Hierarchical Text Segmentation for Medieval Manuscripts

Amir Hazem* Béatrice Daille* Louis Chevalier§ Dominique Stutzmann§
Christopher Kermorvant†‡
* LS2N - Université de Nantes, Nantes
§ IRHT - Institut de recherche et d'histoire des textes, CNRS, Paris
† TEKLIA, Paris
‡ LITIS - Université de Rouen-Normandie, Rouen
{amir.hazem, beatrice.daille}@ls2n.fr
{dominique.stutzmann, louis.chevalier}@irht.cnrs.fr
{kermorvant}@teklia.com

Abstract

In this paper, we address the segmentation of books of hours, Latin devotional manuscripts of the late Middle Ages, that exhibit challenging issues: a complex hierarchical entangled structure, variable content, noisy transcriptions with no sentence markers, and strong correlations between sections for which topical information is no longer sufficient to draw segmentation boundaries. We show that the main state-of-the-art segmentation methods are either inefficient or inapplicable for books of hours and propose a bottom-up greedy approach that considerably enhances the segmentation results. We stress the importance of such hierarchical segmentation of books of hours for historians to explore their overarching differences underlying conception about Church.

1 Introduction

Text segmentation is essential in many downstream applications including document understanding and navigation, summarization, information retrieval and discourse parsing (Purver, 2011; Riedl and Biemann, 2012; Li et al., 2018). Traditional unsupervised approaches assume a high correlation between segments and subtopics. Therefore, based on a prior text decomposition, two adjacent segments are merged if they are highly correlated. On the contrary, if their similarity is below a certain threshold, a shift is determined (Hearst, 1997; Riedl and Biemann, 2012). When sufficient topically annotated training data are available, deep neural approaches based on CNN (Wang et al., 2017) or LSTM (Koshorek et al., 2018) can be efficiently applied (Li et al., 2018; Arnold et al., 2019). Until now, text segmentation methods have exclusively addressed data sets lying within the scope of narrative and expository texts or user dialogues texts and sometimes artificially generated data (Choi, 2000; Jeong and Titov, 2010; Glavaš et al., 2016; Koshorek et al., 2018).

In this paper, we address transcriptions of ancient devotional manuscripts (thereafter also "MS"), from the Middle Ages, known as "books of hours". Books of hours were used by lay people as a guidance in their daily prayers. They represent an important source of information on the late Middle Ages' religious and social practices, and provide opportunities for historical analysis in order to better understand the cultures and faiths of the European society. More than 10,000 manuscripts of ca. 300 pages in average are preserved: they are a specific, standardized, and proto-industrial production (Stutzmann, 2019). However, their textual content is still scarcely studied, because of the lack of transcriptions, the complexity of their liturgical content and their standard appearance. They look very uniform at first glance, but each of them is unique. The content differ from one copy to the other as a large choice of devotional readings are available. Conversely, some prayers are used in several hours of the day and several times within one copy, thus generating section ambiguities. For historians, automating the generation of table of contents is key to understand this complex historical source.

While the building block of the mainstream text segmentation methods is closely related to topical shifts, the liturgical aspect of books of hours exhibits shallow topical relations and a strong correlation between their sections and subsections. As consequence, the topical shift hypothesis becomes inconsistent for this type of data, as has been recently shown in (Hazem et al., 2020). We address the task

Proceedings of the 28th International Conference on Computational Linguistics, pages 6240–6251
Barcelona, Spain (Online), December 8-13, 2020

of books of hours segmentation as a classification problem and propose a greedy two-step bottom-up approach that achieves significant results on books of hours.

2 Related Work

Text segmentation is the task of splitting documents into topically coherent fragments for a better text readability and analysis (Hearst, 1994; Eisenstein, 2009; Glavaš et al., 2016). It is also useful in other NLP and IR (Moens and Busser, 2001) applications such as: summarization, document navigation and indexing, passage retrieval, etc. Segmentation can be content-based where each topic is characterised by a specific vocabulary and each vocabulary change implies a topic change (Hearst, 1994). It can also use topic markers (Fauconnier et al., 2014) whether (i) oral: such as prosody, silence; (ii) written: using connectors, introductory expressions or (iii) visual: using line breaks, bullets, numbering, bold, etc.

A broad range of unsupervised approaches exploit lexical cohesion to detect coherent segments (Hearst, 1997; Choi, 2000) thanks to term repetitions (Hearst, 1994), semantic relations using lexical chains (Morris and Hirst, 1991), dictionary (Kozima, 1993), collocation networks (Ferret et al., 1998), or patterns of lexical co-occurrences (Hearst, 1997) such as discourse structures (Nomoto and Nitta, 1994). Early unsupervised methods include: TextTiling, a TF Cosine based approach (Hearst, 1994), LCSeg, based on lexical chains (Galley et al., 2003), U00, a probabilistic dynamic programming approach (Utiyama and Isahara, 2001), TopicTiling, a topic modeling approach based on Latent Dirichlet Analysis (LDA) (Riedl and Biemann, 2012). Glavaš et al. (2016) proposed a semantic relatedness graph approach that exploits word embeddings. Alemi and Ginsparg (2015) and Naili et al. (2017) studied the contribution of word embeddings on classical segmentation approaches. Text segmentation has also been addressed as a multi-document segmentation problem. Sun et al. (2007) for instance, proposed a method for shared topic detection and topic segmentation of multiple similar documents based on weighted mutual information, while Jeong and Titov (2010) proposed an unsupervised bayesian approach that models both shared and document-specific topics. Supervised approaches have also modeled semantic cohesion. Some methods performed segmentation at the sentence level to discover Elementary Discourse Units (EDU) (Hernault et al., 2010; Joty et al., 2015) while others focused on dialogue.

