Text Segmentation Using N-grams to Annotate Hadith Corpus

Shatha Altammami
School of Computing
University of Leeds
Leeds,UK
scshal@leeds.ac.uk

Eric Atwell
School of Computing
University of Leeds
Leeds,UK
E.S.Atwell@leeds.ac.uk

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

Abstract

In this paper, we exploit natural language processing techniques to build a system that automatically segments Hadith into its two main components, Isnad and Matn. We evaluate the previous attempts to segment Hadith and identified the limitations in these studies. Then a Hadith segmentation system is built and tested with Hadith collections extracted from six different Hadith books. The result demonstrates that bi-grams are effective in identifying Hadith segments with 92.5% accuracy.

Ali bin Mohammed told us, Wakia told him that Younis bin Abi Ishaq heard Mujahid, heard Abu Hurayrah said the Messenger of Allah peace be upon him (PBUH) said: "Jibra'il kept enjoining good treatment of neighbours until I thought he would make neighbours heirs."

حَدَّثَنَا عَلِيُّ بْنُ مُحَمَّدٍ، حَدَّثَنَا وَكِيعٌ، حَدَّثَنَا يُونُسُ بْنُ أَبِي إِسْحَاقَ، عَنْ مُجَاهِدٍ، عَنْ أَبِي هُرَيْرَةَ، قَالَ قَالَ رَسُولُ اللَّهِ صلى الله عليه وسلم " مَا زَالَ جِبْرَائِيلُ يُوصِينِي بِالْجَارِ حَتَّى ظَنَنْتُ أَنَّهُ سَيُوَرِّثُهُ " .

Figure 1: Hadith Example, Isnad in bold

1 Introduction

Advancement in Artificial Intelligence (AI), specifically Natural Language Processing (NLP), encouraged researchers to tackle problems associated with textual data. In this work, we exploit NLP methods to build a system that automatically segments Hadith text, which is the collection of narratives reporting different aspects of Prophet Muhammad's life.

Hadith originated in the 7th century and is considered classical Arabic with unique structure, linguistic features, and patterns that make it suitable for applying computational models. Moreover, Hadith possesses historical importance and is still used by Muslims around the world. This is because not all Islamic laws and regulations are mentioned in the Islamic holy book, the Quran. Hence, producing Hadith resources will be useful for a wider community including Islamic scholars, historians, linguists, and computer scientists.

Before building such Hadith resources, it is crucial to explain what constitutes the Hadith literature. It is a huge collection of Hadiths[1] that record every aspect of the prophet's life. In addition to that, there are supporting documents which include commentaries that explain the Hadith text, and biographic material that identify narrators of Hadith, which is central in studying Hadith authenticity.

Hadith types vary, it could be a short sentence or long paragraph describing what the prophet said in a specific incident, a dialogue of the prophet's conversation with someone, or a story told by the prophet's companions that explain the prophet's actions in a specific matter like 'prayer'. Every Hadith consists of two parts, as shown in Figure 1, the **Isnad** which is a chain of narrators shown in bold face text, followed by **Matn** which is the actual teaching. However, these parts exist as one sentence or paragraph where researchers manually segment Isnad from Matn to focus on one part (Luthfi et al., 2018). For example, researchers analyse Isnad to visualizing the chain of narrators (Muazzam Siddiqui, 2014) and identify how Hadith travelled through time. While other researchers focus on Matn with the aim to categorize Hadith into subtopics (Saloot et al., 2016).

Although there are research in the area of Hadith computation, it is still in its infancy with limit contributions (Bounhas, 2019). In fact, there are no common datasets, benchmarks, or evaluation measures as pointed in a recent survey of research

[1]The plural of Hadith is AHadith, but we will use Hadiths

on Arabic NLP (Guellil et al., 2019). It shows that only 11% of the available Arabic resources are classical-Arabic and none dedicated for Hadith. These resources are mostly dedicated for the Quran, such as Dukes and Atwell (2012). Therefore, we aim to build a segmentation model to produce a Hadith corpus where Isnad and Matn are annotated to serve the wider research community and advance the field.

