INFORMATION EXTRACTION BERBASIS RULE UNTUK SOAL UJIAN

Stefanus Nico Soenardjo, *Teknologi Informasi Institut Sains dan Teknologi Terpadu Surabaya*,
Gunawan, *Teknik Informatika Institut Sains dan Teknologi Terpadu Surabaya*

***Abstrak*—Proses information extraction dapat dilakukan pada beberapa macam media, seperti artikel berita, tanya jawab dan sebagainya. Penelitian ini mencoba untuk melakukan information extraction pada media soal ujian yang dilengkapi dengan jawaban.**

Pendekatan pengolahan informasi yang dibahas dalam penelitian ini adalah information extraction berbasis rule. Informasi yang hendak digali adalah informasi data soal ujian beserta jawabannya. Inputan dalam penelitian ini pasangan file soal dan jawaban milik Cambridge. Ada beberapa mata pelajaran yang digunakan, yaitu Biologi, Matematika dan Ekonomi. Jenis soal yang digunakan juga ada beberapa macam, yaitu pilihan ganda dan esai. Hasil penelitian ini diharapkan bisa menjadi media pembelajaran.

Penelitian dilakukan dengan menggunakan sebanyak 100 pasang data soal dan ujian. Sistem akan menerima 2 inputan file dengan format PDF. Kedua file ini merupakan pasangan soal dan jawaban. Proses yang diakukan adalah file akan dirubah menjadi 2, yaitu file HTML dan file PNG. File HTML mengandung semua teks soal dan file PNG mengandung semua gambar dari soal. Sistem akan mengambil teks dan gambar dari masing-masing soal dan jawaban berdasrkan rule yang sudah ditentukan. Penentuan rule dilakukan secara manual dengan mempelajari pola-pola data yang tedapat dalam tag HTML. Setelah proses ekstraksi, soal dan jawaban ini dipasangkan sesuai dengan nomor urutnya masing-masing. Pasangan soal dan jawaban ini kemudian akan disimpan ke dalam database. Dari hasil penelitian, tingkat akurasi yang didapatkan adalah sekitar 46%. Kendala utama yang dihadapi adalah format soal dan jawaban yang tidak strandar sehingga menimbulkan kesulitan dalam proses ekstraksi informasi.

***Kata Kunci*—Information Extraction, berbasis rule, soal ujian, PDF.**

I. PENDAHULUAN

Pandemi virus corona (Covid-19) pada tahun 2020 sangat mengganggu aktivitas banyak orang. Banyak orang yang mengurangi aktivitas di luar rumah bila me-mungkinkan. Kantor-kantor ada yang menerapkan pola kerja WFH (Work From Home) untuk bagian/divisi yang tidak memerlukan kontak fisik langsung untuk melaksanakan pekerjaannya, kadang juga ada yang menerapkan masuk kantor bergantian untuk menghindari penyebaran virus.

Stefanus Nico Soenardjo, Teknologi Informasi Institut Sains dan Teknologi Terpadu Surabaya, Surabaya, Jawa Timur, Indonesia (e-mail: stefanusnicoid@gmail.com)

Gunawan, Teknik Informatika Institut Sains dan Teknologi Terpadu Surabaya, Surabaya, Jawa Timur, Indonesia (e-mail: gunawan@stts.edu)

Salah satu bagian kehidupan yang sangat terdampak oleh pandemi corona dalam hal ini adalah sektor pendidikan. Dalam keadaan normal, kegiatan pendidikan dilakukan dengan bertatap muka antara pengajar (guru) dengan para murid dalam ruangan kelas di sekolah. Dalam satu kelas ada beberapa murid, jumlah murid beragam, tergantung kebijakan dari masing-masing sekolah. Ditengah pandemi corona ini, pertemuan fisik antara guru dan murid ini tidak memungkinkan, karena berisiko besar untuk menyebarkan virus corona. Oleh karena itu kegiatan belajar dan mengajar dilakukan melalui jalan online. Beberapa aplikasi yang biasa digunakan seperti: Zoom, Micosoft Team dan Google Class.

Dengan pembelajaran online maka guru cenderung akan lebih banyak untuk memberikan tugas lebih banyak bila dibandingkan pembelajaran konvensional. Pemberian penjelasan mengenai suatu materi secara online mempunyai kekurangan bila dibandingkan dengan secara tatap muka langsung. Supaya murid lebih bisa menangkap materi maka diberi tugas yang porsinya lebih banyak, dengan harapan bila murid lebih banyak praktek maka akan lebih mengerti materi yang diberikan.

