deviation of the metric (in %). Stars show statistical difference of mean between modalities.

Results are shown in Table 6 and Figure 8. Compared to the results of the first series of experiments (cf. Section 4.1), the performance is substantially lower. In general, poorer results mean that the examples in the training and test sets are too different. However, it should be noted that all the models using visual and textual features are significantly better than the Image models.

When focusing on the time constraint, it is clear that the Image+Text models perform significantly better for every class, except for *Stocks*. The Image results show that death notices and stocks are visually more stable than serials and weather reports. However, the significant differences between the two models for the classes *Weather* (+23%- 39% at 95% confidence) and *Death Notice* (+16%-35% at 95% confidence) show that textual features are even more stable

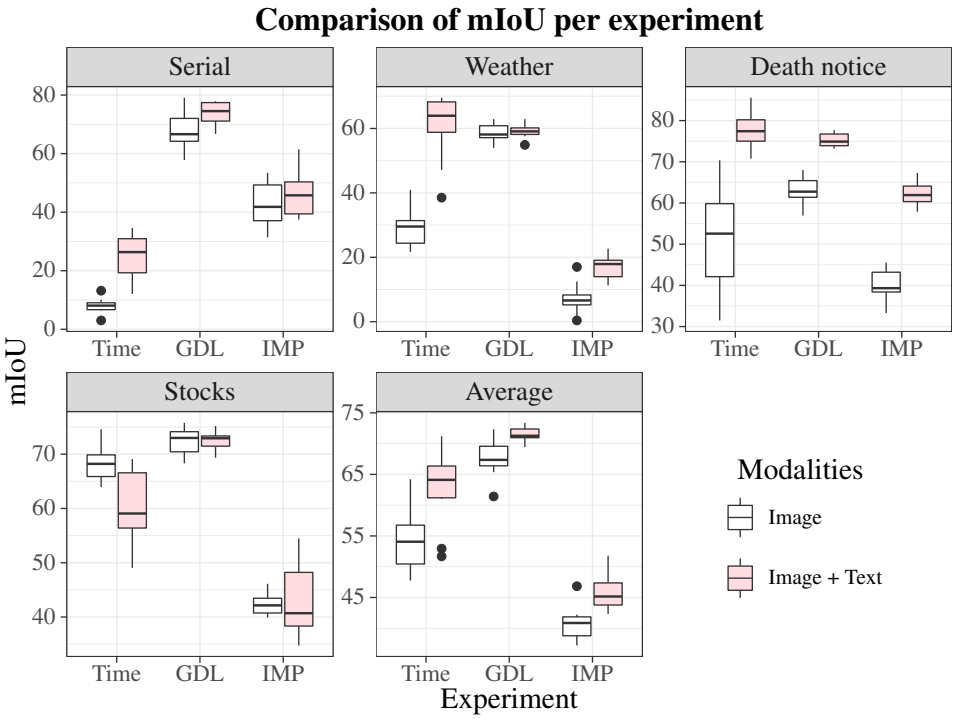


Figure 8: Box plots of the mIoU of the generalization experiments.

for these two classes. The poor results with *Serial* reveal that this class is neither visually nor textually stable over time. Finally, the results of *Stocks* suggest that the visual features are more stable than the textual ones.

When focusing on the model transferability to other newspapers, the overall performance drop is much less pronounced with GDL than with IMP, confirming that JDG and GDL have more common features in terms of layout. In particular, the low score of both models for the class *Weather* in IMP dataset shows that this type of segment has great variability in terms of layout. For IMP, the gain is, once again, particularly significant for *Death Notice* (+23%-31% at 95% confidence), demonstrating that textual features are particularly good at generalizing for this class.

4.2.3 Summary

The overall performance drop between these experiments and the ones in Section 4.1 confirms the variety of newspaper elements, both through time and across newspapers, and stresses the importance of annotated data representatively sampled across time and newspaper. However, it also demonstrates that model generalization and transferability can be improved by the inclusion of textual features.

When considering the scores, most of them are too low to be considered in any practical use case, except maybe for obituaries. This indicates that even though this method shows some promises in terms of raw performance and generalization, it is still too early to use it at large scale without representative annotated data.

