Identifikasi Nama Periwayat Hadis Menggunakan *Name Entity Recognition*

Fauzan Ramadhan - 2301182020[1]

[1]Peminatan Komputasi Sosial, Magister Informatika, Universitas Telkom

June 12, 2020

Abstract

Hadits dalam bahasa, berarti Al-Khabar adalah berita atau sesuatu yang dikatakan atau dipindahkan dari seseorang ke orang lain. Hadis adalah salah satu sumber hukum yang digunakan oleh umat Islam setelah Alquran dan sebelum Ijma juga Qiyas. Hadits terdiri dari sanad, matan dan rawi. Sanad adalah rangkaian orang-orang yang menghubungkan hadits dengan Nabi Muhammad. Matan adalah kata dari Nabi Muhammad yang disebutkan setelah sanad itu. Rawi adalah orang yang diriwayatkan hadits. Dalam penulisan hadits, ada nama-nama yang sama atau nama-nama serupa dalam banyak hadits yang berbeda. Masalah kesamaan nama antara satu orang dengan orang lain tetapi orang yang berbeda dikenal sebagai disambiguasi nama. Proses untuk mengidentifikasi periwayat hadis salah caranya adalah dengan mengidentifikasi nama entitas menggunakan NER. Percobaan dilakukan dengan menggunakan dua jenis model identifikasi yaitu model dengan 1 entitas dan model dengan 3 entitas. Keduanya menghasilkan akurasi diatas 80%. Beberapa perubahan perlu dilakukan agar akurasi yang diberikan dapat ditingkatkan.

Kata Kunci
Named Entity Recognition, hadis, disambiguasi nama

1 Latar Belakang

Hadits dalam bahasa, berarti Al-Khabar adalah berita atau sesuatu yang dikatakan atau dipindahkan dari seseorang ke orang lain. Menurut ulama, hadits sebagai segala sesuatu yang disebarkan kepada nabi Muhammad, menjadi sebuah kata, perbuatan, keputusan atau karakternya. Hadis adalah salah satu sumber hukum yang digunakan oleh umat Islam setelah Alquran dan sebelum Ijma juga Qiyas. Hadits terdiri dari sanad, matan dan rawi. Sanad adalah rangkaian orang-orang yang menghubungkan hadits dengan Nabi Muhammad. Matan adalah kata dari Nabi Muhammad yang disebutkan setelah sanad itu. Rawi adalah orang yang diriwayatkan hadits. Seorang periwayat hadis tidak hanya menceritakan satu hadis, tetapi dapat meriwayatkan lebih dari satu hadis[1].

Shahih Bukhori adalah salah satu dari enam buku hadis terkenal dan digunakan dalam studi hadits oleh umat Islam. Dalam penulisan hadits, ada nama-nama yang sama atau nama-nama serupa dalam banyak hadits yang berbeda. Masalah kesamaan nama antara satu orang dengan orang lain tetapi orang yang berbeda dikenal sebagai disambiguasi nama. Secara umum, disambiguasi nama adalah salah satu masalah yang sering terjadi dalam proses evaluasi penelitian yang menghubungkan catatan bibliografi tertentu dengan masing-masing peneliti. Masalah yang diangkat pada disambiguasi nama ada dua macam, yaitu seorang individu yang dipanggil dapat diidentifikasi sebagai dua atau lebih penulis dan dua atau lebih penulis dapat diidentifikasi sebagai seorang individu saja[2].

Proses untuk mengidentifikasi periwayat hadis dengan menggunakan nama memerlukan proses yang panjang dan cukup lama. Secara ringkas, beberapa proses yang perlu dilakukan adalah adanya identifikasi entitas nama periwayat hadis, identifikasi fitur yang bisa memberikan ciri dari tiap nama (khususnya untuk nama yang sama) dan membandingkan nama periwayat hadis yang sama berdasarkan fitur yang sudah didapatkan sebelumnya. Pada penelitian ini, proses yang akan dijelaskan adalah mengenai identifikasi entitas nama periwayat hadis menggunakan *named entity recognition* (NER).

