# Automatic Hadith Segmentation using PPM Compression

**Taghreed Tarmom**
School of Computing
University of Leeds
Leeds, UK
sctat@leeds.ac.uk

**Eric Atwell**
School of Computing
University of Leeds
Leeds, UK
e.s.atwell@leeds.ac.uk

**Mohammad Alsalka**
School of Computing
University of Leeds
Leeds, UK
M.A.Alsalka@leeds.ac.uk

## Abstract

In this paper we explore the use of Prediction by partial matching (PPM) compression based to segment Hadith into its two main components (Isnad and Matan). The experiments utilized the PPMD variant of the PPM, showing that PPMD is effective in Hadith segmentation. It was also tested on Hadith corpora of different structures. In the first experiment we used the non-authentic Hadith (NAH) corpus for training models and testing, and in the second experiment we used the NAH corpus for training models and the Leeds University and King Saud University (LK) Hadith corpus for testing PPMD segmenter. PPMD of order 7 achieved an accuracy of 92.76% and 90.10% in the first and second experiments, respectively.

## 1 Introduction

Automated text segmentation is the task of building a tool that can automatically identify sentence boundaries in a given text and divide them into their components. The need to convert unstructured text into a structured format is especially important when dealing with unstructured text such as web text or old documents.

One of the most important types of old holy Islamic texts in the Arabic language is Hadith. Hadith—the second source of Islam—refers to any action, saying, order, or silent approval of the holy prophet Muhammad that was delivered through a chain of narrators. Each Hadith has an Isnad—the chain of narrators—and a Matan—the act of the prophet Muhammad. Figure 1 shows an example of Hadith.

While most ordinances of Islam are mentioned in the Quran in general terms, detailed and vivid explanations are often provided in the Hadith. This gives the Hadith importance among Muslims. For example, prayer, 'الصلاة', is mentioned in the Quran, while the Hadith specifies what Muslims should do and say; the Hadith explains the time for each prayer and what Muslims should do before and after the prayer. In contrast to the Quran, some Hadiths, which have been handed down over centuries, have been corrupted by incompetent narrators who transferred them incorrectly. Hadith scholars have classified these as non-authentic Hadiths.

Al-Humaydee `Abdullaah ibn Az-Zubayr narrated to us saying: Sufyaan narrated to us, who said: Yahyaa ibn Sa`eed Al-Ansaree narrated to us: Muhammad Ibn Ibraaheem At-Taymee informed me: That he heard `Alqamah Ibn Waqaas Al-Laythee saying: I heard `Umar ibn Al-Khattaab whilst he was upon the pulpit saying: I heard Allaah's Messenger (salallaahu `alaihi wassallam) saying: *"Indeed actions are upon their intentions"*

حَدَّثَنَا الْحُمَيْدِيُّ عَبْدُ اللَّهِ بْنُ الزُّبَيْرِ قَالَ حَدَّثَنَا سُفْيَانُ قَالَ حَدَّثَنَا يَحْيَى بْنُ سَعِيدٍ الأَنْصَارِيُّ قَالَ أَخْبَرَنِي مُحَمَّدُ بْنُ إِبْرَاهِيمَ التَّيْمِيُّ أَنَّهُ سَمِعَ عَلْقَمَةَ بْنَ وَقَّاصٍ اللَّيْثِيَّ يَقُولُ سَمِعْتُ عُمَرَ بْنَ الْخَطَّابِ رَضِيَ اللَّهُ عَنْهُ عَلَى الْمِنْبَرِ، قَالَ سَمِعْتُ رَسُولَ اللَّهِ صَلَّى اللَّهُ عَلَيْهِ وَسَلَّمَ يَقُولُ
"إِنَّمَا الأَعْمَالُ بِالنِّيَّاتِ"

Figure 1: An example of Hadith, Isnad in black and Matan in green.

Automatic Hadith segmentation of Isnad and Matan can help Hadith researchers, some of whom focus on an Isnad with the aim of studying narrators' reliability, the links between them, or how a specific Hadith has been transferred through the ages, sometimes generating a graphical visualization to represent this (Azmi and Badia, 2010). Other research concentrates on Matan to classify Hadiths into topics (Saloot et al., 2016).