This paper is organized as follows, we survey previous attempts to identify Isnad segments in section 2 and discuss their limitations. Then we introduce the data used to build and test our system in section 3. After that we present our Hadith segmentation model in section 4, and discuss the results in section 5.

2 Literature Review

Surveying the literature, we found a number of attempts to detect Isnad patterns. Table 1 shows summaries of such work and the following paragraphs discuss these papers and their limitations.

Muazzam Siddiqui (2014) attempted to segment Isnad from Matn by using supervised machine learning (ML) algorithms that require an annotated corpus. Thus, a native Arabic speaker annotated Hadith tokens extracted from Sahih Al-Bukhari[2] into five classes (beginning of Person, inside Person, beginning of Narrator, inside Narrator, and Other) where Narrator corresponds to names in Isnad, and Person corresponds to names in Matn.

After annotating the corpus, they studied Hadith contextual patterns to identify features to be used in classification. For example, the word '*told us* -حدثنا' is followed by a narrator's name in Hadith Isnad, and the word '*son of* - بن' is part of a name. Another example is the honorific phrases that always follows a person's name. The classifier takes the training data and features in the form of 'feature, class' where each word is classified as 'beginning/inside Person, beginning/ inside Narrator, or Other'. Then the system classifies new Hadith tokens and segments the Hadith by finding the end of the consecutive list of narrators.

Their system's performance was measured based on two factors. First, its ability to assign each token the correct type irrespective of the boundaries as long as there is an overlap. Second, its ability to correctly find the boundary of each name independent of type assigned (narrator or person). The system produced 90% accuracy in the testing phase. Then for evaluation, they used another manually labeled Hadith book titled 'Musnad Ahmed' which contains 5K tokens that produced 80% accuracy.

Harrag (2014) built a Finite State Transducers (FST) system to extract Hadith segments which include Num-Kitab, Title-Kitab, Num-Bab, Title-Bab, Num-Hadith, Isnad, Matn, Taalik, and Atraf. He used the beginning of words like 'K' for Kitab to identify the book title. Furthermore, he used punctuation to identify other parts of the Hadith assuming that all Matn is surrounded by parenthesis. These features depend on the Hadith book used and how well and correctly it is punctuated. Thus, it cannot be applied to all kinds of Hadith books. His work measures the system's performance to identify several components of Hadith that range from Isnad and going deeper into identifying the narrator's names. However, for the purpose of our study we only report the result of Isnad extraction which was 44% precision.

Azmi and Bin Badia (2010) built a system that aims to draw the tree of narrators, but first Isnad must be extracted. To extract Isnad, they identified pre-processing steps which include removing diacritics and punctuation; apply shallow parsing to handle noise and leave out words which it is not able to parse. Using the shallow parsing output, Noun phrases were the candidate of being a narrator's name. After pre-processing the data, they applied context free grammar to identify each segment in the Hadith by comparing the tokens to the list of Hadith lexicon they compiled earlier. Since the goal of their study is to build the tree, their result reflects the system's success to draw the tree and not the segmentation part.

Maraoui et al. (2018) compiled a list of trigger words that come before, after, and between narrator's names. Furthermore, they identified words that mark the termination of each Hadith which are 'أطرافه' or 'تحفه'. Using these comprehensive lists of words, they were able to segment Isnad from Matn for Sahih Albukhari. However, it is not clear whether it can be used to segment other Hadith books.

Boella (2011) presented HedExtractor system which uses regular expressions (Regex) to extract

[2]Mohammad AlBukhari, Sahih AlBukhari (2002). Damascus: Dar Ibn Kathir.