4.3 Reducing Training Size

The third series of experiments addresses the following question: Do models combining visual and textual features need less training material?

4.3.1 *Experiment Description*

These experiments assess the effect of reducing the training size by 60%. The new training size is of 792 training samples againts 1,387 in the first experiments of Section 4.1. The distribution of the re-sampled datasets can be found in Table 7. Once again it can be seen that the ratios of each class are similar between training and testing sets, and also w.r.t the experiments that used 100% of the training data.

Class	**Train size (ratio)**	**Test size (ratio)**
Serial	56 (7.07%)	81 (6.81%)
Weather forecast	61 (7.70%)	95 (7.98%)
Death Notice	59 (7.45%)	94 (7.90%)
Stock Exchange Table	92 (11.62%)	183 (15.38%)
Pages w/o annotations	578 (72.98%)	815 (68.49%)

Table 7: Distribution of the classes for the training and testing sets.

4.3.2 *Results and Discussion*

Results are presented in Table 8 and in Figure 9.

Modalities	**# pages**	**Serial**	**Weather**	**Death Notice**	**Stocks**	**Average**
Image	1387	74.12± 7.59	81.27±2.18	75.37±2.98	83.11±0.88	79.30±2.29
Image+Text	1387	76.73± 5.90	81.38±3.34	83.58±2.02	84.43±1.84	82.16±1.72
Image	792	****70.27**± 2.79	69.67±2.76	65.88±6.48	74.73±2.25	***70.80**±2.32
Image+Text	792	49.36±14.22	66.71±3.57	***71.20**±4.03	****77.30**±1.07	68.22±2.84

Table 8: Results for the mIoU metric. Mean metric ± standard deviation of the metric (in %). Stars indicate statistical difference of mean with Image. It is only reported for models using 60% of the data.

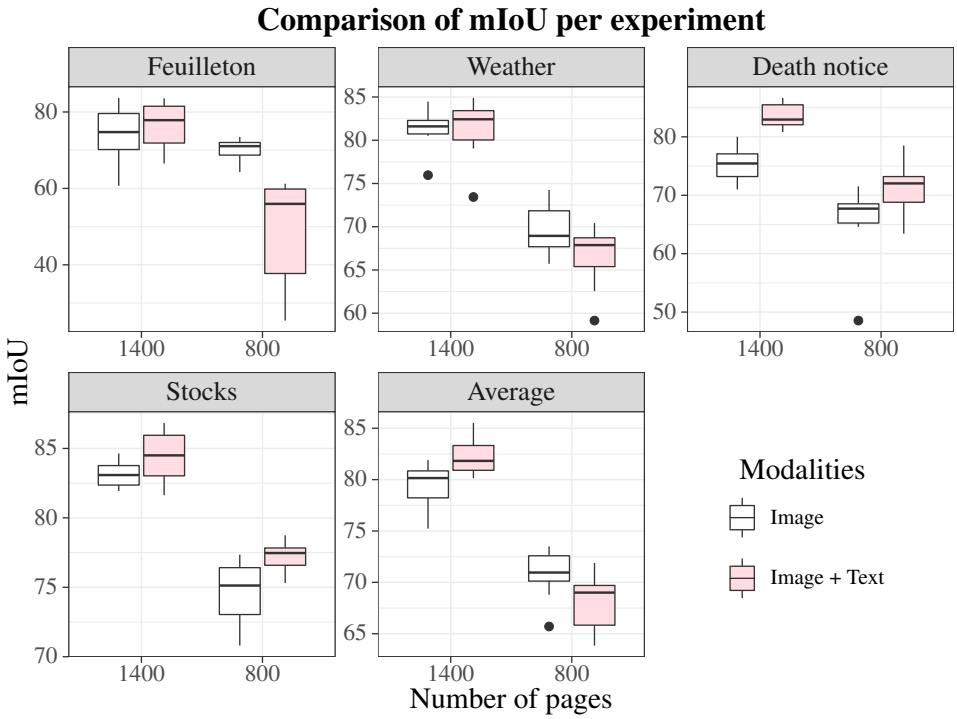


Figure 9: Box plots of the miou of the JDG with 60% training data.