Neural network approaches have also been applied such as: TextTiling-like embedding approach for query-reply dialogue segmentation (Song et al., 2016), multi-party dialogue for EDU using sequential model (Shi and Huang, 2019) and reinforcement learning (Takanobu et al., 2018). Recently, Li et al. (2018) proposed SegBot, a bidirectional RNN coupled with a pointer network that addresses both topic segmentation and EDU. Also, LSTM or CNN based approaches have been proposed, for instance through bidirectional layers (Sheikh et al., 2017), sentence embedding-based with four layers bidirectional LSTM (Koshorek et al., 2018) or through two symmetric CNN (Wang et al., 2017), etc. Finally, Arnold et al. (2019) proposed Sector, the first LSTM-based architecture that combines topical (latent semantic content) and structural information (segmentation) as a mutual task.

From a resource perspective, the used data sets were mainly linear. A sequential analysis of topical changes was usually applied (Hearst, 1994; Choi, 2000). Expository and narrative texts such as stargazer (Hearst, 1994) newspapers (Ferret et al., 1998) or more general interest articles (Morris, 1988) as well as synthetic data sets (Choi, 2000; Galley et al., 2003) were often used. Later on, data sets with hierarchical structure were addressed, which required a more fine-grained subtopic structure analysis (Yaari, 1997; Eisenstein, 2009). Yaari (1997) proposed one of the first approaches for hierarchical text segmentation: a supervised agglomerative bottom-up clustering method exploiting paragraph hierarchy. A pioneer unsupervised approach for hierarchical text segmentation was introduced by Eisenstein (2009) using a bayesian generative model with dynamic programming. Also, Kazantseva and Szpakowicz (2014) proposed a clustering algorithm based on topical trees to perform hierarchical segmentation. Recently, several data sets have been published showing various types of structures: artificial added to automatic speech recognition transcripts of news videos (Sheikh et al., 2017), encyclopedic reflecting Wikipedia article structure (Koshorek et al., 2018; Arnold et al., 2019) or topical in goal-oriented dialogues (Takanobu et al., 2018). Our data set encompasses books of hours, each of them with a complex original structure described in the following section.

3 Books of Hours

Books of hours constitute a challenging data set. Theirs is a complex hierarchical entangled structure, of interest both for NLP and for historians. They appeared in the 13th century in France, Low Countries, England and later on in many other European countries and became by far the best selling book in the Middle Ages. They include and organize thousands of orations, chants and readings. Some of them are common to all books, but others depend on patron and on gender, geographical location, preferred saints and prayers of the commissioner (Clark, 2003). Most texts are in Latin, but some translated from Hebrew (Psalms) or Greek (Gospel lessons), or written in different periods for distinct purposes (chant, meditation...). In the manuscripts, the texts are not grouped according to their origin or linguistic features, but deliberately mixed to compose an ensemble. The textual contents are barely studied, although books of hours have been intensively studied by art historians for their many and often gorgeous miniature paintings and decorations (Wieck et al., 1988; De Hamel, 1994).

3.1 Granularity

Books of hours are a largely standardized devotional prayer book, containing almost always the same list of core contents. Table 1 gives an overview of sections of different levels and the subsections that may appear within them. However, the section order varies or display omissions and additions, as evidenced in Table 2. Generating the list of contents is of the highest interest as well for the history of cultural practices as for book history and codicology. Indeed, changes and differences may be explained by either cultural and religious factors or by physical ones, and are insights into medieval thoughts. Moreover, books of hours are built as a sequence of many sections with subsections in a highly hierarchized manner with inner cross-references (Stutzmann et al., 2019). For example, the Hours of the Virgin (level 1) assemble ca. 400-500 pieces of different length, organized in eight sections of level 2, corresponding to eight "hours" of the day (hence the generic name), and five to ten subsections of level 3.

Level1 1	Level 2	Level 3
Calendar	January to December	Invitatory
Gospel Lections	Gospel of John	Invocation
	Gospel of Luke	Psalms
	Gospel of Matthew	HSL
	Gospel of Mark	Canticle
Hours of the Virgin	Matins	Orationes
Hours of the Cross	Lauds	Preces
Hours of the Holy Spirit	Prime	Hymn
Office of the Dead	Terce	Lessons
	Sext	Short lesson
Penitential Psalms and Litany	None	
	Vespers	Nicolaus
Suffrages	Compline	Barbara
	Penitential Psalms	Margareta
Prayers	Litany	Maria Magdalena
Obsecro Te		Michael
O intemerata		

Table 1: Books of Hours at three levels of granularity. Each of the four blue sections of level 1 (Hours or Office) may contain one or several blue sections of level 2 (names of the hours), in turn containing several blue sections of level 3.

For historians, automating the generation of table of contents is key, to (i) explore and understand overarching differences in choice of arrangements and choice of texts and office (Baroffio, 2011), (ii) trace the leaves that are cut out or misbound, which happens often because of the numerous miniatures to be found in books of hours, (iii) understand how biblical and liturgical texts were perceived, if they are written separately or interspersed with lay devotional prayers or if their use was dictated by the Church or chosen freely by copyists, (iv) study the hierarchy and ranking of saints, including a gendered approach of both sanctity and readership, (v) explore systemic changes in the insertion of optional texts (Clark, 2003), (vi) discover and systematically explore new texts, such as *orationes* and hymns, for which generations of scholars have compiled repertories.