Paper	Approach	Technique	Pre-processing	Manual annotation	Data	Result
(Muazzam Siddiqui, 2014)	Machine Learning	Nave Base, KNN, Decision tree	Remove diacritics, stemming	Person, Narrator, other	Albukhari Musnad Ahmed	80%
(Harrag, 2014)	Finite State Transducer	-	Tokenize,	None	Albukhari	44%
(Azmi and Bin Badia, 2010)	Rule Based	Context Free Grammar	Shallow parsing Remove diacritics and punctuation	Hadith Lexicons	Albukhari	87%
(Maraoui et al., 2018)	Rule Based	Dictionary Lookup	None	Hadith Lexicons	Albukhari	96%
(Boella, 2011)	Rule Based	Regular Expressions	Transliteration	Hadith Lexicons	Albukhari	97%
(Mahmood et al., 2018)	Rule Based	Regular Expressions	None	None	Muslim, Albukhari, Abu Dawud, Imam Malik	98%

Table 1: Review of Previous Research on Hadith Segmentation

Hadith. First, it extracts each Hadith from the book by finding the number of each Hadith. Then the Arabic text is converted to its transliteration to find the words of transmission based on a list they compiled. It assumes any words between these transmission words are the narrator's names. Once there are no transmission terms detected, the system marks the end of Isnad. However, the exact point of Hadith segmentation is sometimes ambiguous even for humans. To overcome this problem, they have set a threshold of 100 characters which they picked based on trial and error. This threshold tells the system if no other transmission word is detected within the next 100 characters then Matn starts.

Mahmood et al. (2018) extracted Hadiths from different sources, but did not mention any Hadith lexicon list compiled to be used in the Regex. In fact, Hadith lexicons were encoded in the Regex which is not fit for re-usability.

2.1 Limitation of Previous Studies

In the previous research, Hadith segmentation was done using three approaches. First, rule-based that consists of whitelists (or gazetteers) to identify names and Isnad specific words, a filtration mechanism (or Blacklists) to identify Matn words, and grammar rules (as a set of regular expressions) to identify the segmentation point. Second is the ML approach which consists of data annotation, feature and algorithm selection, training and classification. The third is the FST which depends on the degree of consistency in a well-structured text.

Looking at Table 1 above, it is evident that rule-based produced better results. However, it is not clear if the rule-based approach designed for one book can be applied to all Hadith books. In fact, researchers in Mahmood et al. (2018) created different regular expressions for the different Hadith books which imply that rule-based approaches cannot be universal. In the other hand, the ML approach is no better since it requires training data that represent all kinds of Hadith to make its performance acceptable when applied to the different Hadith books. For example, the study presented in Muazzam Siddiqui (2014) reported a drop in performance by 10 points once the model was tested with a different book. Another problem is associated with FST, segmentation will not work if the Hadith book does not use unique punctuation that surrounds each segment e.g. parenthesis around Matn.

Although we try to compare systems performance in the table above, it is crucial to clarify that the approaches are not comparable for two reasons. First, the data used to test the systems

are different in terms of size and type. Second, the way system performance was measured is different in every study. For example. The study in Muazzam Siddiqui (2014) measured the precision of the system's ability to annotate the person's name as Narrator or not. That is whether each name is part of Isnad or Matn. Therefore, their system goal is not to segment but rather to identify narrators. To sum up, the results column in the table above for papers (Harrag, 2014)(Maraoui et al., 2018) (Boella, 2011)(Mahmood et al., 2018) reflect the segmentation performance, while the other studies report the performance of named entity recognition(NER) of narrators in Hadith, which can be used in segmentation.

3 Data Preparation

Before building the segmenter, testing data must be prepared. There is a countless number of Hadith books with a varying degree of authenticity. For the purpose of our project, we include the six famous books, commonly referred to *The Authentic Six* or canonical Hadith books. These books are *Sahih Albukhari*, *Sahih Muslim*, *Sunan Abu Dawood*, *Sunan Altarmithi*, and *Sunan Ibn Maja*. From each book, 40 Hadiths were carefully chosen to form 240 Hadiths that include Hadiths with irregular patterns. This is to ensure we accomplish two goals. First, overcome the limitation of previous studies that relied on one book; second, to produce a realistic performance of a segmenter that can deal with all types of Hadith books.