2 Metode dan Dataset

2.1 Identifikasi Nama Orang

Identifikasi nama orang pada sebuah teks dapat dilakukan dengan cara manual dan otomatis. Secara manual, proses identifikasi ini dilakukan dengan membaca keseluruhan teks yang diberikan lalu memberikan tanda ketika ditemukan kata yang merujuk pada nama orang. Proses identifikasi nama ini tentu lebih akurat daripada secara otomatis, tetapi akan memakan waktu yang sangat lama ketika teks yang diberikan banyak atau terdiri dari banyak dokumen. Identifikasi nama orang secara otomatis bisa dilakukan dengan dua cara yaitu dengan pembelajaran (*supervised*) dan dengan menggunakan petunjuk atau ciri-ciri.

Proses identifikasi nama orang menggunakan ciri-ciri secara sederhana dapat dilakukan dengan mencari teks yang mempunyai huruf kapital di awal katanya. Proses ini akan memberikan hasil yang tidak terlalu buruk. Namun, untuk melakukan itu teks yang digunakan harus sudah sesuai dengan aturan yang berlaku. Dalam bahasa Indonesia, cara penulisan nama yang sesuai dengan pedoman umum ejaan bahasa Indonesia (PUEBI) harus diawali dengan huruf kapital. Akan tetapi, tidak semua kata yang diawali dengan huruf kapital adalah kata nama orang. Beberapa penggunaan huruf kapital yang digunakan selain dari merujuk kepada kata nama orang diantaranya adalah huruf pertama kalimat, huruf pertama suatu kalimat langsung, suatu ungkapan dalam agama (kitab suci, Tuhan, malaikat, dll), gelar kehormatan, nama negara, dan seterusnya[3].

Selain itu, nama yang digunakan pada penelitian ini adalah nama dari orang-orang arab yang mempunyai kultur berbeda dengan nama orang Indonesia, seperti adanya kata "bin" atau "binti" yang mengisaratkan "anak dari". Kata "bin" atau "binti" tidak ditulis awal katanya dalam huruf kapital. Hal ini berdampak jika kita menggunakan aturan ini pada kata "Muhammad bin Idris" maka akan dianggap sebagai dua nama orang yaitu "Muhammad" dan "Idris". Padahal seharusnya merujuk pada satu orang saja. Sehingga penggunaan aturan ini tidak menjadi baik jika hanya berpedoman pada aturan 'mencari nama orang dengan cara melihat huruf kapital pada awal katanya saja'. Ada juga *kunyah* (penyebutan nama seseorang dengan merepresentasikan seseorang sebagai bapak atau ibu dari orang lain) dan *laqab* (cara penyebutan seseorang dengan menisbatkan sesuatu pada seseorang) yang sering digunakan oleh orang-orang Arab untuk panggilan nama dari orang lain, seperti "Abu Muhammad" (bermakna bapak dari Muhammad, yaitu Idris) dan "Abdullah" (bermakna hamba Allah)[4].

Cara lain dalam identifikasi nama orang secara otomatis adalah dengan cara *supervised* atau klasifikasi *named entity recognition* (NER), yaitu dengan menyediakan dataset yang akan dipelajari oleh sistem (*data train*) dan data yang akan diuji oleh sistem (*data testing*). Metode ini akan dijelaskan lebih lanjut pada pembahasan mengenai *name entity recognition*.

2.2 *Name Entity Recognition* (NER)

Name entity recognition (NER) adalah suatu bidang kasus (*task*) yang bertugas untuk mengidentifikasi dan mengelompokkan entitas-entitas yang ada pada teks. Entitas dapat berupa kata atau serangkaian kata yang konsisten merujuk pada hal yang sama. Setiap entitas yang terdeteksi dari teks maka akan diklasifikasikan ke dalam kategori yang telah ditentukan. NER merupakan salah satu bentuk dari *natural language processing* (NLP), yaitu salah satu bidang yang dikerjakan pada *artificial intelligence* (AI), yang berfokus pada pemrosesan dan analisis bahasa alami. Secara umum, proses yang dilakukan pada NER adalah mendeteksi suatu nama entitas dan mengkategorisasi nama entitas tersebut. Kategorisasi yang sering dipakai pada NER adalah *person*, *place/location*, *organization* dan *time*[5].