Bagian yang diharapkan bisa dibantu dalam penelitian ini adalah mengenai pemberian latihan atau ujian online. Di internet ada banyak soal ujian yang disediakan oleh beberapa institusi pendidikan. Dataset yang digunakan dalam penelitian adalah milik soal ujian milik Cambridge International Examinations. Dataset ini terdiri dari pasangan file PDF soal dan jawaban. Penelitian ini berusaha untuk melakukan pendekatan berbasis rule untuk mengekstrak soal serta jawaban kemudian berusaha untuk memasangkan antara soal dengan jawabannya. Pasangan antara soal dan jawaban ini kemudian disimpan dalam suatu bentuk database dan kemudian bisa dimanfaatkan lebih lanjut untuk kegiatan pembelajaran.

Lebih lanjut, pada penelitian ini penekanan dilakukan pada pengambilan atau ekstraksi informasi dari soal-soal ujian dalam bentuk pdf. Data soal yang digunakan milik Cambridge International Examinations, yang berupa pasangan antara file Question Paper (QP) dan file Mark Scheme (MS). File QP berisi soal-soal ujian, sedangkan file MS berisi jawaban. Di dalam file QP terdapat beberapa macam jenis soal, seperti esai, pilihan ganda, ada yang memakai gambar dan sebagainya. Pendekatan yang akan dilakukan untuk melakukan Information Extraction (IE) dalam penelitian ini adalah dengan meggunakan sistem berbasis rule. Penentuan rule akan melihat kondisi dari struktur dokumen yang terdapat pada file soal dan jawaban. Hasil ekstraksi yang berupa pasangan antara soal dan

jawaban ini diharapkan bisa dijadikan sebagai bahan pembelajaran.

II. TINJAUAN PUSTAKA

Menurut Wikipedia, Information Extraction (IE) merupakan proses pengambilan informasi dari dokumen yang tidak terstruktur atau semi terstruktur menjadi informasi yang terstruktur. IE [1]–[3] sendiri merupakan salah 1 bentuk dari Natural Language Processing (NLP). Media yang menjadi sumber dari IE bisa berasal dari berbagai macam sumber, seperti: gambar, video, audio, maupun teks.

IE berbasis rule [4], [5] menggunakan satu set atau beberapa aturan dalam proses pengambilan informasi. Proses penentuan aturan ini denngan memperhatikan bagaimana kondisi dalam dokumen, terutama ciri-ciri yang terdapat dalam teks yang menjadi tujuan utama dalam pengambilan informasi. Metode pendekatan yang dilakukan untuk proses IE dibagi menjadi 2 dimensi , yaitu:

1. Hand-coded atau Learning-based

Pada sistem berbasis Hand-coded, diperlukan manusia untuk mendefinisikan rule atau regular expression untuk melakukan proses IE. Untuk bisa menghasilkan rule yang baik, maka diperlukan orang yang mengerti mengenai domain linguistic dan pemrograman untuk bisa menghasilkan rule yang baik. Pada sisi lain, untuk Learning-based diperlukan data yang sudah dilabeli secara manual sebagai bahan pelatihan. Tentu data yang digunakan untuk proses pelatihan harus valid, karena mempengaruhi hasil IE yang akan dilakukan oleh sistem. Pemilihan antara kedua pendekatan ini dipengaruhi oleh kondisi informasi yang akan digunakan, baik dari segi struktur maupun noise-noise yang terdapat dalam data.

2. Rule-based atau Statistical

Rule based menggunakan suatu set aturan yang sudah di prediksi untuk melakukan IE, sedangkan statistical menggunakan suatu system pembobotan di mana bobot yang didapat menjadi dasar acuan untuk melakukan IE. Rule-based lebih mudah untuk diinterpetrasikan dan dikembangkan, sedangkan untuk statistical lebih baik dalam menangani noise dalam data yang tidak tersruktur. Penerapan Rule-based lebih baik untuk data yang terbatas dalam domain tertentu, sedangkan statistical lebih baik untuk domain umum seperti pencarian fakta dari transkrip pidato.

Rule yang digunakan bisa ditentukan sendiri sesuai dengan keperluan ataupun bisa dihasilkan dengan melakukan pelatihan oleh program dengan menggunakan metode-metode tertentu. Tentu saja kondisi data sangat berperan penting dalam pembentukan rule, ada kalanya data memerlukan semacam preprocessing terlebih dahulu. Pada penelitian ini proses pembentukan rule ditentukan sendiri dengan melihat kondisi soal ujian.

