# Low Resource Methods for Medieval Document Sections Analysis

**Petra Galuščáková[ab] and Lucie Neužilová[c]**
[a]UMIACS, University of Maryland
College Park, MD, USA
petra@umiacs.umd.edu
[c]Charles University, Faculty of Arts, Department of Auxiliary Sciences of History
Prague, Czech Republic

**Abstract**

This paper describes a small but unique digitized collection of medieval Latin charters. This collection consists of 57 charters of 7 types illustrating various purposes of issuance by the Royal Chancellery. Sections in these documents were manually annotated for deeper analysis of the structure of issued charters. This paper also describes two baseline methods for an automatic and semi-automatic analysis and detection of sections of diplomatic documents. The first method is based on an information retrieval paradigm, and the second one is an adaptation of Hidden Markov Models. Both methods were proposed to work with respect to a small amount of available train data. Even though these methods were specifically proposed to work with medieval Latin charters, they can be applied to any documents with partially repetitive character.

**Keywords:** medieval documents, document sections detection, information retrieval, Latin

## 1. Introduction

Medieval charters issued by the Royal Chancellery were composed with respect to their typical structure (Guyotjeannin et al., 1993). They often consisted of repetitive sections and phrases, what is frequent in diplomatic language. Type and section ordering depends on the document type, which is defined by its purpose, type of the issuer and period of issuance. The purpose of this work is to 1) create a digital collection of medieval diplomatic charters, 2) provide manual annotations of a common structure of these charters, and 3) propose general methods for automatic detection of sections of these charters. The proposed methods should be trained and tested on this specific collection. As the number of diplomatic charters created using a given structure is often very limited and the size of the available train set is thus restricted, we aim at the methods which use a small amount of manually annotated documents for the training. We expect that providing an annotated collection of medieval charters and tools for automatic or semi-automatic recognition of document sections and their classification will significantly facilitate reading and processing of charters. It will allow further deeper analysis of individual diplomatic sections of charters and provide tools which will enable history researchers to think about relations and classification of material instead of spending time with slow and monotonous work and thus speeds up the whole research process.

### 1.1. Medieval Manuscript Collections

Lately, there is a growing effort towards digitization of medieval manuscripts. A number of such collections are provided as online digital libraries, e.g. Manuscriptorium[1], Digital Scriptorium[2], Manuscripta.at[3], Syriaca[4], E-codices[5], Saramusik[6] and Monasterium[7]. Manuscriptorium is a project of Czech National library which provides access to manuscripts, incunabula, early printed books, maps and charters. Digital Scriptorium is a consortium of American libraries and museums which provides free online access to pre-modern manuscripts. Some of the archives are aimed at specific types of manuscripts, e.g. E-codices provides access to Swiss manuscripts, Syriaca provides and access to manuscripts in Syriac language and Saramusik provides access to Arabic music manuscripts. Closest to our interest is Monasterium which collects archives of medieval and early modern charters. Monasterium provides access to almost 200 collections and more than half a million of documents. Growing effort in the area of processing charters is also reflected by the creation of Charters Encoding Initiative (CEI) (Vogeler, 2010). This coding scheme evolved from a TEI encoding scheme (Burnard and Rahtz, 2013) which is standard for the representation of texts in digital form and it is widely used for encoding historical documents. CEI extension was proposed for charter encoding and it is broadly used in the Monasterium collections.

### 1.2. Document Sections Analysis

This paper presents technical, linguistic and diplomatic problems of automatic analysis of the structure of medieval charters. According to our best knowledge, similar research had not been conducted before. However, our research is connected to research on automatic detection of general document structure, for example, automatic analysis of the discourse structure of medical abstracts (Lin et al., 2006).

[b] Previously Charles University, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics

[1] http://www.manuscriptorium.com

[2] http://www.digital-scriptorium.org

[3] http://manuscripta.at

[4] http://www.syriaca.org/

[5] http://www.e-codices.unifr.ch/en

[6] http://www.saramusik.org

[7] http://monasterium.net

Hidden Markov Models and Support Vector Machines were used to detect the introduction, methods, results and conclusions sections. In contrast to our case, train data for this domain can be relatively easy to acquire and supervised methods can be well applied. Document structure can also be detected using various semantic-based segmentation methods and lexical chain-based algorithms.