Then we gathered data required by our system to segment Hadith, which we refer to as training data. It is a list of Isand and Matn segments extracted from a well-structured Hadith book 'Sahih Albukhari'. To automate this task, we scrutinize Hadith parts to find that Isnad consists of a closed set of words that includes narrators' names and transmission words. The example of Isnad in Figure 2 underlines the unique words in Isnad. A common pattern in Isnad is the narrator's name which takes the form of *'first name - son of- father's name'*, so it is two names connected by a 'relation' word. Narrator's names are usually followed by 'transmission' words that reflect how the Hadith was reported e.g. *x 'heard' y* or *x 'said'*. Hence, transmission words will appear four words apart at most. Using this information, we created a list of Isnad lexicons that consists of 'transmission' and 'relation' words. Then we created

Father of Naim *said* Shaiban *told us*, *from* Yahya, *from* Abdullah *bin* Abi Qatada, *from* his father, *said* the Messenger of Allah peace be upon him *said*, 'If the Iqama is pronounced, then do not stand for the prayer till you see me (in front of you) and do it calmly.' Confirmed by Ali bin Mubarak.

حَدَّثَنَا أَبُو نُعَيْمٍ، قَالَ حَدَّثَنَا شَيْبَانُ، عَنْ يَحْيَى، عَنْ عَبْدِ اللَّهِ بْنِ أَبِي قَتَادَةَ، عَنْ أَبِيهِ، قَالَ قَالَ رَسُولُ اللَّهِ صلى الله عليه وسلم " إِذَا أُقِيمَتِ الصَّلاَةُ فَلاَ تَقُومُوا حَتَّى تَرَوْنِي وَعَلَيْكُمْ بِالسَّكِينَةِ ". تابعه علي بن مبارك.

Figure 2: Isnad Example, Isnad lexicons underlined

a python script that tokenizes a Hadith and takes four words at a time to check if an Isnad lexicons is present. Once it detects a group of four words with no Isnad lexicons, it assumes the beginning of Matn text and separates the Hadith at that point. This approach will automatically detect Isnad with regular patterns only, so this step intends to collect the various names of narrators instead of segmenting the Hadith.

We manually checked the result of this bootstrapping approach that produced a collection of more than four thousand Hadiths to form our gold standard of Isnad and Matn segments.

4 Segmentation Model

In this section, we discuss the techniques and algorithms used to build the Hadith segmenter. Figure 3 shows the structure and components of the Hadith segmenter model which takes in a new Hadith that goes through a preprocessing phase to remove diacritics and punctuations. Then Hadith is tokenized into N-grams depending on the technique applied, next each token is labelled as Isnad, Matn or Neither by comparing it with pre-compiled lists obtained from the gold standard created earlier as explained in section 3. Once every token is labelled, the model finds the best segmentation point of the Hadith. In the following lines, we give details of the techniques used to annotate Hadith tokens.

4.1 Tri-gram Technique

In a previous study, we have shown that considering the meaning of words by using the word em-