Overall the performance of the models with less data is inferior to the ones using 100% of the

training set. Still considering the average, the drop in mIoU is higher for the Image+Text model than for the Image model. However, this is mainly due to the poor performance of the former on *Serial*. Indeed, the Image model performs significantly better (+10%-31% at 95% confidence) than the Image+Text model. This may be due to the fact that Image+Text has not enough data to learn the textual features of *Serial* and is thus more confused.

Regarding *Death Notice* and *Stocks*, the Image+Text model improves over the Image model. This may indicate that the textual features of these two classes are easier to learn than for *Serial*, and that the model using both text and image features leverages better the small amount of training data for these classes by combining the two signals.

Finally, the fact that *Weather* results have no significant difference between the two models may show that textual features are not as complex to learn as they are for *Serial*, but do not influence much the results.

4.3.3 Summary

This experiment shows that not all classes are equal when the training size is reduced. It indicates that for content items with non focused textual content, such as *Serial*, Image+Text requires more data to efficiently combine both signals. In contrast, it suggests that for more domain specific content items, such as *Stocks* and *Death Notice*, Image+Text model easily leverages the additional signal provided by the textual information.

4.4 Assessing the Benefits of In-domain Embeddings

The fourth series of experiments addresses the following three questions: (a) How big is the advantage if we train in-domain textual embeddings instead of using off-the-shelf embeddings trained on contemporary text data (without noisy OCR)? (b) Can adding more training data compensate for the expected benefit of in-domain embeddings? (c) What is the impact of adding more training material on the performance of the models and is it possible to identify the point at which adding more data becomes ineffective, i.e. a plateau is reached?

4.4.1 Experiment Description

These experiments make uses of several training sets with different numbers of pages, while the test set is kept constant. Training sets of different sizes are iteratively built starting from biggest to smallest by sampling, at each iteration, half of the number of issues per year: the first dataset has 26 issues per year, the next one 13 (therefore a subset of the previous one), the next 6, and so on. Training set statistics are shown in Table 9. The embeddings used are the *BPEmb-flair-multi* and the *fastText-flair-luxwort* stacks (cf. Section 3.2). Each experiment is thus characterized by its amount of training data and the embeddings used (or lack of it).

This experiment uses the newspaper *Luxemburger Wort* and focuses on the *Death Notice* class only since, as seen in previous experiments, it is the one that benefits most from the addition of textual features.

The results are presented in Table 10 and Figure 10. Overall the results show that there is a significant gain in using textual embeddings maps, both in terms of performance and variance.

The Image+Text out-domain and Image+Text in-domain models always significantly beat the Image model. Moreover, the performance of those models increases and the variance decreases with the size of the training set. However, the performance difference between 550 pages and 14,100 pages is only around 5%, even though more than 25 times as many annotated pages are used.

Dataset	# issues/year	# issues	# pages	# death notices (ratio)
Training pages *550*	1	103	562	49 (8.72%)
1100	2	206	1098	100 (9.11%)
1650	3	309	1642	155 (9.44%)
3300	6	618	3290	315 (9.57%)
7000	13	1339	7029	683 (9.72%)
14100	26	2678	14128	1344 (9.51%)
test set	8	824	4825	421 (8.73%)

Table 9: Distribution of the classes for the training and test sets.

# pages	Image	Image+Text out-domain	Image+Text in-domain
550	72.77± 5.11	**80.06±1.75	****83.49**±2.18
1100	76.20± 3.05	****83.12±1.53	*****85.99**±0.79
1650	69.31±13.17	**82.86±1.77	*****85.93**±0.64
3300	69.43± 7.94	***84.38±1.45	***85.88**±0.91
7000	75.20± 5.19	***85.28±0.46	******87.24**±0.31
14100	66.85± 9.48	**85.63±0.54	******87.55**±0.24

Table 10: Results for the mIoU metric. Mean metric ± standard deviation of the metric (in %). The stars of the Image+Text out-domain column indicate the statistical difference of mean w.r.t to Image, and the ones of Image+Text in-domain the difference w.r.t Image+Text out-domain.