NER juga merupakan salah satu alat bantu untuk melakukan ekstraksi informasi (IE), prosesnya adalah mencari dan mengklasifikasikan entitas yang disebutkan pada teks[6]. Metode dalam pengembangan NER terbagi pada tiga jenis, yaitu pembelajaran terbimbing (*supervised*), pembelajaran tidak terbimbing (*unsupervised*) dan pembelajaran semi terbimbing (*semi-supervised*) [7]. Penelitian ini berfokus pada pembelajaran terbimbing (*supervised*) untuk proses identifikasi nama periwayat. Proses ini dibantu oleh aplikasi NER dari Stanford yang sudah bisa digunakan secara bebas dan hanya tinggal menyediakan data latihnya saja. Adapun proses pengidentifikasian nama entitas dengan menggunakan Stanford NER adalah dengan cara berikut ini.

1. Pastikan komputer atau laptop yang digunakan sudah ter-*install* Java.
2. Unduh Stanford NER dari https://nlp.stanford.edu/software/CRF-NER.html#Download. Setelah itu, *extract* file tersebut.
3. Buatlah sebuah data latih. Proses pembuatan data latih ini akan dijelaskan pada

pembahasan mengenai Sahih Bukhori.

4. Buatlah sebuah file yang berfungsi untuk melatih file data latih tadi. Perintah untuk melatih file data latih dapat diunduh juga pada alamat https://nlp.stanford.edu/software/ner-example/austen.prop. Simpan file tersebut dalam satu folder dengan file data latih tadi. Setelah itu, ubah nama data latihnya dengan data latih yang akan digunakan (pada baris ke 2). Lakukan perubahan nama pada

```
1  #location of the training file
2  trainFile = coba_trn2_hadis.tsv
3  #location where you would like to save
   (serialize to) your
4  #classifier; adding .gz at the end
   automatically gzips the file,
5  #making it faster and smaller
6  serializeTo = ner-model-coba2-hadis.ser.gz
7
8  #structure of your training file; this
   tells the classifier
9  #that the word is in column 0 and the
   correct answer is in
```

Figure 1: File untuk melatih data latih

baris 6, '`ner-model-coba2-hadis.ser.gz`' sesuai dengan keinginan. Perubahan nama dilakukan pada kata sebelum '`.ser.gz`'. Ini dilakukan untuk memberikan nama *classifier* hasil dari melatih data latih tadi. Nama untuk file tersebut bisa diganti sesuai keinginan.

5. Buka *command prompt*. Akses folder tempat disimpannya file tadi dari *command prompt*. Lakukan pelatihan data latih dengan memberikan perintah '`java -cp stanford-ner.jar edu.stanford.nlp.ie.crf.CRFClassifier -prop austen.prop`'. Kata '`austen.prop`' merupakan file untuk melatih data latih yang sebelumnya sudah diinisialisasi.

6. Pada folder `stanford-ner-4.0.0`, jalankan '`stanford-ner-4.0.0.jar`'. Pada aplikasi itu, pilih menu 'Classifier', pilih 'Load CRF from File', lalu pilih file `classifier` yang pada poin sebelumnya telah dilakukan. File dipilih adalah file dengan format '`.ser.gz`'.

7. Masukkan kalimat yang ingin diidentifikasi entitas namanya. Setelah itu, tekan tombol 'Run NER'.

2.3 Sahih Bukhori

Sahih Bukhori merupakan buku hadis yang disusun oleh Imam Bukhori (wafat 256 H). Para ulama pada masa Imam Bukhori dan setelahnya bersepakat bahwa kumpulan hadits yang dimuat dalam kitab Shahih al-Bukhari adalah kitab yang lebih utama dibanding kitab hadits lain. Dua kitab ini dinilai sebagai kitab hadits yang menetapkan syarat-syarat kesahihan hadits yang ketat. Imam an Nawawi (wafat 676 H) memberikan komentar bahwa dua kitab shahih–yaitu Shahih al-Bukhari dan Shahih Muslim–merupakan kitab yang disepakati kesahihan haditsnya oleh ulama ahli hadits. Begitu pula Imam Ibnu Shalah (wafat 643 H), menyebutkan dalam karyanya tentang ilmu hadits yang berjudul Muqaddimah Ibnu Shalah bahwa hadits-hadits yang dihimpun oleh Imam al-Bukhari dan Imam Muslim dalam Shahihain, merupakan hadits sahih yang derajatnya paling tinggi, atau kerap disebut *muttafaq 'alaih*.
Demikian mengapa posisi kitab Shahih al-Bukhari dan Shahih Muslim dipandang istimewa di kalangan ulama ahli hadits, begitu pula ahli fiqih. Melalui syarat-syarat kesahihan yang ketat, Shahih al-Bukhari dan Shahih Muslim telah menghasilkan karya yang istimewa dan selalu dirujuk serta dikaji umat Islam[8]. Buku sahih Bukhori mengandung kurang lebih 6000 hadis yang telah didapatkan dari guru-gurunya. Hadis-hadis ini terbagi menjadi 9 volume dan terdiri dari 93 bab yang meliputi keseluruhan aspek kehidupan.
Dari keseluruhan hadis pada buku sahih Bukhori, penelitian ini hanya menggunakan 100 hadis sebagai data latih dan 20 hadis sebagai data uji. Data uji diambil secara langsung dari buku sahih Bukhori tanpa perlu ada proses penyuntingan redaksi atau format. Sedangkan untuk data latih akan disesuaikan dengan format yang bisa dibaca oleh sistem. Adapun proses penyesuaian yang dimaksud adalah sebagai berikut.