Dalam pembuatan penelitian ini dgunakan beberapa tinjauan pustaka untuk membantu dalam proses penelitian [6], [7]. Tinjauan pustaka yang digunakan tidak sama persis dengan penelitian ini, tetapi memiliki prinsip pendekatan yang mirip. Tinjauan Pustaka ini membantu dalam proses untuk menetapkan pendekatan yang mau digunakan dalam melakukan ekstraksi terhadap soal-soal ujian.

III. INFORMATION EXTRACTION BERBASIS RULE UNTUK SOAL UJIAN

A. *Analisa Permasalahan*

Masalah utama dalam penelitian ini adalah melakukan information extraction terhadap soal-soal ujian beserta dengan jawabannya. Inputan berupa pasangan file soal ujian dan jawaban dengan format PDF. Pendekatan yang dipilih dalam penelitian ini adalah information extraction berbasis rule. Pada Tabel 1 dapat dilihat keuntungan dan kekurangan dari penggunann information extraction yang berbasis rule dengan yang berbasis machine learning.

TABEL I
PERBANDINGAN RULE BASED DAN MACHINE LEARNING BASED

Rule Based	Machine Learning Based
Keuntungan 1. Deklaratif 2. Mudah dipahami 3. Mudah dimaintain 4. Mudah untuk memasukkan suatu domain pengetahuan 5. Mudah untuk melacak penyebab kesalahan dan memperbaikinya	Keuntungan 1. Dapat dilatih 2. Beradaptasi 3. Mengurangi kerja manual untuk mengadaptasi suatu domain penge- tahuan
Kekurangan 1. Heuristik 2. Membutuhkan kerja manual yang banyak	Kekurangan 1. Membutuhkan data berlabel 2. Membutuhkan pelatihan ulang 3. Membutuhkan keahlian machine learning untuk menggunakan atau maintain

Proses ektraksi akan di lakukan pertama kali pada file soal. Bentuk soal yang terdapat dalam file pdf ini ada beberapa macam, seperti soal esai, pilihan ganda dan sebagainya. Pada masing-masing jenis soal mempunyai tingkat kesulitan sendiri-sendiri. Hal yang paling menyulitkan adalah tidak adanya standarisasi mengenai format soal. Walaupun dalam satu mata pelajaran yang sama dan jenis soal yang sama, belum tentu format penulisan pada file soal dan jawabannya sama. Beberapa pendekatan yang dilakukan untuk mengatasi masalah ini:

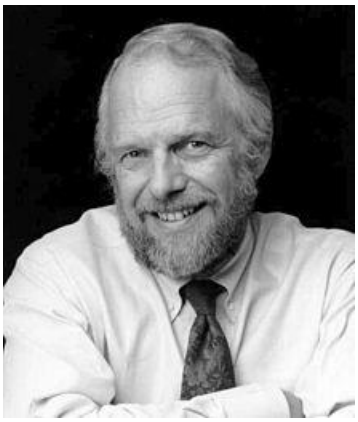
1. Membatasi jumlah mata pelajaran. Mata pelajaran yang digunakan dalam penelitian ini adalah Biologi, Matematika dan Ekonomi. Dengan pembatasan ini, maka setidaknya akan mengurangi banyaknya variasi pola soal dan jawaban yang harus dilakukan secara manual.
2. Mencari ciri format penulisan sebanyak mungkin, sehingga sistem dapat mengidentifikasi soal dan jawaban sebanyak mungkin. Hal ini penting karena mempengaruhi dari tingkat keberhasilan dalam melakukan information extraction.
3. Rule dibentuk hardcode dan statis dalam program, karena melihat kerumitan dalam parsing dokumen HTML. Bentuk rule akan mengikuti pola dari tag HTML hasil ekstraksi.

Untuk proses mengakses file PDF akan digunakan beberapa library tambahan. Pada fase ini file PDF akan diubah menjadi HTML sehingga lebih mudah untuk diproses. Dari file HTML akan dimulai proses untuk

mengidentifikasi soal dan jawaban. Pada proses ini, sistem akan berusaha untuk mencari pola dalam tag HTML untuk mencari soal dan jawaban. Soal dan jawaban mempunyai pola tag HTML sendiri-sendiri. Pencarian pola dalam hal ini mempunyai peranan dalam pembentukan rule, sebagai dasar dalam information extraction. Setelah berhasil mendapatkan soal dan jawaban maka langkah selanjutnya adalah memasangkan soal ke jawaban yang sesuai. Setelah mendapatkan pasangan soal dan jawaban, maka sistem akan menyimpan file tersebut ke dalam database sesuai dengan format yang telah ditentukan.