Teahan (2000) used prediction by partial matching (PPM) to solve several NLP problems, such as text classification and segmentation. Altamimi and Teahan (2017) and Tarmom et al. (2020b) pointed out that us-



*Proceedings of the 17th International Conference on Natural Language Processing*, pages 22–29
Patna, India, December 18 - 21, 2020. 

ing a character-based compression scheme for tasks such as detecting code-switching and gender/authorship categorization is more effective than word-based machine learning approaches. Many current Hadith studies use a word-based method to segment Hadith from the six canonical Hadith books, but the method of this paper uses a character-based PPM compression method to automatically segment the Isnad and Matan. Our goals are to evaluate PPM segmenter on (1) unstructured Hadith text from lesser-known Hadith books and (2) well-structured Hadith text from the six canonical Hadith books.

This paper explains the data sets chosen for our experiments and outlines the experiments performed on Arabic Hadith text to evaluate the PPM compression method. Finally, we draw conclusions and suggest future work based on this study.

## 2 Related Work

There have been relatively few studies on the segmentation of Hadith into Isnad and Matan. One study was carried out by Harrag (2014), who developed a finite state transducers-based system to detect the different parts of a Hadith, such as *Title-Bab*, *Num Hadith*, *Sanad* '*Isnad*', and *Matn* '*Matan*'. The disadvantage of this system is that it was built to depend on the Hadith structure in *Sahih Al-Bukhari* book (the most trusted Hadith book), which cannot be used for other Hadith books. Figure 2 shows the Hadith structure in the *Sahih Al-Bukhari* book. This system achieved a precision of 0.44 for Isnad extraction and 0.61 for Matan extraction.

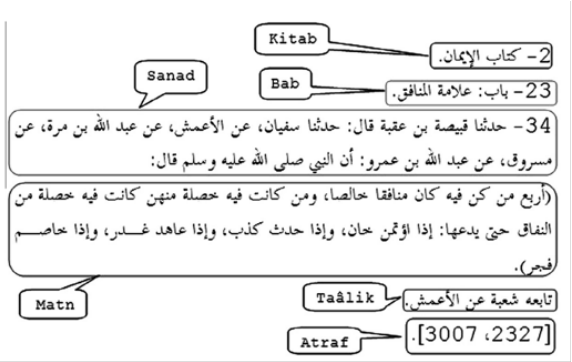


Figure 2: An example of a Hadith structure in the *Sahih Al-Bukhari* book (Harrag, 2014).

| Book Name | Precision | Recall | F1 Measure |
|---|---|---|---|
| Sahih Muslim English | 96% | 91% | 93% |
| Sahih Bukhari English | 99% | 99% | 99% |
| Sunan Abudawud | 100% | 100% | 100% |
| Mawta Imam Malik | 100% | 100% | 100% |

Table 1: Results of different Hadith books (Mahmood et al., 2018).

Mahmood et al. (2018) selected authentic and reliable Hadith sources such as *Sahih Al-Bukhari*, *Sahih Muslim English*, and *Sunan Abu Dawud*. Since these books differ in format, structure, length, and content, the researchers used different kinds of regular expressions (Regex) for data extraction. However, Hadith patterns extracted by their system lack detail. The results obtained by their system are summarized in Table 1.

Maraoui et al. (2019) implemented a segmentation tool to automatically segment Isnad and Matan from each text from the *Sahih Al-Bukhari* book. First, they analysed the *Sahih Al-Bukhari* corpus and identified the words that distinguish Isnad from Matan. These words were then added to the trigger word dictionary. This tool achieved a precision of 96%.

Altammami et al. (2019) built a Hadith segmenter using N-grams. The *Sahih Al-Bukhari* book was selected as a training set, and the testing set was manually extracted from the six canonical Hadith books. Their result showed that using bi-grams achieved a much higher accuracy (92.5%) than tri-grams (48%).

Most Hadith segmentation research works have used the six famous Hadith books, called The Authentic Six 'الصحاح الستة'. Hence, there is a shortage of research on lesser-known Hadith books, such as Fake Pearls of the Non-Authentic Hadiths 'اللآلئ المصنوعة في الأحاديث الموضوعة'. These books contain a mixture of authentic and non-authentic Hadiths and do not have a clear structure, which makes the segmentation task more complex. Also, character-based text compression methods have not been used in previous Hadith segmentation studies. Our work seeks to fill these gaps in research.

## 3 Data Collection

For this study, we selected a non-authentic Hadith (NAH) corpus built by Tarmom et al. (2020a) as a training and testing set. The main feature of this corpus is that it contains

452,624 words from different lesser-known Hadith books. It also included several annotated Hadith books, which help to determine the switch points between the Isnad and the Matan, and thus provide a ground truth. Table 2 shows the NAH corpus contents.