1. Buatlah sebuah data latih. Data latih tersebut dibentuk menjadi dua kolom dengan kolom pertama adalah kata dari data latih dan kolom kedua adalah keterangan entitasnya.

2. Keterangan entitas diisi dengan 'PERSON' jika kata tersebut adalah kata yang merujuk ke nama seseorang, 'ORGANIZATION' jika kata tersebut adalah kata yang merujuk ke nama organisasi. Dalam beberapa format disebutkan sampai 7 keterangan entitas, yaitu 'person', 'organization', 'location', 'date', 'money', 'percent', dan 'time'. Adapun untuk pemisah antara kolom pertama dengan kolom kedua adalah dengan menggunakan tab pada *keyboard*. Setelah jadi, file tersebut disimpan dalam format '.tsv'. Berikut adalah contoh dari implementasi pembuatan data latih.

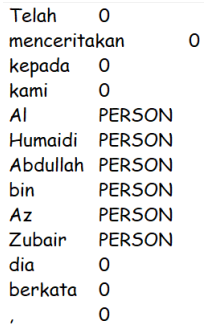

Figure 2: Format data latih

File ini nantinya akan digunakan pada proses identifikasi entitas nama yang sudah dijelaskan pada pembahasan mengenai *name entity recognition*.

3 Hasil dan Analisa

Percobaan identifikasi nama entitas untuk hadis dengan menggunakan aplikasi NER dari Stanford ini dilakukan dengan menggunakan 100 data hadis sebagai data latih dan 20 data hadis sebagai data uji. Data latih dibagi menjadi 2 bagian, yaitu data latih dengan 1 entitas yaitu hanya menggunakan entitas nama orang (*PERSON*) dan data latih dengan 3 entitas, yaitu entitas nama orang (*PERSON*), nama tempat (*LOCATION*) dan nama kaum (*ORGANIZATION*). Selanjutnya akan dilakukan perbandingan antar keduanya dan akan dilihat data latih yang paling baik antara keduanya. Berikut adalah contoh data akan digunakan sebagai data latih dengan tiga jenis entitas dan satu entitas.

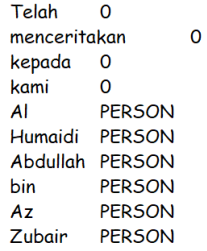

Figure 3: Data latih dengan 1 entitas

Data uji yang digunakan hanya berbentuk kalimat dari 20 hadis saja, yang diambil secara langsung dari file hadis Bukhori dan disimpan pada sebuah file tanpa ada perubahan atau penyesuaian formatn terlebih dahulu. Teks ini selanjutnya ditempelkan ke tempat teks yang ada pada aplikasi Stanfor NER. Tabel 1 adalah lima data teratas hasil identifikasi entitas dari data uji dengan menggunakan model yang telah dibangun sebelumnya. Tabel 2 adalah keseluruhan

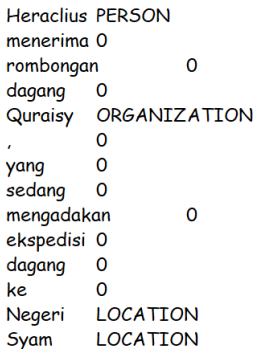

Figure 4: Data latih dengan 3 entitas

Table 1: 5 Data teratas perbandingan hasil model 1 entitas

	Model 1 Entitas				
	101	102	103	104	105
Kenyataan	7	5	5	5	11
Prediksi	5	4	4	5	10

hasil pegujian 20 data uji terhadap model identifikasi entitas untuk 1 entitas dan 3 entitas. Pada