The two most often used algorithms for semantic segmentation are *TextTiling* (Hearst, 1997) and *C99* (Choi, 2000). Both measure segment similarity by calculating the cosine distance between neighbouring segments. *C99* calculates the similarity between all sentence pairs using the cosine measure and identifies regions with high intra-similarities. *TextTiling* calculates the similarity for adjacent segments of predefined size and points with the lowest values are designated as boundaries. Lexical chain-based algorithms detect boundaries based on the fact that the number of lexically related words within one segment is typically higher than the number of related words between adjacent paragraphs. Repetition of the lexical items can be detected easily and this approach may be improved by using synonyms and subordinates. Morris and Hirst (1988) determine lexically close words from thesaurus, Nguyen et al. (2011) further utilize word collocations, Mohri et al. (2010) calculate co-occurrence statistics and Kozima (1993) estimates similarities for pairs of words and uses them to find a sequence of lexical cohesiveness. Ponte and Croft (1997) propose a method for detection of small segments which share few common words using Local Content Analysis. In contrast to these approaches, our approach is supervised, and in addition to segmentation of the text, we are also interested in detection of the section type.

## 2. Medieval Charters Collection

Documents created in the era of John the Blind, King of Bohemia (1310–1346) and Count of Luxembourg are used in this work. We work only with charters written in Latin due to its much more consistent orthography than vernacular languages. Medieval charters also have a high level of uniformity. The presence and absence of charter sections and the level of their expression are frequently in correspondence with the type of the charter and with its historical importance.

Fealty deeds of John the Blind (von Estgen et al., 2009) are mainly used in this work. Fealty deeds are representative, typically contain all common sections and were not influenced by the literary creativity of the author. They are also very common in the High Middle Age. The advantage of this particular charter edition is also available transcription rules and uniformed orthography. The charters from printed edition were scanned, automatically recognized by a scanner built-in OCR system and manually corrected. Manual corrections were especially necessary as the applied OCR system did not support Latin, though the quality of the scanned documents was reasonable. Finally, sections in these charters were manually marked using CEI encoding scheme. All annotations were done by a single annotator who was a doctoral student of medieval history with a knowledge of Latin.

In most of the cases, we followed the transcription rules of the mentioned edition, except some simple rules, as interchanging *u* and *v*, which used to form the same grapheme in medieval Latin. These rules helped us to unify different word variants and achieve higher data consistency. The list of applied rules is given in Table 1.

| Original | Replaced |
|---|---|
| ae | e |
| oe | e |
| y | ii |
| j | i |
| ci + vowel | ti + vowel |
| uu + vowel | w + vowel |
| vowel + u | vowel + v |
| u + vowel | v + vowel |

Table 1: A list of applied transcription rules.

The collection consists of 57 manually annotated documents in total. These documents were divided into train, held-out and test set. Apart from fealty deeds (*lenni slib*), the collection also contains a limited number of acquaintances (*kvitance*), debt reliefs (*zprosteni dluhu*), donation (*donace*), reformation (*polepseni*) and request (*zadost*)) charters. Some documents were not able to be well categorized. Collection statistics are tabulated in Table 2. All these types of documents are on the same level in the hierarchy of importance of documents issued by the Royal Chancellery. This diversity of the document types allows us to test our methods on a different type of charters.

The annotated collection was published together with section detection framework[8] and is available under Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

### 2.1. Charter Sections

Sections and phrases of diplomatic charters often have repetitive character, and similar phrases and sections occur in similar document types. However, this kind of stability does not involve always using the same words and word phrases. Meaning and the role of the charter is often unified, but it was on the author how to express this meaning. Especially in the Royal Chancellery, the forms were relatively stabilized. For different authors, chancellery locality and era, it is characteristic to use different forms and phrases. Thanks to automatic recognition of these sections, we can study more about the process of creating of charters and about the chancellery personnel.

Some of these sections and phrases were composed very practically using only a few necessary words. For example, *intitulatio*, where the issuer introduces himself, *publicatio*, where the issuer expresses his intention to issue the charter, or *corroboratio*, where the issuer announces means of the sealing. Other sections and phrases leave more space for creativity and literary art of the author. *Dispositio*, the main

[8] http://ufal.mff.cuni.cz/Medieval-Charter-Sections-Corpus

| Collection Set | Fealty deeds | Acquaintance | Debt relief | Donation | Reformation | Request | None | Total |
|---|---|---|---|---|---|---|---|---|
| Train | 15 | 1 | 0 | 0 | 1 | 0 | 6 | 23 |
| Held-out | 3 | 0 | 0 | 0 | 0 | 1 | 0 | 4 |
| Test | 18 | 6 | 3 | 3 | 0 | 0 | 0 | 30 |

Table 2: Document type statistics

section of the charter, expresses the primary meaning of the charter as a legal document which codifies a legal act of the issuer. *Narratio*, which express issuer's intention to do something, is frequently formulated even more freely and is thus more complicated to detect automatically.