These books were downloaded from Hadith websites such as `islamweb.net` and `almeshkat.net` as Word files and converted to csv files. Some of these books have both Hadiths (authentic and NAH), while others only contain NAH. The annotating process was done to determine eight primary features for each Hadith in this corpus. These are *No.*, *Full Hadith*, the *Isnad*, the *Matan*, the *Authors Comments*, the *Hadith Type*, *Authenticity* and *Topic*. A description of the NAH corpus features is shown in Table 3.

## 4 PPM Compression-based Segmenter

The PPM text compression algorithm is a character-based model that predicts an upcoming symbol by using the previous symbols with a fixed context. Every possible upcoming symbol is assigned a probability based on the frequency of previous occurrences. If a symbol has not been seen before in a particular context, the method will 'escape' to another lower-order context to predict the symbol. This is called the escape method and is used to combine the predictions of all character contexts (Cleary and Witten, 1984). Different variants of PPM have been produced in order to give better compression results, such as PPMC (Moffat, 1990) and PPMD (Howard, 1993). Howard (1993), who invented PPMD, showed that PPMD gives better results for text compression than PPMC.

Equation 1 defines how PPMD estimates the probability P for the next symbol $\phi$:

$$P(\phi) = \frac{2C_d(\phi) - 1}{2T_d} \quad (1)$$

where $d$ is the coding order, $T_d$ indicates how many times that the current context, in total, has existed, and $C_d(\phi)$ is the total number of instances for the symbol $\phi$in the current context. Equation 2 defines how PPMD estimates the escape probability $e$:

$$e = \frac{t_d}{2T_d} \quad (2)$$

where $t_d$ represents how many times that a unique character has existed following the current context.

Table 4 describes how PPM handles the string 'PXYZXY' with order $k$=2. For illustration purposes, two has been chosen as the model's maximum order. In order two, if the symbol 'Z' follows the context 'PXYZXY', its probability will be $\frac{1}{2}$ because it has been found before ($XY \rightarrow Z$). The encoding of the symbol 'Z' requires -log($\frac{1}{2}$)=1 bit.

If the symbol 'T' follows the context 'PXYZXY', an escape probability of $\frac{1}{2}$ will be arithmetically encoded because it has not been found after 'XY' in order two. Then the PPM algorithm will move to the lower order, which is order one. In order one, because the symbol 'T' has not been found after the symbol 'Y', an escape probability of $\frac{1}{2}$ will be also encoded. Then, it will be repeated in order zero and an escape probability of $\frac{4}{10}$ will be be encoded because the symbol 'T' has not been found in order zero. Finally, the algorithm will move to order –1. In this order, all symbols are found and the probability will be $\frac{1}{|A|}$, where A = 256 (the alphabet size for ASCII), so its probability will be $\frac{1}{256}$ . The encoding of the symbol 'T' requires -log($\frac{1}{2}\times\frac{1}{2}\times\frac{4}{10}\times\frac{1}{256}$)=11.32 bits.

Tawa is a compression-based toolkit that adopts the PPM algorithm. It consists of nine main applications, such as `classify`, `codelength`, `train`, `markup`, and `segment` (Teahan, 2018). This study concentrates on two applications provided by the Tawa toolkit: building models and text segmentation.

### 4.1 Viterbi Algorithm

For text segmentation using Tawa, we used the toolkit's `train` tool to train multiple models on representative text under research. We then used the markup tool, which utilizes the Viterbi algorithm (Viterbi, 1967). This uses a trellis-based search (Ryan and Nudd, 1993) to find the segmentation with the best compression with all possible segmentation search paths extended at the same time, discarding the poorly performing alternatives (Teahan, 2018).