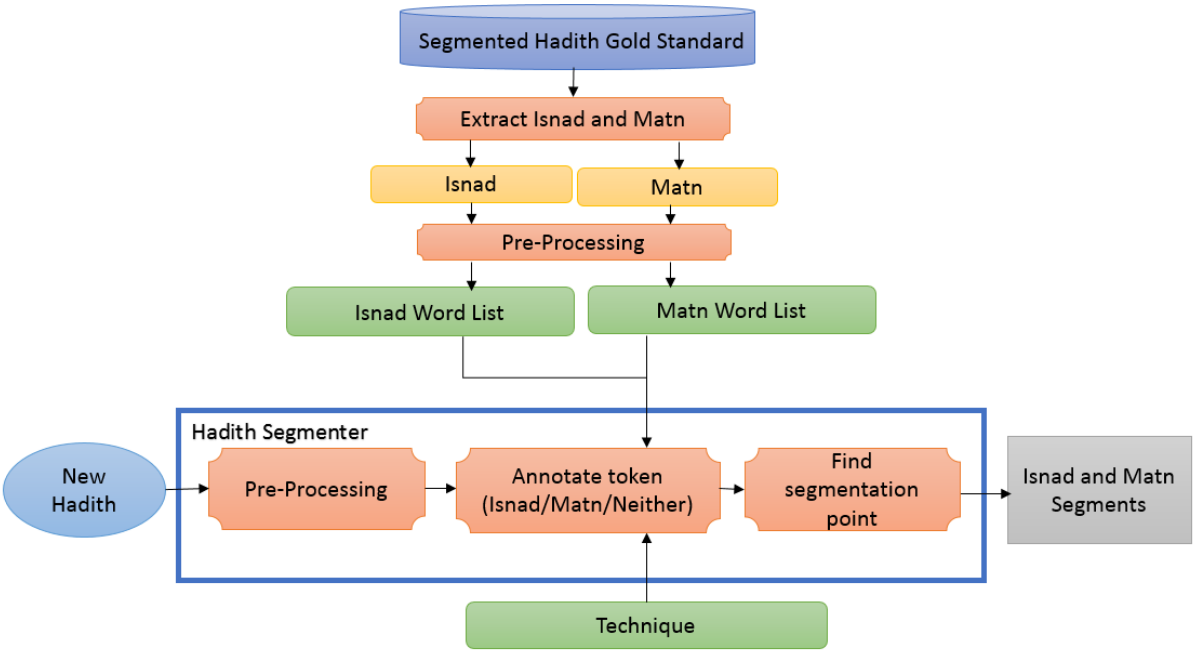

Figure 3: Hadith Segmenter Model

bedding technique does not perform well in Hadith segmentation. This is because such an approach relies on uni-grams that do not capture the unique pattern in Isnad. Furthermore, some words exist in both Isnad and Matn segments, Hence, in this experiment, we aim to capture Isnad patterns by using the N-gram technique. As illustrated in Figure 3, Isnad and Matn segments are extracted from the gold standard of segmented Hadiths. Then they are pre-processed to remove diacritics and punctuations. Finally, tri-gram, bi-gram and uni-gram lists for Isnad and Matn are created to be the evaluation lists in the annotation phase. The reason three lists are created is that a back-off approach can handle irregularity and missing information. For example, if an encountered tri-gram has no match in the tri-gram list of the training data, it can be annotated according to its components. Consider a narrator's full name is not captured in the lists, then Hadith lexicons like 'بن - *son of*' will enable the system to identify this tri-gram as part of the narrator's chain and label it 'Isnad'. This approach is detailed in Algorithm 1. Once every token is labelled, the system finds the segmentation point as detailed in Algorithm 2.

This approach produced 48% accuracy where only 115 out of 240 Hadiths were correctly segmented. To understand this disappointing result, we inspect the incorrectly segmented Hadith and found that the system rarely used the tri-gram fea-

Algorithm 1 Annotate Tri-gram tokens

Tokenize Hadith into Tr-igrams T
for $t \in T$ **do**
  if $t \in IsnadTrigramList$ **then**
    $Label\ t$ **Isnad**
  else if $t \in MatnTrigramList$ **then**
    $Label\ t$ **Matn**
  else
    Convert t to Bigrams b
    if $b \in IsnadBigramList$ **then**
      $Label\ t$ **Isnad**
    else if $b \in MatnBigramList$ **then**
      $Label\ t$ **Matn**
    else
      Convert t to Unitgram u
      if $u \in IsnadUnigramList$ **then**
        $Label\ t$ **Isnad**
      else if $u \in MatnUnigramList$ **then**
        $Label\ t$ **Matn**
      else
        $Label\ t$ **Niether**

Output:

Token1	Lable1	Token2	Label2	...