In-domain text embeddings are beneficial since they provide a consistent gain (+2-3%) in performance over the out-domain text embeddings. Moreover Image+Text in-domain already surpasses with 1100 examples the performance that Image+Text out-domain achieves with 14100 examples. This is certainly due to the fact that the in-domain embeddings have been trained on enough data to capture the particular semantics of newspapers and the OCR errors present in them.

The results of the model using only the image are quite surprising since the performance decrease with the amount of training data which is counter-intuitive. In order to eliminate the hypothesis that a visual model that uses more data needs more steps for converging, the model with 14,100 pages was trained for the double of training steps, however, it only improved the results by 5%, while still being 5% lower than the model with 1,100 pages and having three times its variance.

4.4.2 Summary

In this experiment, the usage of textual features brings a gain in performances, reduces the variance of the model which converges better even with large amount of data. However, when considering the scores, the addition of training data has little impact and this extra annotation does not seem worth the effort. The use of in-domain embeddings shows a decent improvement that cannot be compensated by more training samples over out-domain embeddings while not requiring additional annotations. Indeed, training in-domain embeddings is done completely unsupervised, making it a worthwhile option if the amount of available text data is sufficiently large.

CONCLUSION AND OUTLOOK

We believe these series of experiments led to a better understanding of the interplay between visual and textual features for semantic segmentation of newspapers document images.

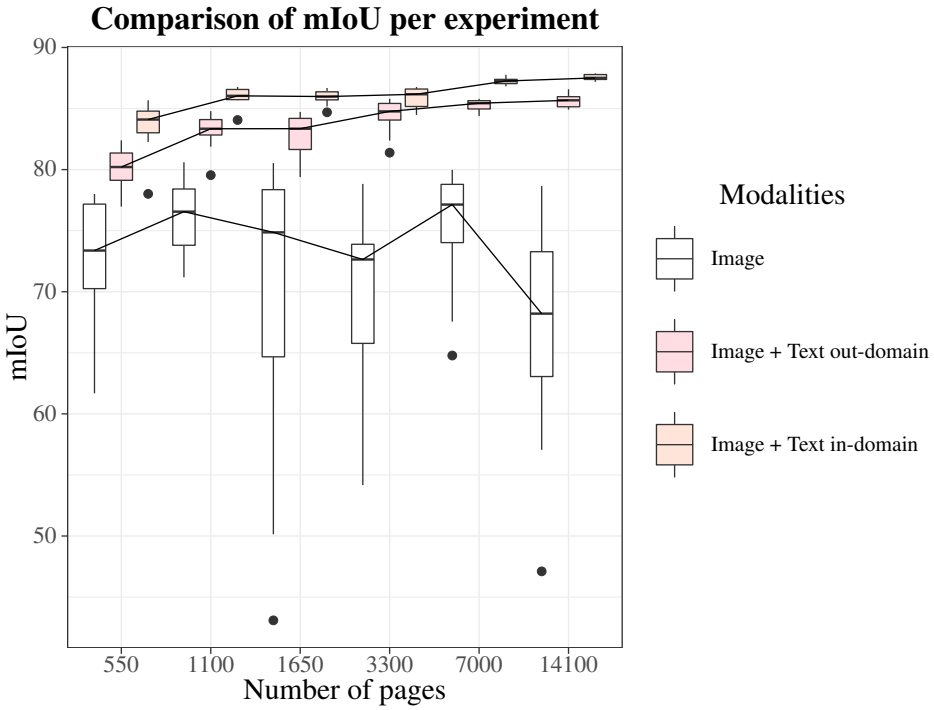


Figure 10: Box plots of the mIoU of the Luxwort experiment.

The first series of experiments using annotation data that was representatively sampled show a consistent improvement for models that combine textual and visual features relative to a strong baseline using visual features only. Textual features also help to mitigate the problem of high variance that purely visual models have with the varying and diverse material in newspapers published over a long period of time.

The second series of experiments on the generalization ability over time and across newspapers showed that a simple transfer of models leads to a stark drop of performance. However, on average, models with textual features show substantially better results than the ones without. We can conclude that text characteristics are indeed more stable than layout characteristics and that they are vital for improving the model's robustness. For practical applications, these experiments make clear that annotation efforts need to be carefully distributed over the diachronic variety and diversity of the material present in historic newspapers.