Figure 3 shows an illustrative example of the search tree for the text segmentation problem in the Tawa toolkit. In this example, the

| No. | Book Reference Name | Book's Title | Author | Book's Contents | Hadith's Type | No. of words |
|---|---|---|---|---|---|---|
| 1 | N1 | الأباطيل والمناكير والصحاح والمشاهير | أبو عبد الله الهمذاني الجورقاني | Isnad/Matan/Comments | Authentic and NAH | 121,080 |
| 2 | N2 | مائة حديث ضعيف وموضوع منتشرة بين الخطباء والوعاظ | إحسان العتيبي | Matan/Comments | NAH | 2,898 |
| 3 | N3_1 | اللآلئ المصنوعة في الأحاديث الموضوعة الجزء الثاني ط دار المعرفة | جلال الدين السيوطي | Isnad/Matan/Comments | Authentic and NAH | 15,421 |
| 4 | N3_2 | اللآلئ المصنوعة في الأحاديث الموضوعة الجزء الثاني ط دار المعرفة | جلال الدين السيوطي | Isnad/Matan/Comments | Authentic and NAH | 151,382 |
| 5 | N4 | الأحاديث الضعيفة في كتاب رياض الصالحين | إحسان العتيبي | Isnad/Matan/Comments | NAH | 5,675 |
| 6 | N5 | الجد الحثيث في بيان ما ليس بحديث ت بو زيد دار الراية | أحمد بن عبد الكريم العامري | Matan/Comments | NAH | 16,382 |
| 7 | N6 | الفوائد المجموعة في الأحاديث الموضوعة ط العلمية | الإمام محمد بن علي الشوكاني | Matan/Comments | NAH | 139,786 |

Table 2: The NAH corpus contents.

| Features | Description |
|---|---|
| No. | The Hadith reference number. |
| Full Hadith | The Hadith as it appears in the book without annotations |
| Isnad | The chain of narrators |
| Matan | The act of the Prophet Muhammad |
| Authors Comments | The author describes the authenticity of each Hadith |
| Hadith Type | The Hadith Type (Maqtu' مقطوع, Mawquf موقوف and Marfo مرفوع) or Hadith degree (موضوع،ضعيف and so on) |
| Authenticity | Whether this Hadith is authentic or non-authentic |
| Topic | The chapter title |

Table 3: Features of the NAH corpus.

| *Order k=2* | | | *Order k=1* | | | *Order k=0* | | | Order k=-1 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Prediction | $c$ | $p$ | Prediction | $c$ | $p$ | Prediction | $c$ | $p$ | Prediction | $c$ | $p$ |
| $PX \rightarrow Y$ | 1 | 1/2 | $P \rightarrow X$ | 1 | 1/2 | $P$ | 1 | 1/10 | $A$ | 1 | 1/\|A\| |
| $\rightarrow Esc$ | 1 | 1/2 | $\rightarrow Esc$ | 1 | 1/2 | $X$ | 2 | 2/10 | | | |
| $XY \rightarrow Z$ | 1 | 1/2 | $X \rightarrow Y$ | 2 | 2/3 | $Y$ | 2 | 2/10 | | | |
| $\rightarrow Esc$ | 1 | 1/2 | $\rightarrow Esc$ | 1 | 1/3 | $Z$ | 1 | 1/10 | | | |
| $YZ \rightarrow X$ | 1 | 1/2 | $Y \rightarrow Z$ | 1 | 1/2 | $\rightarrow Esc$ | 4 | 4/10 | | | |
| $\rightarrow Esc$ | 1 | 1/2 | $\rightarrow Esc$ | 1 | 1/2 | | | | | | |
| $ZX \rightarrow Y$ | 1 | 1/2 | $Z \rightarrow X$ | 1 | 1/2 | | | | | | |
| $\rightarrow Esc$ | 1 | 1/2 | $\rightarrow Esc$ | 1 | 1/2 | | | | | | |

Table 4: Handling the string 'PXYZXY' using PPM with order 2.

tree has a branching of two, since two labels have been used: Isnad and Matan. The label <I> (used for the Isnad model) and <M> (used for the Matan model) show the transformed sequences within each node. If a character switches from one model to the other, the sentinel character is encoded. The compression codelength is also calculated for the transformed sequence, which it appears on the right of each node and below the last nodes. The smallest one, which is the best segmented, is shown in bold font.

## 5 Evaluation Experiments

Two experiments were performed as part of the evaluation of the compression-based method (provided by Tawa) (Teahan, 2018) to automatically separate Hadith into two components, Isnad and Matan. In the first experiment we used the NAH corpus for training models and testing, and in the second experiment we used the NAH corpus for training models and the Leeds University and King Saud University (LK) Hadith corpus, built by Altammami et al. (2020), for testing PPM segmenter.