Books of Hours									
Abbeville	**Auxerre**	**Beaune54**	**Beaune55**	**Caen**	**Arsenal637**	**Arsenal651**	**Versailles**	**Geneva**	**Zurich**
Calendar	Calendar	Calendar	Calendar	Calendar	Calendar	Calendar	Calendar	-	Calendar
Gospels	Gospels	Gospels	Gospels	Virgin	Gospels	Gospels	Gospels	Virgin	Gospels
Cross	Cross	Virgin	Virgin	Cross	Obsecro	Cross	Cross	PP L	Obsecro
Spirit	Spirit	Cross	Cross	Spirit	O intem	Spirit	Spirit	Prayers	O Intem
Virgin	Virgin	Spirit	Spirit	Prayers/Suffrages	Virgin	Virgin	Virgin	Gospels	Prayers
PP L	PP L	Suffrages	PP L	Virgin	Cross	PP L	PP L	Prayers	Virgin
Dead	Dead	PP L	Dead	Cross	Spirit	Dead	Obsecro	-	Cross
-	Obsecro	Dead	Suffrages	Spirit	PP L	Suffrages	-	-	Spirit
-	O Intem	Obsecro	Prayers	PP L	Dead	Obsecro	-	-	PP L
-	-	O Intem	Suffrages	Gospels	-	O Intem	-	-	Dead
-	-	Prayers	Prayers	Obsecro	-	Suffrages	-	-	Prayers
-	-		Obsecro	Suffrages/Prayers	-	-	-	-	-
-	-	-	O Intem	Dead	-	-	-	-	-

Table 2: Structure comparison at the first Level. Gospels for Gospel Lections, Obsecro for Obsecro Te, O Intem for O intemerata, Virgin for Hours of the Virgin (same annotations for Cross and Spirit for Holy Spirit), Dead for Office of the Dead, PP L for Penitential Psalms and Litany.

3.2 Bible, Clergy, Private Persons: selection and order of texts of level 1

Some higher-level sections do not appear equally over time. The selection reflects either a personal taste or broader habits, it also mirrors underlying conception about Church, clergy, Bible and prayers in relation to the Divine. Here, some examples. (i) The gospel lessons, excerpts of the Bible that are used in Mass, are a set of texts that enjoyed a fashion in the very late Middle Ages (mostly from 1470 onwards), but the appearance and position are not explained. The gospels settle at the start, between Calendar and Hours of the Virgin, but some early examples prove that they were first a subsidiary addition as other "average" prayers. (ii) Prayers and suffrages are, in Church theory, completely different sets of texts: the former express a personal creation while the later are liturgical pieces that are also pronounced by clerics in collective settings. In many books of hours, suffrages are interspersed in a larger prayer section. This may reveal a shift in the perception of suffrages that are taken out of the liturgical ensemble and adopted for a more familiar use. (iii) Some books of hours add or remove specific prayers such as *Obsecro Te*, or add one or several votive masses, as an indication of wealth and diversity of devotional practices.

3.3 Rationality and Saints: selection and order of texts of level 2

In their organization at a lower level, tracing the list of texts helps to uncover the mindset of medieval readers. (i) "Mixed hours" demonstrate a complete reorganization of the content, divided not by office (Virgin, Cross, Holy Spirit), but by hour (Matins, Lauds, etc.). They are another conception of the organization of texts and may correspond to other ways of reading texts. Mostly known in Western France, their origin and developments can only be traced on the larger corpus with an automated description. (ii) Several examples show inversions of some hours, be it with a standard text and an erroneous rubric, or with inverted segments. The reason for this is still unknown. (iii) Suffrages are devoted to saints or divine mysteries. These texts are expressions of particular preferences of patrons. Each manuscript contains on average 15 different ones, but may contain several dozens. Each time, it is a singular selection among thousands of saints, allowing to trace specific fears and hopes of individuals, such as the suffrage of saint Margaret, known at that time to help pregnant women for the delivery. (iv) Moreover, the saints are hierarchized and ordered, and it is assumed that the most important ones are first, typically with male saints first. Exceptions help tracing female patrons and idiosyncratic conceptions.

3.4 Discovering New Texts: texts of level 3

For several decades, cataloguing guidelines have directed scholars to record the outlines of offices and hours. Indicating in a systematic manner where sections of hymns and orationes start will help building a larger survey of these pieces of which generations of scholars have compiled repertories. This research logic would then be extended to additional text sorts, such as absolutions in order to discover new texts.

	Use		Counts				Number of segments		
Books	Virgin	Dead	Images	Pages	Ps.-Lines	Ps.-Words	Level1	Level2	Level3
Abbeville 15	Amiens	Paris	130	238	3,475	12,730	7	29	68
Auxerre Cathedral 14	Troyes	Troyes	164	318	4,520	17,165	9	33	77
Beaune 54	Autun	Autun	152	292	3,806	16,729	11	36	85
Beaune 55	Autun	Autun	151	290	4,302	21,615	11	35	89
Caen MBA FMM.273	Coutances	Coutances	133	250	3,145	15,196	10	34	89
Paris BnF Arsenal 637	Rome	Rome	315	294	4,323	15,652	9	33	80
Paris BnF Arsenal 651	Troyes	Troyes	267	259	3,972	17,191	10	33	85
Versailles 1688	Troyes	Troyes	182	340	1,344	4,659	7	19	26
Geneva Comites Lat. 38	Rome	Rome	206	166	3,927	20,407	4	11	45
Zurich Rheinau 169	Besançon	Besançon	360	353	4,444	16,562	10	34	80

Table 3: Illustration of the number of pseudo-lines and pseudo-words and their corresponding number of segments per level.