The third series of experiments on the reduction of training data gave mixed results. On the one hand, the model using both visual and textual still improves on most classes. On the other, there seems to be, for some classes, a lower bound on the number of samples for the proposed method in order to be able to extract the relevant signals from textual features.

The fourth series of experiments on the benefit of using large amounts of training data—by taking advantage of the fact that a lot of material with useful semantic classification already exists in digitized archives— showed that purely visual models have more difficulties to exploit a larger amount of data than combined models. Another important outcome of these experiments is the fact that more training material can not fully compensate for the availability of text embeddings specifically built on in-domain text data, especially with noisy OCR texts. As in-domain text embeddings can be computed without human annotation and are therefore cost-effective, they should always be considered.

Although proof of concept, the present approach can already support two main use cases. First, similarly as *dhSegment*, the released framework can help scholars and/or non-specialists to easily process document images, provided they can be associated with text (thanks to e.g. an open-source OCR software), and that embeddings are available. Second, even though not perfect, the models can already be used as support for manual annotation (users only need to correct false positives or negatives) and, for some classes, be used to segment real newspaper collections to offer further search facets and/or filter unwanted material.

As future works, we intend to compare this approach with pure text classification in order to bridge the comparison spectrum from pure pixel to pure text, as well as to apply it to other documents than newspapers. It could also be interesting to integrate a region proposal module (such as in Mask R-CNN [He et al., 2017b]) in order to segment at instance level.

AUTHOR CONTRIBUTIONS

RB designed and carried out the experiments, wrote the paper; ME designed and supervised the project and experiments, wrote the paper; SC designed and supervised the project and experiments, wrote the paper; SO supervised the project and experiments, helped with the paper writing; FK supervised main directions.

ACKNOWLEDGMENTS

We warmly thank the journal *Le Temps* and the Swiss and Luxembourgish National Libraries for giving us access to their newspaper archive collections in the context of the *impresso* project. We also thank Julien Nguyen Dang for his contribution to the annotation of part of the data. Finally, the second and third authors also gratefully acknowledge the financial support of the Swiss National Science Foundation (SNSF) for the project '*impresso* – Media Monitoring of the Past' under grant number CR-SII5_173719.

References

Željko Agić and Ivan Vulić. JW300: A wide-coverage parallel corpus for low-resource languages. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3204–3210, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1310. URL https://www.aclweb.org/anthology/P19-1310.

Alan Akbik, Duncan Blythe, and Roland Vollgraf. Contextual string embeddings for sequence labeling. In *COLING 2018, 27th International Conference on Computational Linguistics*, pages 1638–1649, 2018.

Sofia Ares Oliveira, Benoit Seguin, and Frédéric Kaplan. dhSegment: A generic deep-learning approach for document segmentation. In *2018 16th International Conference on Frontiers in Handwriting Recognition (ICFHR)*, pages 7–12, August 2018. doi: 10.1109/ICFHR-2018.2018.00011.

Galal M. Binmakhashen and Sabri A. Mahmoud. Document layout analysis: A comprehensive survey. *ACM Comput. Surv.*, 52(6), October 2019. ISSN 0360-0300. doi: 10.1145/3355610. URL https://doi.org/10.1145/3355610.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146, 2017.

Hassina Bouressace and Janos Csirik. Recognition of the logical structure of arabic newspaper pages. In *International Conference on Text, Speech, and Dialogue*, pages 251–258. Springer, 2018.

Vincent Buntinx, Frédéric Kaplan, and Aris Xanthos. Layout analysis on newspaper archives. In *Digital Humanities 2017*, pages 409–412, 2017.

Kai Chen, Mathias Seuret, Jean Hennebert, and Rolf Ingold. Convolutional Neural Networks for Page Segmentation of Historical Document Images. In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, volume 01, pages 965–970, November 2017. doi: 10.1109/ICDAR.2017.161.