Isnad	Matn
حدثنا قتيبة حدثنا مروان بن معاوية الفزاري عن أبي يعفور عن الوليد بن العيزار عن أبي عمرو الشيباني أن رجلا قال لابن مسعود أي العمل أفضل قال سألت عنه رسول الله صلى الله عليه وسلم فقال الصلاة على مواقيتها قلت وماذا	يا رسول الله قال وبر الوالدين
Qutaiba told us Marwan bin Muawiya al-Fizari from Abu Yafour from Al-Walid bin Al-Azar from Abu Amr AlShibani that a man said to Ibn Masood, which work is better? He said I asked the Messenger of Allah (PBUH): "Which action is dearest to Allah?" He (PBUH) replied, "Performing the prayer at its earliest fixed time." I asked, "What is next ?"	O prophet, He said, "Kindness towards parents."

Table 2: Example of incorrect segmentation when applying tri-gram technique.

Algorithm 2 Find Segmentation point

```
for every token in Output List do
    if label is Matn then
        if followed by Matn or Neither then
            Mark it as Segmentation Point
            Break
        else if label is Neither then
            if followed by two labels Matn
              or Neither                    then
                Mark it as Segmentation Point
                Break
```

ture, but rather relied on the bi-gram and uni-gram features to annotate tokens. Consider the example in Table 2, feeding this Hadith to the system produces 79 tri-grams, of which only 15 found a match in the tri-gram training set. The remaining 64 tri-grams relied on the bi-gram and uni-gram training set to be annotated. This dependency on bi-gram/uni-gram features to annotate Hadith tri-grams produced unreliable results as illustrated in the example. The phrase 'أن رجلا قال - *that a man said*' should mark the beginning of Matn, instead it was labelled as Isnad. This is because when the system did not find a match in the tri-gram training set, it applied the back-off approach and searched in the bi-gram and uni-gram lists. Since it found a match for the term 'قال *said* ' in the Isnad lists, it labelled the phrase accordingly. Therefore, using tri-grams did not prove useful in our case for two reasons. First, the training data is not large enough to cover all known narrators. Second, it is obtained from only one Hadith book which does not include all Hadith lexicons and patterns.

4.2 Bi-gram Technique

To improve the system performance, we omit tri-gram features and use bi-grams and uni-grams only.

The bi-gram technique produced better results as expected with 222 Hadith out of 240 were correctly segmented, showing 92.5% accuracy. In fact, it is able to segment Hadiths having different structures. For example, the traditional ones where a Matn start with a prophetic saying as shown in Table 3. Other Hadith structures include those containing irregular patterns where Matn starts with an introductory phrase followed by the prophetic saying as shown in Table 4, a dialogue with the prophet as shown in Table 5, or an explanation of a prophetic deed as in Table 6.

Then we analyse the faulty output and found that our system incorrectly segmented some Hadiths with irregular patterns. For example, a Hadith may contain a parallel Isnad, which is a chain of narrators that ends at the prophet followed by another chain of narrators that ends at the prophet again, as shown in Table 7. Another example of an irregular pattern in Isnad is shown in Table 8 which illustrates that Isnad may contain Matn patterns. Finally, Table 9 shows that some Hadith posses a vague segmentation point. Note that for space issues some Hadiths in the examples were truncated as indicated by (...).

Isnad	Matn
حدثنا كثير بن عبيد الحمصي حدثنا محمد بن خالد عن عبيد الله بن الوليد الوصافي عن محارب بن دثار عن عبد الله بن عمر قال Kathir bin Obeid Al-Homsi told us Mohammed bin Khalid from Obidallah bin Walid Al-Wasafi from Moharib bin dathar from Abdullah bin Omar said	قال رسول الله صلى الله عليه وسلم ابغض الحلال الى الله الطلاق The prophet (PBUH) said, Of all the lawful acts the most detestable to Allah is divorce.