### 5.1 First Experiment

In this experiment, the first book in the NAH corpus, *N1*, was chosen for training purposes. This book is called the *False, Disreputable, and Well-known Hadith Texts* 'الأباطيل والمشاهير والمناكير للجورقاني'. It consists of 732 Hadiths and 121,080 words. Isnads and Matans were manually extracted from *N1* for Isnad and Matan training models, which were 52,221 and 33,489 words long, respectively. The testing text was manually extracted from the third book in NAH corpus, *N3_1*, which contained just Isnad and Matan and is 6,339 words long.

For automatic Hadith segmentation, different orders of PPMD were performed, from order 3 to order 10. As shown in Table 5, Order 7 obtained a higher accuracy of 92.76%, a higher average recall of 0.9365, a higher average precision of 0.9231, and a higher average F-measure of 0.9288. A sample output from the first experiment is shown in Figure 4. Figure 5 shows the last part of Isnad texts were predicted as Matan such as 'عن جابر عن النبي صلى الله عليه وسلم It has been narrated on the authority of Jabir on the authority of the Prophet, may God bless him and grant him peace'(highlighted in blue).

We noticed that the structure of the Isnad texts used in the training set and the testing set differed, creating some confusion in the result. The type of Hadith is given at the beginning of each Hadith in *N1*, for example حديث مرفوع 'Marfo Hadith', which was not labelled as

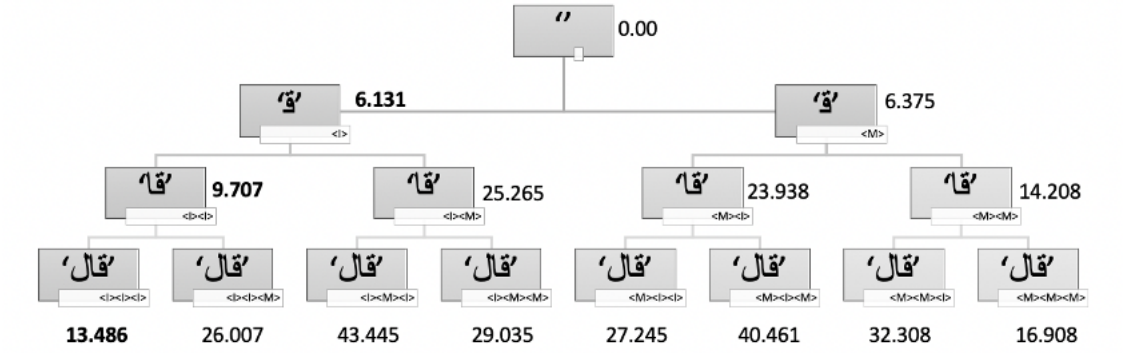


Figure 3: An illustrative example of the search tree for the text segmentation problem in the Tawa toolkit.

| Orders | Accuracy (%) | Recall | Precision | F-measure |
|---|---|---|---|---|
| 2 | 83.34 | 0.8580 | 0.8555 | 0.8568 |
| 3 | 87.20 | 0.8914 | 0.8801 | 0.8858 |
| 4 | 88.04 | 0.8996 | 0.8843 | 0.8919 |
| 5 | 87.02 | 0.8881 | 0.8800 | 0.8840 |
| 6 | 88.58 | 0.9022 | 0.8901 | 0.8961 |
| **7** | **92.76** | **0.9365** | **0.9231** | **0.9288** |
| 8 | 92.68 | 0.9356 | 0.9222 | 0.9345 |
| 9 | 92.67 | 0.9350 | 0.9215 | 0.9282 |
| 10 | 91.78 | 0.9275 | 0.9127 | 0.9200 |

Table 5: Hadith segmentation using PPMD.

<Isnad>حدثنا المقدام بن داود حدثنا أسد بن
موسى حدثنا يوسف بن زياد عن عبدالمنعم بن
إدريس عن أبيه إدريس عن جده وهب بنم منبه عن
أبي هريرة<\Isnad><Matan> أن رجلاً من اليهود
أتى النبي صلى الله عليه وسلم فقال يا رسول الله هل
احتجب الله من خلقه بشيء غير السموات قال نعم
بينه وبين الملائكة الذين حول العرش سبعون
حجاباً من نور وسبعون حجاباً من نار وسبعون
حجاباً من ظلمة وسبعون حجاباً من رفارف الإستبرق
وسبعون حجاباً من رفارف السندس وسبعون حجاباً
من در أبيض وسبعون حجاباً من در أحمر وسبعون
حجاباً من در أصفر وسبعون حجاباً من در أخضر
وسبعون حجاباً من ضياء استضاء من ضوئه النار
والنور وسبعون حجاباً من ثلج وسبعون حجاباً من
ماء وسبعون حجاباً من برد غمام وسبعون حجاباً
من برد وسبعون حجاباً من عظمة الله التي لا توصف
قال فأخبرني عن ملك الله الذي يليه فقال النبي
صلى الله عليه وسلم أصادقت فيما أخبرتك يا يهودي
قال نعم قال فإن الملك الذي يليه إسرافيل ثم
جبريل ثم ميكائيل ثم ملك الموت
<\Matan>