4 Data Set

4.1 Ten Books of Hours

The data set is composed of 10 books of hours, mostly of the 15th c., of diverse geographic origins and containing texts which are adapted for different places in France (Amiens, Autun, Paris...) and Italy (Rome), which are not always their place of production. One MS such as Abbeville 15 has the specific Hours of the Virgin for the use of Amiens and an Office of the Dead for the use of Paris. Table 2 illustrates the order of the first level sections in all 10 books. We first observe that some structure similarities can be shared by several books. For instance, most of books of hours start with Calendars, followed by Gospel Lections, *Obsecro Te* is often followed by *O intemerata*. A group combines the Hours of the Virgin, of the Cross, and of the Holy Spirit (either as Virgin-Cross-Spirit or as Cross-Spirit-Virgin), followed by the Penitential psalms and the office of the Dead. However, many exceptions can be observed. For instance, some books do not contain some sections. The manual observation of level 1 sections clearly indicates that the overall structure is proper to each book.

Table 3 indicates the number of pages[1], pseudo-words/lines as well as of segments per level. At the first level, the number of shifts is stable[2]. It varies much more for the second and third levels.

4.2 Automated Transcription

The automated transcription of the text from the manuscripts images is done in two steps: automated text line detection and automated handwritten text recognition (HTR). The text line detection is performed using a deep neural network trained on multiple databases of historical documents (Boillet et al., 2020) which predicts the text line positions at pixel level. A post-processing is then used to define an oriented rectangle for each text line. The handwritten text recognition is performed with a model built using the Kaldi library (Arora et al., 2019). The optical part of the model, which is responsible for recognizing the shape of the letters, is trained using 7,224 text lines manually transcribed, extracted from several manuscripts in French and Latin with similar handwriting and content but no books of hours. The language model, which is responsible for reconstructing plausible sentences, was trained on 903,281 words from prayers and the Bible both in Latin and French. The estimated Character Error Rate is 9.9% and the corresponding Word Error Rate is 26.5%. The automatic processing of the 10 manuscripts produced 37,258 pseudo-lines and 157,906 tokens or pseudo-words[3]. Figure 2 illustrates line segmentation and text recognition. The recognized text contains errors. The original text is ”tuum. Deo gr(ati)as. S(e)c(un)d(u)m Matheum”, containing several abbreviations, here in brackets, while the automated transcription is: ”tuum , sigras secundum infitheum”. Some abbreviations are correctly expanded, but words may be incorrectly recognized and even agglutinated.

[1]We distinguish between the number of digital images listed in the IIIF manifest, including technical signs, color palette and bindings, and the actual number of pages of the MS, without flyleaves, but with blank pages belonging to the original quires.

[2]A shift is a boundary between two segments or categories.

[3]Tokens are a sequence of one or more letters, which would ideally correspond to the words written in the manuscript.

(a) Line detection (in red) and automated HTR (in tooltip "Domine labia me")

(b) Manual annotation (in green) of sections (in tooltip "Preces")

Figure 1: Examples of line detection, Handwritten Text Recognition, and manual annotation of sections in MS Beaune 54.

Figure 2: Line detection (in red) and HTR transcription ("tuum, sigras secundum infitheum") in MS Paris, BnF, Arsenal 637.

4.3 Reference Annotation

Annotating the sections and subsections is a difficult task due to the characteristics and complexity of books of hours (length, text, Gothic script). Moreover, manuscripts often do not indicate the start and end of sections and subsections. Human annotators have to identify, based on their first words, the first items within a section to annotate them. Yet, those texts may vary or be omitted, which requires a large flexibility and is immensely helped by prior knowledge. The annotation interface was provided by Arkindex[4], a web platform dedicated to automated processing of large collections of scanned documents. The annotation work consisted in drawing rectangular boxes around the first line of each (sub)section and adding the name and level of the (sub)section. More than 4,500 annotations were done on the 10 manuscripts. For the start of a section of level 1, annotators had to annotate several times the first line to indicate the start of all three levels of (sub)sections (e.g. L1 = Hours of the Virgin > L2 = Matins > L3 = Invocation). The manual annotation involved four Humanists, two experts (IA1 and IA3) at post doctoral level and two students (IA2 and IA4) with a master's degree in history. A 54-pages annotation guide supported the work, indicating the clues to identify the beginning of each section at every level. The duration varied from approx. 1 hour to more than 2h30 per MS and per annotator. As expected, more expert annotators were faster (11 and 13 hours vs. 19 and 26 hours) and more accurate.

4.4 Inter-Annotator Agreement

Given that manual annotations (free rectangular shapes) and automated layout segmentation and text recognition are not connected, the evaluation of inter-annotator agreement as well as the gold standard for the present task of text segmentation were created as follows. An automatically segmented line and the corresponding recognized text are considered as the start of a section if the center of an annotation (rectangular shape) lies within its boundaries.Two annotations are considered equal if their centers lie in the same automatically segmented line. The inter-annotator agreement is represented by a Fleiss kappa of 0.9 which is an almost perfect agreement. Overall, IA2 and IA4 missed each around 100 section starts (8.4%) and unexpected changes, for manuscripts whose structure is more complex than usual or when misbound pages created an unorderly sequence.

[4]https://arkindex.teklia.com/

5 Greedy Approach

To segment books of hours we propose a semi-supervised approach that takes advantage of the sequential structure of book of hours which models the liturgical cycles of prayers. We perform a hierarchical segmentation adopting a bottom up strategy starting from lines and inferring superior structure levels. This strategy is motivated by the lack of large amounts of annotated sections of books of hours (see Table 3)[5]. Inspired by sequence labeling approaches, we consider each section as a set of lines for which we assign the corresponding label. We first train a classifier to identify the label of each line, then, we merge nearby blocks with the same label so that single misclassified lines do not introduce spurious section breaks (See Figure 3).