These differences between sections are fundamental for our work. Since *publicatio*, *intitulatio* and *corroboratio* typically contain only a few words and have unified meaning, they can be easily recognized by information retrieval-based methods. However, *dispositio* and *narratio* are almost never composed in the same way and they are thus much harder to recognize. As the ordering of the sections is also often standardized, positional information in the text can be expected to be helpful for recognition of these sections.

Another problem is that not all the sections are contained in each charter. Presence and absence of the sections and phrases are in connection with a type of charter. Typically, all mentioned sections are present only in the most important charters. Documents in our collection, contain all the necessary phrases like *intitulatio*, *publicatio*, and *dispositio*. More freely formulated phrases such as *narratio* are present only in some of them.

Presence of these section types in our data is displayed in Table 3. We also show an average length of each section type in terms of words, which is often in correspondence with quality of automatic detection of the specific section.

## 2.2. Additional Metadata

Apart from manually marked sections, the documents in the train set also contain manually marked named entities. Documents contain 68 personal names (*persName*), 54 roles or positions in society, (*rolename*), 43 place names (*placename*), 66 single letter characters (*c*), 59 measure labels (*measure*), 23 date labels (*date*), and 4 expansion of an abbreviation (*expan*). All documents in all sets also contain short manually crafted abstract in Czech.

# 3. Charter Sections Detection

We experimented with two methods for automatic detection of charter sections. The first method is based on information retrieval and the second one uses Hidden Markov models. Both methods were adapted to be able to be trained using a small amount of data.

To further reduce language variance, we also experimented with lemmatized forms of words. Lemmatized word forms were created using the online version of the Lemlat lemmatizer (Ruffolo et al., 2017) for Latin. In addition to this, we also trained our models with and without manually marked named entities.

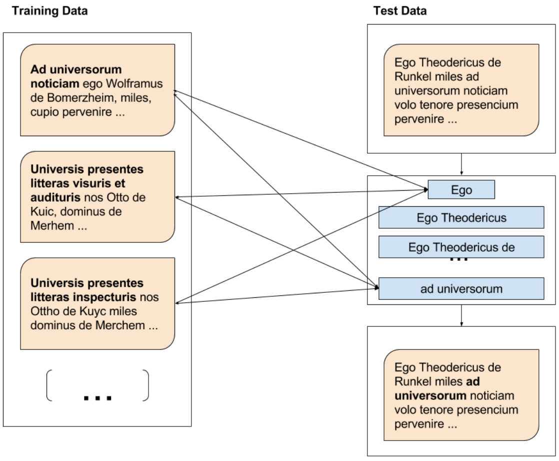
Training Data
Test Data
**Ad universorum noticiam** ego Wolframus de Bomerzheim, miles, cupio pervenire ...
**Universis presentes litteras visuris et audituris** nos Otto de Kuic, dominus de Merhem ...
**Universis presentes litteras inspecturis** nos Ottho de Kuyc miles dominus de Merchem ...
...
Ego Theodericus de Runkel miles ad universorum noticiam volo tenore presencium pervenire ...
Ego
Ego Theodericus
Ego Theodericus de
...
ad universorum
Ego Theodericus de Runkel miles **ad universorum** noticiam volo tenore presencium pervenire ...

Figure 1: An overview of the information retrieval-based method. The scheme presents searching for the *intitulatio* section. Three examples of the *intitulatio* section are displayed in the train data. These are then compared with all the possible sections in the test data document. The section in the test document, which is the most similar to one of the sections in the train data is then marked as the *intitulatio* section.

## 3.1. Section Detection Approaches

Thanks to the reduced vocabulary of some sections, **information retrieval**-based approach is expected to work well on them. For each pre-defined section, we find a sequence of words in the test document which is the most similar to any phrase manually marked as this pre-defined section in the train data. A similarity is calculated between each sequence of words in the test document and each marked phrase in all train documents. Then, a sequence of words with a maximal cosine similarity is selected to belong to the pre-defined section (the *Cosine* method). We also used the TF IDF score in addition to the cosine similarity (the *TF IDF* method). Employing the cosine similarity and TF IDF also enables us to detect the most similar section in the annotated document collection, which can be very helpful for the history researchers. To further improve the precision, we also calculate a sum of distances between the sequence of words from the test document and three most similar phrases manually marked to belong to the pre-defined section (the *Cosine Max 3* method). Information retrieval-based method is explained in Figure 1.