Christian Clausner, Christos Papadopoulos, Stefan Pletschacher, and Apostolos Antonacopoulos. The ENP image and ground truth dataset of historical newspapers. In *2015 13th International Conference on Document Analysis and Recognition (ICDAR)*, pages 931–935. IEEE, 2015.

Christian Clausner, Apostolos Antonacopoulos, Stefan Pletschacher, Lotte Wilms, and Steven Claeyssens. PRImA, DMAS2019, Competition on Digitised Magazine Article Segmentation (ICDAR 2019), 2019. URL https://www.primaresearch.org/DMAS2019/.

Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. Natural language processing (almost) from scratch. *Journal of machine learning research*, 12(Aug):2493–2537, 2011.

Tuan Anh Nguyen Dang and Dat Nguyen Thanh. End-to-End Information Extraction by Character-Level Embedding and Multi-Stage Attentional U-Net. In *2019 British Machine Vision Conference*, page 13, 2019.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

Timo I. Denk and Christian Reisswig. BERTgrid: Contextualized embedding for 2d document representation and understanding. In *Workshop on Document Intelligence at NeurIPS 2019*, 2019.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://www.aclweb.org/anthology/N19-1423.

Marco Dinarelli and Sophie Rosset. Tree-structured named entity recognition on OCR data: Analysis, processing and results. In *LREC*, pages 1266–1272, 2012.

Abhishek Dutta and Andrew Zisserman. The VIA annotation software for images, audio and video. In *Proceedings of the 27th ACM International Conference on Multimedia*, MM '19, New York, NY, USA, 2019. ACM. ISBN 978-1-4503-6889-6/19/10. doi: 10.1145/3343031.3350535. URL https://doi.org/10.1145/3343031.3350535.

Sébastien Eskenazi, Petra Gomez-Krämer, and Jean-Marc Ogier. A comprehensive survey of mostly textual document segmentation algorithms since 2008. *Pattern Recognition*, 64:1–14, April 2017. doi: 10.1016/j.patcog.2016.10.023.

Basilios Gatos, SL Mantzaris, KV Chandrinos, A Tsigris, and Stavros J Perantonis. Integrated algorithms for newspaper page decomposition and article tracking. In *Proceedings of the Fifth International Conference on Document Analysis and Recognition. ICDAR'99 (Cat. No. PR00318)*, pages 559–562. IEEE, 1999.

Karim Hadjar and Rolf Ingold. Arabic newspaper page segmentation. In *ICDAR*, volume 3, pages 895–899, 2003.

Dafang He, Scott Cohen, Brian Price, Daniel Kifer, and C. Lee Giles. Multi-scale multi-task FCN for semantic page segmentation and table detection. In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, page nil, November 2017a. doi: 10.1109/icdar.2017.50.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017b.

David Hebert, Thomas Palfray, Stephane Nicolas, Pierrick Tranouez, and Thierry Paquet. Automatic article extraction in old newspapers digitized collections. In *Proceedings of the First International Conference on Digital Access to Textual Cultural Heritage*, pages 3–8, 2014.

Benjamin Heinzerling and Michael Strube. BPEmb: Tokenization-free Pre-trained Subword Embeddings in 275 Languages. In Nicoletta Calzolari (Conference chair), Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Koiti Hasida, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, Stelios Piperidis, and Takenobu Tokunaga, editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May 7-12, 2018 2018. European Language Resources Association (ELRA). ISBN 979-10-95546-00-9.

Sergey Ioffe. Batch renormalization: Towards reducing minibatch dependence in batch-normalized models. In *Advances in neural information processing systems*, pages 1945–1953, 2017.

Frédéric Kaplan and Isabella di Lenardo. Big Data of the Past. *Frontiers in Digital Humanities*, 4, 2017. ISSN 2297-2668. doi: 10.3389/fdigh.2017.00012. URL https://www.frontiersin.org/articles/10.3389/fdigh.2017.00012/full.