Transcription	Gold	Prediction
obumbrabit tibi ideoque et quod	Gospel Lections	Gospel Lections
nascetur ex te sanctum vocabitur	Gospel Lections	Gospel Lections
filius dei et ecce elezabeth cogna	Gospel Lections	Gospel Lections
ta tua et ipse concepit filium in	Gospel Lections	Gospel Lections
senectutem suam et hic mensis est	Gospel Lections	Gospel Lections
sextus illi que vocatur sterilis	Gospel Lections	Gospel Lections
quia non erit nepossibile apud	Gospel Lections	Gospel Lections
deum domine uerbum dixit au	Gospel Lections	Gospel Lections
teni maria lacte ancilla domini	Gospel Lections	Obsecro Te
fiat michi secundum uerbum	Gospel Lections	Gospel Lections
tuum sigras secundum infitheum	Gospel Lections	Prayers
ecce magi abb oriente venerunt	Gospel Lections	Gospel Lections
iherosolymam dicentes ubi est	Gospel Lections	Gospel Lections
qui natus est rex iudeorum diduntur	Gospel Lections	Gospel Lections
enum stellam eius in oriente et	Gospel Lections	Gospel Lections

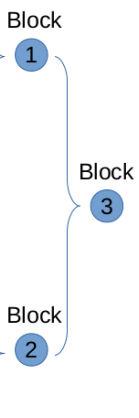

Figure 3: Illustration of the bottom-up greedy segmentation of an extract of the MS Arsenal 637. The first column refers to transcriptions, the second column to the gold reference of level 1 (Gospel Lections), and the third to the predicted labels of the SVM classifier.

5.1 Line Classification

We consider the task of segmentation as a classification problem at the line break level. Each line is represented by its corresponding section's label. We assume that if enough lines of a given section are correctly classified, segmentation can be efficiently performed thanks to a greedy merging approach. To perform line classification, we chose to experiment with Support Vector machines (SVM) and BERT (Devlin et al., 2019)[6]. For SVM, Tf-Idf features are calculated over unigrams and bigrams at the line level. We also used BERT for multi-class classification (the multi-class resides in all the section labels as depicted in Table 1). Each line is associated with its corresponding class (of level 1, level 2 or level 3). Then, BERT is trained to predict the section label of each line. We also, used BERT for sentence pair classification assuming that more information can be captured if we take advantage of the next line of books of hours to predict the current line label. We refer to this approach as BERT* by contrast to BERT as single line classification.

5.2 Segmentation

Once section labels have been assigned to each line (see for example the third column of Figure 3), we perform segmentation using a greedy forward-backward strategy. Our approach is bottom-up, which means that it starts at the line level to reach the segment level using a greedy process. We consider a change in line's label as a new section boundary. Assuming that lines classification may convey errors due to short lines and transcription errors (See the "Prediction" column of Figure 3 where the prediction errors are shown in red), we introduce an error tolerance threshold, that we call a relaxation factor. The relaxation factor is the number of misclassified lines that appear between two blocks of the same label and that we decide to ignore in our merging process. Basically, we first identify all the lines sequences of the same label (See Figure 3, blocks 1 and 2), then, starting from the longest sequence of a given label, we walk forward (respectively backward) and merge the next block of the same label if the distance between the blocks is lower than the relaxation factor. In Figure 3 for instance, we first detect block 1 which has the longest *Gospel Lections* sequence (8 consecutive similar labels), then block 2 with 4 similar labels of *Gospel Lections*. Between these two blocks, we see two miclassified lines (Obsecro Te and Prayers). Based on the relaxation factor (empirically fixed between 50 and 100 in our experiments),

[5]It is expected that traditional classifiers will not be able to perform classification at the section level.

[6]Other classifiers such as: Naive Bays, fastText and others were addressed with no prominent results.

we ignore these lines and merge blocks 1 and 2 and produce a third block that corresponds to the new *Gospel Lections* section. We repeat this procedure recursively for each label of each level of granularity. Our approach takes into account the hierarchy because when dealing with sections of level 2 or 3, we create a new set of labels which combines level 1 and 2 if we deal with level 2 and combines levels 1, 2 and 3 if we deal with level 3.

6 Experiments and Results

6.1 Experimental Setup

Data To test SVM and BERT classifiers as well as books of hours segmentation, we used four books of hours (Arsenal 637, Beaune 55, Caen FMM.273 and Zurich Rh.169). The 6 remaining books were used for training. As books of hours are mostly written in Latin, we used the bert-base-multilingual-cased model. For the fine-tuning phase of BERT, we used the simpletransformers[7] library and its default parameters setting with 50 epochs[8].

Baselines We evaluated: (i) five unsupervised approaches: TextTiling (Hearst, 1994), C99 (Choi, 2000), U00 (Utiyama and Isahara, 2001), MinCut (Malioutov and Barzilay, 2006), HierBays (Eisenstein, 2009). Due to the lack of large annotated training data, we did not evaluate other classifiers-based and deep learning-based approaches on the book of hours corpus.

Evaluation Metrics The approaches are evaluated in terms of P_k (Beeferman et al., 1999) and Windowdiff (WD) (Pevzner and Hearst, 2002) metrics. P_k is an error metric which combines precision and recall to estimate the relative contributions of the different feature types. Nonetheless, it exhibits several drawbacks. P_k is affected by segment size variation. It also penalizes more heavily false negatives than false positives and overpenalizes near misses. Hence, a second measure, WD, a variant of P_k, is also used as it equally penalizes false positives and near misses.