Information retrieval-based approach is supposed to work exceptionally well for short and well-defined sections, but it cannot be expected to work well on more freely formulated sections. Therefore, we also use **Hidden Markov Models** (HMM), which are supposed also to include infor-

| Section Phrase | Train | Held-out | Test | Avg. Length [words] |
|---|---|---|---|---|
| Corroboratio | 23 | 2 | 29 | 15.1 |
| Datatio | 22 | 3 | 30 | 7.8 |
| Dispositio | 20 | 3 | 30 | 94.7 |
| Inscriptio | 15 | 4 | 5 | 3.7 |
| Intitulatio | 22 | 4 | 30 | 4.1 |
| Narratio | 4 | 1 | 2 | 20.7 |
| Publicatio | 21 | 3 | 30 | 4.4 |

Table 3: Statistics of manually marked section phrases.

| Method | Named Entities | Lemmatized | Precision | Recall | F-score |
|---|---|---|---|---|---|
| Cosine | No | No | 0.92 | 0.19 | 0.31 |
| Cosine | No | Yes | 0.93 | 0.20 | 0.33 |
| Cosine | Yes | No | 0.86 | 0.27 | **0.41** |
| Cosine Max 3 | No | No | **0.94** | 0.17 | 0.29 |
| TF IDF | No | No | 0.88 | 0.23 | 0.36 |
| TF IDF | No | Yes | 0.83 | 0.22 | 0.35 |
| TF IDF | Yes | No | 0.83 | 0.25 | 0.38 |
| HMM | No | No | 0.67 | 0.27 | 0.38 |
| HMM | No | Yes | 0.65 | 0.20 | 0.31 |
| HMM | Yes | No | 0.52 | **0.30** | 0.38 |

Table 4: Comparison of information retrieval and HMM setting trained with and without marked named entities and with and without lemmatization. If named entities are set to "No", then all manually marked entities were removed from the train data. Best results are in **bold** type.

| Train Documents Type | Test Documents Type | Precision | Recall | F-score |
|---|---|---|---|---|
| All | All | 0.92 | 0.19 | 0.31 |
| All | Fealty deeds | 0.91 | **0.22** | **0.35** |
| All | Acquaintance | **0.96** | 0.14 | 0.24 |
| Fealty deeds | All | 0.91 | 0.20 | 0.32 |
| Fealty deeds | Fealty deeds | 0.91 | 0.21 | 0.34 |
| Fealty deeds | Acquaintance | **0.96** | 0.15 | 0.25 |

Table 5: Comparison of the cosine similarity method trained and tested on different document types. Best results are in **bold** type.

mation about sections ordering into the decision. Hidden states in the proposed HMM are formed by individual section types. These hidden states then generate charter text. As the amount of train data is small and the output probabilities for individual words can be skewed, cosine distance is used instead of the output probability. This distance is calculated between continuous word sequence which ends in the current state and the closest phrase manually assigned to this state (i.e. section type) in the train data. More precisely, all words in the word sequence need to be assigned to the same state as the current one. Finally, each word has an assigned section type. As the sections created in this way can be non-continuous and several word sequences can be marked to belong to a single section type, we assign only the continuous word sequence with the highest cosine similarity score to each section type.

## 4. Results and Discussion

Results of tested approaches are tabulated in Table 4 and Table 5. All methods are evaluated using precision, recall and F-score.

The highest precision scores are achieved using information retrieval-based method with cosine measure. The overall highest F-score of 0.41 is obtained for cosine measure when it is trained with all available named entities and no lemmatization is employed. The overall highest precision is acquired when this measure is calculated as a sum of cosine distances over three most similar phrases. We confirm that information retrieval-based methods perform especially well on more standardized section types. For example, the precision of retrieval *intitulatio* is 1, but the returned sequences are very short and they often contain only a few words.

As predicted, HMM-based method achieve better results for less formally standardized sections. It produces longer phrases and thus also achieves higher recall. The recall is highest when all named entities are used and no lemmatization is used. Though, the HMM-based method does not perform well when a particular section is missing from the test document.