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

Mike Kestemont, Folgert Karsdorp, and Marten Düring. Mining the twentieth century's history from the time magazine corpus. *EACL 2014*, page 62, 2014. URL http://www.aclweb.org/anthology/W/W14/W14-06.pdf#page=72.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Thomas Lansdall-Welfare, Saatviga Sudhahar, James Thompson, Justin Lewis, and Nello Cristianini. Content analysis of 150 years of british periodicals. *Proceedings of the National Academy of Sciences*, 114(4):E457–E465, 2017. ISSN 0027-8424. doi: 10.1073/pnas.1606380114. URL https://www.pnas.org/content/114/4/E457.

Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, page nil, June 2015. doi: 10.1109/cvpr.2015.7298965.

Elizabeth Lorang, Leen-Kiat Soh, Maanas Varma Datla, and Spencer Kulwicki. Developing an image-based classifier for detecting poetic content in historic newspaper collections. *D-Lib Magazine*, 21(7/8), 2015.

Benjamin Meier, Thilo Stadelmann, Jan Stampfli, Marek Arnold, and Mark Cieliebak. Fully convolutional neural networks for newspaper article segmentation. In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, page nil, November 2017. doi: 10.1109/icdar.2017.75.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

Jean-Philippe Moreux. Innovative Approaches of Historical Newspapers: Data Mining, Data Visualization, Semantic Enrichment. In *Proceedings of IFLA WLIC 2016*, page 17, Columbus, OH, 2016. URL http://library.ifla.org/id/eprint/2076.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. In *Proc. of NAACL*, 2018.

Nils Reimers and Iryna Gurevych. Why comparing single performance scores does not allow to draw conclusions about machine learning approaches. *arXiv preprint arXiv:1803.09578*, 2018.

Mia Ridge, Giovanni Colavizza, Laurel Brake, Maud Ehrmann, Jean-Phillipe Moreux, and Andrew Prescott. The past, present and future of digital scholarship with newspaper collections. In *DH 2019 Book of Abstracts*, page 9, 2019. URL http://infoscience.epfl.ch/record/271329. Multi-paper panel presented at the 2019 Digital Humanities Conference, Utrecht, July 2019.

Martin Riedl, Daniela Betz, and Sebastian Padó. Clustering-based article identification in historical newspapers. In *Proceedings of the 3rd Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 12–17, 2019.

Christophe Rigaud, Antoine Doucet, Mickaël Coustaty, and Jean-Philippe Moreux. ICDAR 2019 Competition on Post-OCR Text Correction. In *15th International Conference on Document Analysis and Recognition*, Sydney, Australia, September 2019. URL https://hal.archives-ouvertes.fr/hal-02304334.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. *CoRR*, 2015.

Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*, 2015.

Melissa M Terras. The Rise of Digitization. In Ruth Rikowski, editor, *Digitisation Perspectives*, pages 3–20. SensePublishers, Rotterdam, 2011. ISBN 978-94-6091-299-3. doi: 10.1007/978-94-6091-299-3_1. URL http://dx.doi.org/10.1007/978-94-6091-299-3_1http://www.emeraldinsight.com.ezproxy.lancs.ac.uk/doi/full/10.1108/OIR-06-2015-0193.

Melvin Wevers. Using word embeddings to examine gender bias in Dutch newspapers, 1950-1990. In *Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change*, pages 92–97, Florence, Italy, August 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-4712. URL https://www.aclweb.org/anthology/W19-4712.

Christoph Wick and Frank Puppe. Fully convolutional neural networks for page segmentation of historical document images. In *2018 13th IAPR International Workshop on Document Analysis Systems (DAS)*, page nil, April 2018. doi: 10.1109/das.2018.39.

Yue Xu, Wenhao He, Fei Yin, and Cheng-Lin Liu. Page Segmentation for Historical Handwritten Documents Using Fully Convolutional Networks. In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, volume 01, pages 541–546, November 2017. doi: 10.1109/ICDAR.2017.94.

Tze-I. Yang, Andrew J. Torget, and Rada Mihalcea. Topic modeling on historical newspapers. In *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTecH)*, pages 96–104, 2011.

Xiao Yang, Ersin Yumer, Paul Asente, Mike Kraley, Daniel Kifer, and C. Lee Giles. Learning to extract semantic structure from documents using multimodal fully convolutional neural networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, page nil, July 2017. doi: 10.1109/cvpr.2017.462.