6.2 Results

	SVM			**BERT**			**BERT***		
	Level 1	**Level 2**	**Level 3**	**Level 1**	**Level 2**	**Level 3**	**Level 1**	**Level 2**	**Level 3**
Arsenal637	**73.89**	60.02	48.79	67.56	51.54	40.37	73.41	**62.02**	**51.20**
Beaune55	**68.39**	**54.74**	47.85	61.17	45.53	39.65	65.53	54.22	**50.23**
Caen FMM.273	**67.78**	57.30	53.14	62.06	49.68	44.15	67.41	**58.53**	**53.41**
Zurich Rh.169	66.63	46.42	43.63	63.10	41.54	36.36	**69.13**	**47.61**	**45.26**

Table 4: Classification results (Accuracy) at line level on four test manuscripts

Table 4 reports line classification results of SVM, BERT and BERT* at three levels of granularity. The accuracy differs with regard to each book of hours and to the addressed level. Overall, the best performing classifiers are SVM and BERT*. BERT obtains lower results, this may suggest that considering the next sentence in the line classification model is of some help in detecting lines class. However, the results are low in many cases, especially for levels 2 and 3. This could be either due to the transcription quality or to the fact the books of hours lines are often of small size (around 10 tokens per line) which limits the conveyed sequence information.

The results of segmentation on Books of hours are shown in Table 5. Our proposed model (Greedy) is examined in three different settings: SVM (Greedy (SVM)), BERT as single sentence classification (Greedy (BERT)) and BERT as sentence pair classification (Greedy (BERT*)). Our greedy models achieved better results compared to the baselines. This is particularly remarkable on the first level of segmentation where the Greedy (SVM) approach obtained a P_k of 0.01% and a WD of 0.09% for Arsenal 637. Overall, our best performing model is Greedy (SVM). Greedy (BERT) and Greedy (BERT*) obtain

[7] https://github.com/ThilinaRajapakse/simpletransformers

[8] The used data sets as well as our greedy approach can be found at https://github.com/hazemAmir/Greedy_Text_Segmentation

	Arsenal 637						Beaune 55					
	Level 1		**Level 2**		**Level 3**		**Level 1**		**Level 2**		**Level 3**	
	P_k	WD	P_k	WD	P_k	WD	P_k	WD	P_k	WD	P_k	WD
TextTiling	0.67	0.87	0.51	0.54	0.38	0.40	0.67	0.87	0.54	0.60	0.42	0.44
C99	0.69	0.96	0.62	0.81	0.47	0.56	0.61	0.90	0.50	0.63	0.36	0.41
U00	0.26	0.30	0.30	0.30	0.31	0.31	0.31	0.32	0.36	0.37	0.37	0.37
MinCut	0.43	0.50	0.45	0.50	0.44	0.50	0.52	0.58	0.42	0.48	0.44	0.52
HierBayes	0.22	0.30	0.28	0.30	0.31	0.32	0.31	0.40	0.35	0.38	0.37	0.38
Greedy (SVM)	**0.01**	**0.09**	0.10	**0.15**	0.16	0.25	**0.06**	**0.12**	0.16	0.22	0.21	0.25
Greedy (BERT)	0.02	0.11	0.10	0.16	0.18	0.29	0.09	0.16	0.19	0.24	0.21	0.26
Greedy (BERT*)	0.10	0.22	**0.09**	0.17	**0.14**	**0.24**	0.12	0.17	**0.12**	**0.18**	**0.18**	**0.24**
	Caen FMM.273						**Zurich Rh.169**					
	Level 1		**Level 2**		**Level 3**		**Level 1**		**Level 2**		**Level 3**	
	P_k	WD	P_k	WD	P_k	WD	P_k	WD	P_k	WD	P_k	WD
TextTiling	0.54	0.60	0.47	0.51	0.35	0.38	0.65	0.88	0.51	0.55	0.38	0.40
C99	0.64	0.84	0.63	0.82	0.42	0.50	0.66	1.0	0.57	0.89	0.50	0.62
U00	0.32	0.33	0.30	0.32	0.36	0.37	0.25	0.30	0.34	0.35	0.36	0.37
MinCut	0.39	0.48	0.37	0.46	0.41	0.48	0.43	0.50	0.46	0.50	0.39	0.46
HierBayes	0.27	0.29	0.29	0.31	0.35	0.36	0.14	0.31	0.33	0.38	0.36	0.37
Greedy (SVM)	**0.21**	**0.25**	0.14	0.22	**0.21**	**0.27**	**0.06**	**0.11**	**0.12**	**0.19**	0.18	**0.25**
Greedy (BERT)	0.22	0.26	0.13	0.21	0.23	0.28	**0.06**	0.14	0.14	0.22	0.21	0.27
Greedy (BERT*)	**0.21**	0.26	**0.09**	**0.16**	0.22	0.30	**0.06**	0.12	0.15	0.22	**0.16**	**0.25**

Table 5: Segmentation results using P_k and WD for segmentation. P_k and WD are penalties, so lower scores are better.

competitive results. Greedy (BERT*) is sometimes better than Greedy (SVM) (For Caen FMM.273 at level 2 with a P_k of 0.09% and a WD of 0.16% for instance). The lower results for levels 2 and 3 can be partially explained by the higher line classification errors (See Table 4) and also by the higher number of segments to detect. Conversely to the baselines which are unsupervised and only draw boundaries without section identification, our approach can assign to each section the granularity level required for historian studies. Despite low classification results, especially for levels 2 and 3, our proposed greedy algorithm obtained good segmentation results. This confirms the usefulness of using a relaxation factor to reduce the impact of wrong classifications. Nonetheless, our relaxation factor is fixed empirically on a development set. In future work , we plan to develop a strategy that allows the use of a dynamic relaxation factor for better accuracy. Finally, one of the main advantages of our strategy is that it can be used with any classifier. Improving the classification results will undoubtedly has a positive impact on the segmentation.