To test retrieval of specific charter type, we compare retrieval trained on full train set and the *Fealty deeds* documents only. We test retrieval on full dataset, *Fealty deeds* documents and *Acquaintance* documents as we only have maximally three documents from other charter types in the test set. These results show that the differences between different document types are small and thus confirm the uniformity of our collection.

## 5. Conclusion and Future Work

In the paper, we presented the digitized and annotated collection of medieval Latin charters. We provided a detailed inspection of sections of such documents with respect to different issuance properties. In addition to manual annotations of the sections available in our collection, we also provided manually crafted named entities and Czech abstracts.

We also presented methods for detection of sections of documents and applied them to our collection. The highest precision was achieved when we used word sequences with the highest sum of cosine similarity scores calculated over three most similar phrases assigned to particular section type in the train data. The highest recall was achieved for the HMM-based algorithm.

We confirm that used methods can substantially reduce the effort needed by historians to process medieval charters, what enables them to work faster and thus to analyze larger amounts of data. As the methods were initially been not proposed to replace manual work completely but to allow researchers to work in a semi-automatic way, achieved precision is especially encouraging, as it enables reliable detection of probable word sequences assigned to sections. Researchers can then easily manually correct these results, if needed, by including additional words.

However, we would still further like to improve relatively low recall and thus enlarge the length of detected sections. We believe that better post-processing of the HMM algorithm can improve this. Moreover, we also plan to combine both methods to be able to make an advantage of the high precision achieved by information retrieval- based methods on short sections and recall achieved by the HMM-based methods on more freely formulated sections.

## 6. Acknowledgements

This research is supported by the Czech Science Foundation, grant number P103/12/G084 and NSF grant number 1618695.

## 7. Bibliographical References

Burnard, L. and Rahtz, S. (2013). A Complete Schema Definition Language for the Text Encoding Initiative. Presented at XML London 2013.

Choi, F. Y. Y. (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, WA, USA.

Guyotjeannin, O., Pycke, J., and Tock, B. (1993). *Diplomatique Médiévale*. L'atelier du médiéviste. Brepols.

Hearst, M. A. (1997). TextTiling: Segmenting Text into Multi-paragraph Subtopic Passages. *Computational Linguistics*, 23(1):33–64.

Kozima, H. (1993). Text Segmentation Based On Similarity Between Words. In *Proceedings of the 31st annual meeting on Association for Computational Linguistics ACL '93*, pages 286–288, Columbus, Ohio.

Lin, J., Karakos, D., Demner-Fushman, D., and Khudanpur, S. (2006). Generative Content Models for Structural Analysis of Medical Abstracts. In *Proceedings of the Workshop on Linking Natural Language Processing and Biology: Towards Deeper Biological Literature Analysis*, pages 65–72, New York City, New York.

Mohri, M., Moreno, P., and Weinstein, E. (2010). Discriminative Topic Segmentation of Text and Speech. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pages 533–540, Chia Laguna Resort, Sardinia, Italy.

Morris, J. and Hirst, G. (1988). Lexical Cohesion, the Thesaurus, and the Structure of Text. Technical report, Computer Systems Research Institute, University of Toronto, Toronto, Canada.

Nguyen, V. C., Nguyen, L. M., and Shimazu, A. (2011). Improving Text Segmentation with Non-systematic Semantic Relation. In *Computational Linguistics and Intelligent Text Processing: 12th International Conference, CICLing 2011, Part I*, pages 304–315, Tokyo, Japan.

Ponte, J. M. and Croft, W. B. (1997). Text Segmentation by Topic. In *Proceedings of the First European Conference on Research and Advanced Technology for Digital Libraries ECDL '97*, volume 1324 of *LNCS*, pages 113–125, Pisa, Italy.

Ruffolo, P., Passarotti, M., Budassi, M., and Litta, E. (2017). The Lemlat 3.0 Package for Morphological Analysis of Latin. In *Proceedings of the NoDaLiDa 2017 Workshop on Processing Historical Language*, number 133, pages 24–31.

Vogeler, G. (2010). Charters Encoding Initiative Overview. In *Digital Proceedings of the Lawrence J. Schoenberg Symposium on Manuscript Studies in the Digital Age*, volume 2.

von Estgen, A., Pauly, M., Pettiau, H., and Schroeder, J. (2009). *Urkunden- und Quellenbuch zur Geschichte der altluxemburgischen Territorien IX. Die Urkunden Graf Johanns des Blinden (1310–1346). Teil 2. Die Urkunden aus den Archives Générales du Royaume Brüssel*. Luxemburg.