Table 3: Correct segmentation, regular pattern.

Isnad	Matn
حدثنا أبو معمر قال حدثنا عبد الوارث عن عبد العزيز قال أنس Abu Muammar told us that Abdul Warith told us from Abdul Aziz said that Anas said	إنه ليمنعني أن أحدثكم حديثا كثيرا أن النبي صلى الله عليه وسلم قال من تعمد علي كذبا فليتبوأ مقعده من النار I refrain from telling you many things about the prophet because I heard the prophet (PBUH) said, ”He who deliberately forges a lie against me let him have his abode in the Hell.”.

Table 4: Correct segmentation, introductory statement.

Isnad	Matn
حدثنا قتيبه قال حدثنا الليث عن يزيد بن ابي حبيب عن ابي الخير عن عبد الله بن عمرو Qaytibah told us Alith from Yazid ibn Abi Habib from Abi Al-Khair from Abdullah bin Amr	ان رجلا سال رسول الله صلى الله عليه وسلم اي الاسلام خير قال تطعم الطعام وتقرا السلام على من عرفت ومن لم تعرف A man asked the Messenger of Allah (PBUH): ”Which act in Islam is the best?” He (PBUH) replied, ”To give food, and to greet everyone, whether you know or you do not.”

Table 5: Correct segmentation, conversation of the Prophet.

Isnad	Matn
حدثنا إسمعيل بن موسى الفزاري حدثنا شريك عن أبي إسحق عن الحارث عن علي بن أبي طالب قال Ismail bin Musa al-Fazari told us Sharik said Abu Ishaq from AlHarith from Ali bin Abi Talib said	من السنة أن تخرج إلى العيد ماشيا وأن تأكل شيئا قبل أن تخرج It is the Sunnah (prophetic tradition) to go out to the Eid prayer walking and eat something before you go out.

Table 6: Correct segmentation, no prophetic words.

Isnad	Matn
حدثنا مسدد قال حدثنا يحى عن شعبة عن قتادة عن أنس رضي الله عنه عن النبي Mosadad said Yahya told us Shoba heard Qatada from Anas may Allah be pleased with him, the Prophet	صلى الله عليه وسلم وعن حسين المعلم قال حدثنا قتادة عن أنس عن النبي صلى الله عليه وسلم قال لا يؤمن أحدكم حتى يحب لأخيه ما يحب لنفس (PBUH), and from Husayn al-Muallim said Qatada told us from Anas that the Prophet (PBUH) said: "No one of you becomes a true believer until he likes for his brother what he likes for himself".

Table 7: Incorrectly segmented, Parallel Isnad.

Isnad	Matn
حدثنا نصر بن علي الجهضمي وأبو عمار والمعنى Nasser bin Ali Juhadhmi and Abu Ammar told us and the meaning	واحد واللفظ لفظ أبي عمار قالا أخبرنا سفيان بن عيينة عن الزهري عن حميد بن عبد الرحمن عن أبي هريرة قال أتاه رجل فقال يا رسول الله هلكت... Is the same but the words are of Ammar they said, Sufian bin Aayneh from Alzahri from Hamid bin Abdul Rahman on the authority of Abu Hurayrah said a man came and said, "O Allah's Apostle! I have been ruined." ...

Table 8: Incorrectly segmented, Isnad contain Matn lexicons.