7 Conclusion

The present research evidences that hierarchical text segmentation can be achieved for medieval manuscripts and that our models can overcome the difficulties of both an error-prone text and a very correlated text with many repetitions. From the historians perspective, the results are already deemed very useful. The level of expertise and time required for manual annotation are very high. Even if some shifts are slightly misplaced (sometimes a few lines above or below), users can rely on the automated creation of tables of contents, pointing correctly to the right page (and not to the line), as is usual in Medieval studies. Historians are already enthusiastic about these results, because this text segmentation is applied for the first time to medieval manuscripts, based on HTR and not on time-consuming scholarly edition. Combined with an expected increase in HTR accuracy, the proposed hierarchical text segmentation approach supports textual criticism and cultural studies.

Acknowledgments

This work is part of the HORAE project (Hours - Recognition, Analysis, Editions) and is supported by the French National Research Agency under grant ANR-17-CE38-0008.

References

Alexander A. Alemi and Paul Ginsparg. 2015. Text segmentation based on semantic word embeddings. CoRR, abs/1503.05543.

Sebastian Arnold, Rudolf Schneider, Philippe Cudré-Mauroux, Felix A. Gers, and Alexander Löser. 2019. SECTOR: A neural model for coherent topic segmentation and classification. Transactions of the Association for Computational Linguistics, 7:169–184.

Ashish Arora, Chun Chieh Chang, Babak Rekabdar, Daniel Povey, David Etter, Desh Raj, Hossein Hadian, Jan Trmal, Paola Garcia, Shinji Watanabe, Vimal Manohar, Yiwen Shao, and Sanjeev Khudanpur. 2019. Using ASR methods for OCR. In Proceedings of the 15th International Conference of Document Analysis and Recognition (ICDAR 2019), pages 663–668, Sydney, Australia.

Giacomo Baroffio. 2011. Testo e musica nei libri d'ore. Rivista Italiana di Musicologia, (46):18–77.

Doug Beeferman, Adam Berger, and John Lafferty. 1999. Statistical models for text segmentation. Machine Learning, 34(1-3):177–210.

Mélodie Boillet, Christopher Kermorvant, and Thierry Paquet. 2020. Multiple document datasets pre-training improves text line detection with deep neural networks. In Proceedings of the 25th International Conference on Pattern Recognition (ICPR 2020), Milan, Italy.

Freddy Y. Y. Choi. 2000. Advances in domain independent linear text segmentation. In Proceedings of the 1st North American Chapter of the Association for Computational Linguistics Conference (NAACL 2000), pages 26–33, Seattle, Washington.

Gregory T. Clark. 2003. The Spitz Master. A Parisian Book of Hours. Getty Museum Studies on Art publications, Los Angeles.

Christopher De Hamel. 1994. A history of illuminated manuscripts. Phaidon P, London.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 17th Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (NAACL-HLT 2019), pages 4171–4186, Minneapolis, Minnesota, USA.

Jacob Eisenstein. 2009. Hierarchical text segmentation from multi-scale lexical cohesion. In Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, pages 353–361, Boulder, Colorado, USA.

Jean-Philippe Fauconnier, Laurent Sorin, Mouna Kamel, Mustapha Mojahid, and Nathalie Aussenac-Gilles. 2014. Détection automatique de la structure organisationnelle de documents à partir de marqueurs visuels et lexicaux. In Actes de la 21e conférence sur le Traitement Automatique des Langues Naturelles (TALN 2014), pages 340–351, Marseille, France.

Olivier Ferret, Brigitte Grau, and Nicolas Masson. 1998. Thematic segmentation of texts: Two methods for two kinds of texts. In Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics - Volume 1 (ACL 1998 / COLING 1998), pages 392–396, Montreal, Quebec, Canada.

Michel Galley, Kathleen McKeown, Eric Fosler-Lussier, and Hongyan Jing. 2003. Discourse segmentation of multi-party conversation. In Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1 (ACL 2003), pages 562–569, Sapporo, Japan.

Goran Glavaš, Federico Nanni, and Simone Paolo Ponzetto. 2016. Unsupervised text segmentation using semantic relatedness graphs. In Proceedings of the 5th Joint Conference on Lexical and Computational Semantics, pages 125–130, Berlin, Germany.

Amir Hazem, Beatrice Daille, Christopher Kermorvant, Dominique Stutzmann, Marie-Laurence Bonhomme, Martin Maarand, and Mélodie Boillet. 2020. Books of hours. the first liturgical data set for text segmentation. In Proceedings of The 12th Language Resources and Evaluation Conference (LREC 2020), pages 776–784, Marseille, France.

Marti A. Hearst. 1994. Multi-paragraph segmentation expository text. In Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics (ACL 1994), pages 9–16, Las Cruces, New Mexico, USA.

Marti A. Hearst. 1997. Texttiling: Segmenting text into multi-paragraph subtopic passages. Computational Linguistics, 23(1):33–64.

Hugo Hernault, Danushka Bollegala, and Mitsuru Ishizuka. 2010. A sequential model for discourse segmentation. In Proceedings of the 11th International Conference on Computational Linguistics and Intelligent Text Processing (CICLing 2010), pages 315–326, Iasi, Romania.

Minwoo Jeong and Ivan Titov. 2010. Multi-document topic segmentation. In Proceedings of the 19th ACM International Conference on Information and Knowledge Management (CIKM 2010), pages 1119–1128, Toronto, Canada.

Shafiq Joty, Giuseppe Carenini, and Raymond T. Ng. 2015. Codra: A novel discriminative framework for rhetorical analysis. Computational Linguistics, 41(3):385–435.

Anna Kazantseva and Stan Szpakowicz. 2014. Hierarchical topical segmentation with affinity propagation. In Proceedings of the 25th International Conference on Computational Linguistics (COLING 2014), pages 37–47, Dublin, Ireland.