Figure 4: Sample output using PPMD with Order 7.

Isnad (see Figure 6). In the *N3_1* book, each type of Hadith has been written at the end of the Isnad (see Figure 7). Figure 8 shows that Isnad and Matan are correctly predicted but the word مرفوعا 'Marfo' was wrongly predicted

<Isnad>وقال الخطيب أخبرني القيني أنبأنا محمد
بن العباس أنبأنا أبو أيوب سليمان بن إسحق
الحلاب قال سئل إبراهيم الحربي عن حديث موسى بن
إبراهيم عن ابن لهيعة عن أبي الزبير<\Isnad>
<Matan> عن جابر عن النبي صلى الله عليه وسلم من
قال القرآن مخلوق فقد كفر
<\Matan>
<Isnad>حدثنا محمد بن أحمد الوراق حدثنا سعيد
بن محمد ثواب بكر بن عيسى عن محمد بن عثمان
الحراني عن مالك بن دينار عن الحسن ع<\Isnad>
</\Isnad><Matan>ن أنس مرفوعاً أن لله لوحاً أحد
وجهيه درة والآخر ياقوتة قلمه النور فيه يخلق
وبه يرزق وبه يحيي وبه يميت ويعز ويذل ويفعل
ما يشاء في يوم وليلة
<\Matan>

Figure 5: An example of confusion between an Isnad and a Matan using PPMD with Order 7.

as belonging to a Matan since it did not appear in the Isnad training set (highlighted in blue).

**حديث مرفوع** فقال: فيما أخبر أبو الفضل محمد بن طاهر بن علي المقدسي، رضي الله عنه، قال: أخبرنا علي بن أحمد بن البندار، قال: حدثنا أبو طاهر محمد بن العباس المخلص، قال: حدثنا عبد الله بن محمد بن عبد العزيز البغوي، قال: حدثنا أبو خيثمة زهير بن حرب، قال: حدثنا إسماعيل بن إبراهيم، عن عبد العزيز بن صهيب، عن أنس، عن النبي صلى الله عليه وسلم، قال: "" من كذب علي متعمدا فليتبوأ مقعده من النار "". هذا حديث صحيح، أخرجه الإمام أبو الحسين مسلم بن الحجاج النيسابوري في صحيحه، عن أبي خيثمة زهير بن حرب هكذا

Figure 6: An example of Hadith from *N1* book (Hadith's type is in bold).

ابن عدي) حدثنا أحمد بن محمد بن حرب حدثنا ابن حميد عن جرير عن الأعمش عن أبي صالح عن أبي هريرة **مرفوعاً** القرآن كلام الله لا خالق ولا مخلوق من قال غير ذلك فهو كافر. موضوع: آفته ابن حرب وشيخه أيضاً كذاب وهو محمد بن حميد بن حبان.

Figure 7: An example of Hadith from *N3_1* book (Hadith's type is in bold).

<Isnad>حدثنا محمد بن إسماعيل حدثنا مكي بن إبراهيم حدثنا موسى بن عبيدة عن عامر بن الحكم بن ثوبان عن عبدالله بن عمرو بن العاص عن أبي حازم عن سهل بن سعد<\Isnad><Matan> دون مرفوعاً الله تعالى سبعون ألف حجاب من نور وما تسمع نفس شيئاً من حسن تلك الحجب إلا زهقت نفسها قال أبو الشيخ في العظمة ذكر حجب ربنا تبارك وتعالى فبدأ بهذا الحديث ثم بعده<\Matan>

Figure 8: An example of confusion between an Isnad and a Matan, from the first experiment, because of different Hadith structures in training and testing sets.