Isnad	Matn
أخبرنا محمد بن منصور قال حدثنا سفيان قال حدثنا يحى بن سعيد عن مسلم بن أبي مريم شيخ من أهل المدينة ثم لقيت الشيخ فقال سمعت علي بن عبد الرحمن يقول صليت إلى جنب ابن عمر فقلبت الحصى فقال لي ابن عمر Muhammad bin Mansour told us, that Sufian said Yahya bin Said told us about Muslim bin Abi Maryam a Sheikh from Madinah then I met the Sheikh and he said he heard Ali bin Abdul Rahman say I prayed beside Ibn Omar, while I turned the gravel he said	لا تقلب الحصى فإن تقليب الحصى من الشيطان وافعل كما رأيت رسول الله صلى الله عليه وسلم يفعل قلت وكيف رأيت رسول الله صلى الله عليه وسلم يفعل قال هكذ ... Do not fluctuate the gravel, turning the gravel is from the devil and do as I saw the Messenger of Allah peace be upon him do...

Table 9: Incorrectly segmented, names should be part of Matn.

5 Discussion

The findings of this study clearly show that using bi-grams for Hadith segmentation works better than tri-grams specifically because our training data is limited. Although the segmenter result is promising, not all Hadiths with irregular patterns were correctly segmented. In fact, this can be vague even for human annotators who are not experts in Hadith studies. For this reason, we argue that Hadith can be segmented to fine-grained segments that go beyond Isnad and Matn. This is because some Hadith contain information in the Isnad that was identified as Matn segments by our system. For example, a Hadith Isnad may include information about where a specific narrator comes from, then it continues the chain of narrators. An-

other example is a Hadith that starts a Matn segment with a piece of introductory information containing names of people which was identified as Isnad pattern by our segmenter as in Table 9. Thus, we aim to make an enhancement to Algorithm 2 to enable the segmenter output several segments instead of two, then apply probabilistic measures to identify the exact point of segmentation.

6 Conclusion

In this paper, we demonstrate the need for Hadith common datasets and our initiative to bridge the gap by automatically annotating Hadith corpus using NLP. The main objective of this study is to build a system that segments and annotates Hadith components, Isnad and Matn. Before building our system, we evaluated previous attempts to segment Hadith and found that the successful techniques rely on hand-crafted tools that cannot be generalized to segment all Hadith types. Furthermore, these systems were tested on a limited number of Hadiths from a single book. To address these limitations, our segmenter rely on NLP techniques and tested with Hadiths extracted from six books to ensure coverage of all Hadith types. Although it was successful in segmenting Hadith with 92.5% accuracy, examining the incorrect results points us to ways of improvements discussed in the paper.

References

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.

Marco Boella. 2011. Regular expressions for interpreting and cross-referencing hadith texts. *Langues et Littératures du Monde Arabe (LLMA)*, 9(3):25–39.

Ibrahim Bounhas. 2019. On the usage of a classical arabic corpus as a language resource: related research and key challenges. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, 18(3):23.

Kais Dukes and Eric Atwell. 2012. Lamp: a multimodal web platform for collaborative linguistic analysis. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC12)*, pages 3268–3275. European Language Resources Association (ELRA).

Imane Guellil, Houda Saâdane, Faical Azouaou, Billel Gueni, and Damien Nouvel. 2019. Arabic natural language processing: An overview. *Journal of King Saud University-Computer and Information Sciences*.

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

Emha Taufiq Luthfi, Nanna Suryana, and Adbulsamad Hasan Basari. 2018. Digital hadith authentication: A literature review and analysis. *Journal of Theoretical & Applied Information Technology*, 96(15).

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):165–172.

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

Abobakr Bagais Muazzam Siddiqui, Mostafa Saleh. 2014. Extraction and visualization of the chain of narrators from hadiths using named entity recognition and classification. *Int. J. Comput. Linguist. Res*, 5(1):14–25.

Mohammad Arshi Saloot, Norisma Idris, Rohana Mahmud, Salinah Jaafar, Dirk Thorleuchter, and Abdullah Gani. 2016. Hadith data mining and classification: a comparative analysis. *Artificial Intelligence Review*, 46(1):113–128.