Omri Koshorek, Adir Cohen, Noam Mor, Michael Rotman, and Jonathan Berant. 2018. Text segmentation as a supervised learning task. In Proceedings of the 16th Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (NAACL-HLT 2018), pages 469–473, New Orleans, Louisiana.

Hideki Kozima. 1993. Text segmentation based on similarity between words. In Proceedings of the 31st Annual Meeting on Association for Computational Linguistics (ACL 1993), pages 286–288, Columbus, Ohio, USA.

Jing Li, Aixin Sun, and Shafiq Joty. 2018. Segbot: A generic neural text segmentation model with pointer network. In Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence (IJCAI 2018), pages 4166–4172, Stockholm, Sweden.

Igor Malioutov and Regina Barzilay. 2006. Minimum cut model for spoken lecture segmentation. In Proceedings of the 21st International Conference on Computational Linguistics (COLING 2006) and 44th Annual Meeting of the Association for Computational Linguistics (ACL 2006), pages 25–32, Sydney, Australia.

Marie-Francine Moens and Rik De Busser. 2001. Generic topic segmentation of document texts. In Proceedings of the 24th Annual International Conference on Research and Development in Information Retrieval (SIGIR 2001), pages 418–419, New Orleans, Louisiana, USA.

Jane Morris and Graeme Hirst. 1991. Lexical cohesion computed by thesaural relations as an indicator of the structure of text. Computational Linguistics, 17(1):21–48.

Jane Morris. 1988. Lexical cohesion, the thesaurus, and the structure of text. Technical report CSRI-219. Computer system research institute.

Marwa Naili, Anja Habacha Chaibi, and Henda Hajjami Ben Ghezala. 2017. Comparative study of word embedding methods in topic segmentation. In Proceedings of the 21st International Conference on Knowledge-Based and Intelligent Information Engineering Systems (KES 2017), pages 340 – 349, Marseille, France.

Tadashi Nomoto and Yoshihiko Nitta. 1994. A grammatico-statistical approach to discourse partitioning. In Proceedings of the 15th Conference on Computational Linguistics - Volume 2 (COLING 1994), pages 1145–1150, Kyoto, Japan.

Lev Pevzner and Marti A. Hearst. 2002. A critique and improvement of an evaluation metric for text segmentation. Computational Linguistics, 28(1):19–36.

Matthew Purver. 2011. Topic segmentation. Spoken Language Understanding: Systems for Extracting Semantic Information from Speech. Wiley.

Martin Riedl and Chris Biemann. 2012. Topictiling: A text segmentation algorithm based on lda. In Proceedings of the 50th Annual Meeting on Association for Computational Linguistics (ACL 2012), Student Research Workshop, pages 37–42, Jeju Island, Korea.

Imran Sheikh, Dominique Fohr, and Irina Illina. 2017. Topic segmentation in ASR transcripts using bidirectional rnns for change detection. In Proceedings of IEEE Automatic Speech Recognition and Understanding Workshop (ASRU 2017), Okinawa, Japan.

Zhouxing Shi and Minlie Huang. 2019. A deep sequential model for discourse parsing on multi-party dialogues. In Proceedings of the 33rd Conference on Artificial Intelligence (AAAI 2019), pages 7007–7014, Honolulu, Hawaii.

Yiping Song, Lili Mou, Rui Yan, Li Yi, Zinan Zhu, Xiaohua Hu, and Ming Zhang. 2016. Dialogue session segmentation by embedding-enhanced texttiling. In Proceedings of Interspeech 2016, pages 2706–2710, San Francisco, California, USA.

Dominique Stutzmann, Jacob Currie, Béatrice Daille, Amir Hazem, and Christopher Kermorvant. 2019. Integrated DH. rationale of the HORAE research project. In Proceedings of Digital Humanities (DH 2019), Utrecht, The Netherlands.

Dominique Stutzmann. 2019. Résistance au changement ? Les écritures des livres d'heures dans l'espace français (1200-1600). In Proceedings of the 19th Colloquium of the Comité international de paléographie latine. Change in medieval and Renaissance scripts and manuscripts, pages 97–116, Berlin, Germany.

Bingjun Sun, Prasenjit Mitra, C. Lee Giles, John Yen, and Hongyuan Zha. 2007. Topic segmentation with shared topic detection and alignment of multiple documents. In Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2007), pages 199–206, Amsterdam, The Netherlands.

Ryuichi Takanobu, Minlie Huang, Zhongzhou Zhao, Fenglin Li, Haiqing Chen, Xiaoyan Zhu, and Liqiang Nie. 2018. A weakly supervised method for topic segmentation and labeling in goal-oriented dialogues via reinforcement learning. In Proceedings of the 27th International Joint Conference on Artificial Intelligence (IJCAI 2018), pages 4403–4410, Stockholm, Sweden.

Masao Utiyama and Hitoshi Isahara. 2001. A statistical model for domain-independent text segmentation. In Proceedings of the 39th Annual Meeting on Association for Computational Linguistics (ACL 2001), pages 499–506, Toulouse, France.

Liang Wang, Sujian Li, Yajuan Lv, and Houfeng Wang. 2017. Learning to rank semantic coherence for topic segmentation. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP 2017), pages 1340–1344, Copenhagen, Denmark.

Roger S. Wieck, Lawrence R. Poos, Virginia Reinburg, John H. Plummer, and Walters art museum. 1988. Time sanctified: the Book of Hours in medieval art and life. G. Braziller, New York.

Yaakov Yaari. 1997. Segmentation of expository texts by hierarchical agglomerative clustering. CoRR, cmp-lg/9709015.