We classified some Hadiths as hard Hadiths owing to having a story in the Isnad or between Isnad and Matan which makes the segmentation task more complex. There are two different type of these stories: a narrative story and a chronology story. The narrative story refers to any story related to the narrator such as describing where did he live, his age, who did he meet and so on. The chronology story means telling a sequence of events in order (Sternberg, 1990) such as describing the first event which is the prophet Muhammad and his companions' scene, why did he say a certain Hadith or the person/ group of people who came to ask him and then the following event will be the Matan. We labelled the narrative story as Isnad and the chronology story as Matan. Figure 9 shows an example of the narrative story wrongly predicted as Matan.

### 5.2 Second Experiment

In this experiment, we used Isnad and Matan training models that were produced from the first experiment. The LK Hadith corpus was chosen for testing purposes. It is a parallel corpus of English-Arabic Hadith, containing 39,038 annotated Hadiths from the six canonical Hadith books.

From the LK corpus, we manually extracted

<Isnad>وقال الطبراني حدثنا علي بن سعيد الرازي حدثنا محمد بن حاتم المؤدب حدثنا القاسم بن مالك المزني حدثنا سفيان بن زياد عن عمه سليم بن زياد <\Isnad><Matan>قال لقيت عكرمة مولى ابن عباس فقال لا تبرح حتى أشهدك على هذا الرجل ابن لمعاذ بن عفراء فقال أخبرني بما أخبرك أبوك عن قول رسول الله صلى الله عليه وسلم فقال حدثني أبي أن رسول الله صلى الله عليه وسلم حدثه أنه رأى رب العالمين عز وجل في حظيرة من القدس في صورة شاب عليه تاج <\Matan>

Figure 9: An example of the narrative story wrongly predicted as Matan using PPMD with Order 7 (highlighted in blue).

chapters two and three from the *Sahih Al-Bukhari* book, comprising a testing file of 10,539 words. We noticed that the last part of Isnads, such as قال وسلم عليه الله صلى النبي 'the Prophet, may God bless him and grant him peace, said', were labelled as Matan so we relabelled these parts as Isnad for consistency with the labelling throughout. Then we removed Arabic diacritics (Al-Tashkeel) and quotation marks.

Order 7 was chosen since it had a higher accuracy rate in the first experiment. The Hadith segmentation using PPMD produced an accuracy of 90.10%, an average precision of 0.9249, an average recall of 0.8607, and an average F-measure of 0.8914. Figure 10 shows the confusion matrix of this experiment and Figure 11 shows an example of the chronology story correctly predicted as Matan from this experiment.

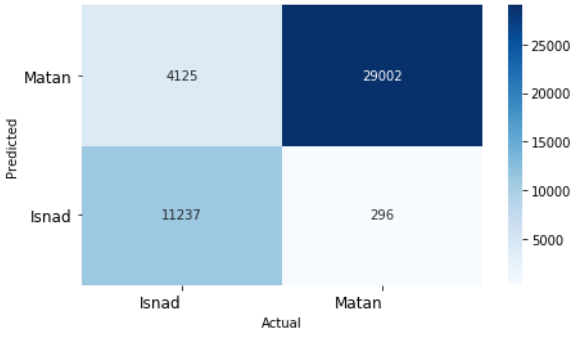


Figure 10: Confusion matrix of the second experiment's results.

## 6 Conclusion and Future Work

In this paper, we evaluated PPM compression-based method for automatic segmentation of

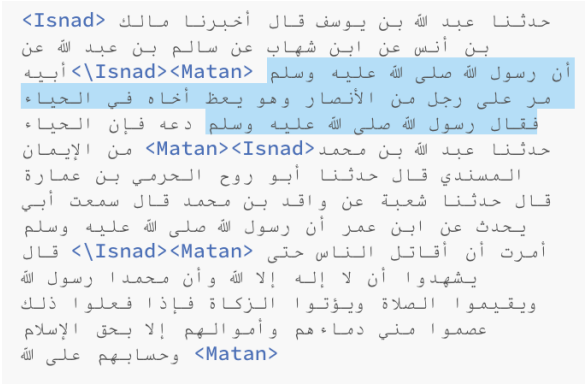


Figure 11: An example of the scene's story correctly predicted as Matan from the second experiment (highlighted in blue).

Arabic Hadith. The experiments showed that PPMD is effective in segmenting Hadith into its two main components (Isnad and Matan), having been tested on Hadith corpora (NAH and LK) that have different structures. The main innovation in these experiments is their use of a character-based text compression method to segment Hadith.