Table 2: Keseluruhan hasil pengujian menggunakan model 1 entitas dan 3 entitas

	1 Entitas	3 Entitas		
	Person	Person	Location	Organization
Kenyataan	179	179	4	6
Prediksi	160	161	4	5
Perbedaan	19	18	0	1
Akurasi	89.4%	89.9%	100%	83.3%

kenyataannya, proses identifikasi suatu nama entitas tidak selalu berjalan dengan benar, sehingga kadang ada entitas yang tidak teridentifikasi dengan benar. Sebagai contoh, nama `Zaid bin Khalid al-Juhani` hanya dikenali sebagai `Zaid bin Khalid` saja. Hal ini ditemui pada contoh nama yang serupa. Contoh lain adalah nama `asy-Sya'bi` tidak dikenali sebagai suatu entitas nama dari seseorang. Analisis yang sampai saat ini dilakukan adalah karena kata yang tidak dikenali itu tidak diawali dengan huruf kapital dan adanya strip (-) pada pertengahan katanya. Dalam data latih, kata-kata yang bersifat pengulangan (seperti kata bintang-bintang) itu dibuat menjadi terpisah (bintang, -, bintang) dan masing-masing bernilai 0 atau tidak teridentifikasi sebagai sebuah entitas, sehingga memberikan anggapan bahwa ketika ada dua kata dengan strip (-) ditengahnya akan dianggap sebagai selain dari suatu entitas. Selain itu, ada juga kata selain entitas nama diklasifikasikan sebagai nama, contohnya adalah kata `Al Waqiah`.

Analisis pada kasus seperti ini adalah adanya dua kata yang tersusun dan keduanya diawali oleh huruf kapital, sehingga aplikasi Stanford NER menganggap bahwa kata ini adalah entitas nama. Kesimpulannya adalah aplikasi Stanford NER ini tidak buruk karena bisa menghasilkan akurasi diatas 80% walaupun menggunakan data latih dan data uji dengan bahasa lain. Namun, perlu ada sedikit penyesuaian pada algoritmanya sehingga beberapa kesalahan kecil dalam identifikasi nama entitas dapat diperbaiki dan dapat menyesuaikan dengan data latih dan data uji yang digunakan.

4 Pengakuan

Ucapan terima kasih atas dedikasi yang tinggi pada semua elemen yang sudah membantu pada percobaan kali ini. Elemen-elemen tersebut adalah sebagai berikut.

1. Tulisan mengenai NER Bahasa Indonesia dengan Stanford NER yang ditulis oleh bapak Yudi Wibisono pada *wordpress*-nya di https://bit.ly/yudiNER
2. Penjelasan lebih lanjut terkait cara proses pelatihan data latih, ditulis oleh The Stanford Natural Language Processing Group pada *website*-nya di https://nlp.stanford.edu/software/crf-faq.html#a
3. File sahih Bukhori yang didapat dari https://bit.ly/dutaShohih
4. Data latih dan data uji yang digunakan, model yang sudah dibangun untuk 1 entitas dan 3 entitas, *gold standard*, dan laporan dapat diunduh pada https://github.com/KangOjan/teksmin-12Juni2020.git

References

[1] Zainul Arifin. Studi kitab hadis, 2013.

[2] Staša Milojević. Accuracy of simple, initials-based methods for author name disambiguation. *Journal of Informetrics*, 7(4):767–773, 2013.

[3] Dika. Pedoman penulisan huruf kapital (besar) yang benar sesuai puebi. https://pedomane.com/penulisan-huruf-kapital/, 2019. Accessed: 2020-06-08.

[4] Bertrand Lisbach and Victoria Meyer. *Linguistic identity matching*. Springer, 2013.

[5] Christopher Marshall. What is named entity recognition (ner) and how can i use? https://bit.ly/mediumNER, 2019. Accessed: 2020-06-09.

[6] Susan Li. Named entity recognition with nltk and spacy. https://bit.ly/towardsNER, 2018. Accessed: 2020-06-09.

[7] David Nadeau and Satoshi Sekine. A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 30(1):3–26, 2007.

[8] Mahbib Khoiron. Keutamaan 'shahih al-bukhari' dan 'shahih muslim'. https://bit.ly/shahihNU, 2018. Accessed: 2020-06-08.