For training Isnad and Matan models we used the first book in the NAH corpus. In the first experiment, we used the third book in the NAH corpus, which lacks a clear structure, as a testing set. We found that PPMD of order 7 obtained a higher accuracy (92.76%) than other orders. In the second experiment, we aimed to evaluate PPMD segmentation on a different Hadith corpus so we used the *Sahih Al-Bukhari* book of the LK Hadith corpus for testing purposes, which produced an accuracy of 90.10%.

The first experiment showed that the Hadith's type is not in the same place between the training and testing set, which leads to some confusion between Isnad and Matan. Possible ways to reduce this confusion that could be undertaken in future work may be to (1) extend the Isnad training set to have different Isnads structured from different Hadith books, (2) clean the testing set from all non-Isnad words.

## Acknowledgments

The first author is grateful to the Saudi government for their support.

## References

Mohammed Altamimi and William Teahan. 2017. Gender and authorship categorisation of arabic text from twitter using ppm. *International Journal of Computer Science and Information Technology*, 9:131–140.

Shatha Altammami, Eric Atwell, and Ammar Alsalka. 2019. Text segmentation using n-grams to annotate hadith corpus. In *Proceedings of the 3rd Workshop on Arabic Corpus Linguistics*, pages 31–39, Cardiff, United Kingdom. Association for Computational Linguistics.

Shatha Altammami, Eric Atwell, and Ammar Alsalka. 2020. The arabic–english parallel corpus of authentic hadith. *International Journal on Islamic Applications in Computer Science And Technology*, 8(2).

Aqil Azmi and Nawaf Bin Badia. 2010. itree - automating the construction of the narration tree of hadiths (prophetic traditions). In *Proceedings of the 6th International Conference on Natural Language Processing and Knowledge Engineering(NLPKE-2010)*, pages 1–7.

John Cleary and Ian Witten. 1984. Data compression using adaptive coding and partial string matching. *IEEE Transactions on Communications*, 32(4):396–402.

Fouzi Harrag. 2014. Text mining approach for knowledge extraction in sahîh al-bukhari. *Computers in Human Behavior*, 30:558–566.

Paul Glor Howard. 1993. The design and analysis of e cient lossless data compression systems. *Diss. PhD thesis, Brown University.*

Ahsan Mahmood, Hikmat Ullah Khan, Fawaz K Alarfaj, Muhammad Ramzan, and Mahwish Ilyas. 2018. A multilingual datasets repository of the hadith content. *International Journal of Advanced Computer Science and Applications*, 9(2).

Hajer Maraoui, Kais Haddar, and Laurent Romary. 2019. Segmentation tool for hadith corpus to generate tei encoding. In *Proceedings of the International Conference on Advanced Intelligent Systems and Informatics 2018*, pages 252–260, Cham. Springer International Publishing.

Alistair Moffat. 1990. Implementing the ppm data compression scheme. *IEEE Transactions on Communications*, 38(11):1917–1921.

Matthew S Ryan and Graham R Nudd. 1993. The viterbi algorithm. *University of Warwick. Department of Computer Science.*

Mohammad Arshi Saloot, Norisma Idris, Rohana Mahmud, Salinah Ja'afar, Dirk Thorleuchter, and Abdullah Gani. 2016. Hadith data mining

and classification: a comparative analysis. *Artificial Intelligence Review*, 46(1):113–128.

Meir Sternberg. 1990. Telling in time (i): Chronology and narrative theory. *Poetics Today*, 11(4):901–948.

Taghreed Tarmom, Eric Atwell, and Mohammad Alsalka. 2020a. Non-authentic hadith corpus: Design and methodology. *International Journal on Islamic Applications in Computer Science And Technology*, 8(3).

Taghreed Tarmom, William Teahan, Eric Atwell, and Mohammad Ammar Alsalka. 2020b. Compression versus traditional machine learning classifiers to detect code-switching in varieties and dialects: Arabic as a case study. *Natural Language Engineering*, page 1–14.

William John Teahan. 2000. Text classification and segmentation using minimum cross-entropy. In *Content-Based Multimedia Information Access - Volume 2*, RIAO '00, page 943–961, Paris, FRA. Le Centre De Hautes Etudes Internationales D'informatique Documentaire.

William John Teahan. 2018. A compression-based toolkit for modelling and processing natural language text. *Information*, 9(12):294.

A. Viterbi. 1967. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Transactions on Information Theory*, 13(2):260–269.