B. File PDF

PDF adalah kependekan untuk Portable Document Format, PDF merupakan format file yang dikembangkan oleh Adobe. Pada tahun 1991, salah satu pendiri Adobe Dr. John Warnock meluncurkan revolusi kertas-ke-digital dengan ide yang disebutnya, Camelot Project. Tujuannya adalah agar siapa pun dapat mengubah dokumen ke dalam bentuk digital, sehingga kemudian dapat mengirim nya ke mana pun dan mencetak kembali dokumen tersebut. Pada tahun 1992, Camelot telah berkembang menjadi PDF. Saat ini, format tersebut dipercaya oleh bisnis di seluruh dunia.



Gambar. 1. Dr. John Warnock

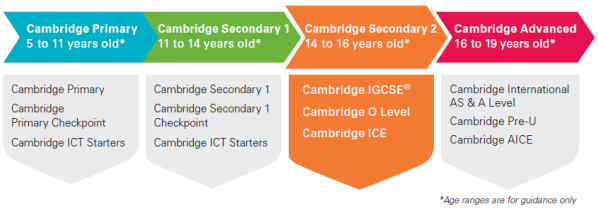
PDF mampu menampung berbagai format seperti teks, gambar, tabel dan sebagainya. PDF bisa dikirim secara digital seperti melalui email atau media komunikasi yang lain. Pihak penerima dapat membuka file tersebut dan mendapat tanpilan yang sesuai dengan kondisi sama dengan yang dibuat oleh pihak pengirim (mempertahankan jenis huruf (font), gambar, grafik serta tata letak yang tepat dari berkas aslinya). PDF juga menyediakan enkripsi dan tanda tangan digital.

C. Input dan Output Sistem

Dataset yang digunakan dalam penelitian adalah milik soal ujian milik Cambridge International Examinations, yang terdiri dari:

1. Ordinary Level (O)

O Level merupakan program sertifikasi yang dipelopori oleh University of Cambridge. Soal ujian ini untuk anak usia 14-16 tahun. Setara dengan IGCSE. Cambridge O Level menerima 620.000 pendaftaran setahun di lebih dari 50 negara di seluruh dunia.



Gambar. 2. Cambridge International Examinations

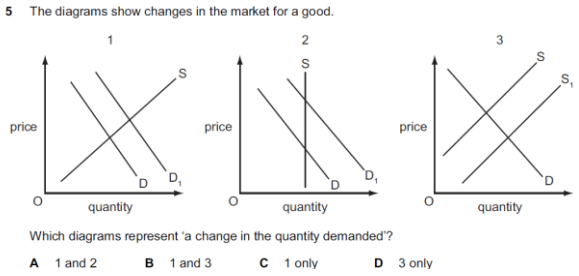
2. International General Certificate of Secondary Examinations (IGCSE)

IGCSE merupakan program sertifikasi yang dipelopori oleh University of Cambridge. Soal ujian ini untuk anak usia 14-16 tahun. Urutan jenjang program bisa dilihat pada Gambar3.2. IGCSE juga merupakan dasar yang ideal untuk program tingkat yang lebih tinggi seperti A Level. Setara dengan O Level. Daftar grup mata pelajaran IGCSE adalah Languages, Humanities and Social Sciences, Sciences, Mathematics, Creative and Professional.

3. Advance Level (A)

A Level merupakan program sertifikasi yang dipelopori oleh University of Cambridge, ditujukan untuk siswa berusia 16–19 tahun sebelum masuk ke jenjang universitas, Cambridge International AS & A Levels diikuti oleh lebih dari 175.000 siswa di lebih dari 125 negara setiap tahun. Daftar grup mata pelajaran A Level adalah English, Mathematics, Science, Languages, Humanities, Technology, Social Sciences, The Arts, General Studies.

Untuk setiap level mempunyai ujian dalam berbagai mata pelajaran, daftar bisa dlihat di Lampiran A. Data yang digunakan berupa sepasang file pdf yang terdiri dari Question Paper (QP) dan Mark Scheme (MS). Contoh pasangan antara QP dan MS dapat dilihat pada Gambar 3.1.



Gambar. 3. Contoh Question Paper

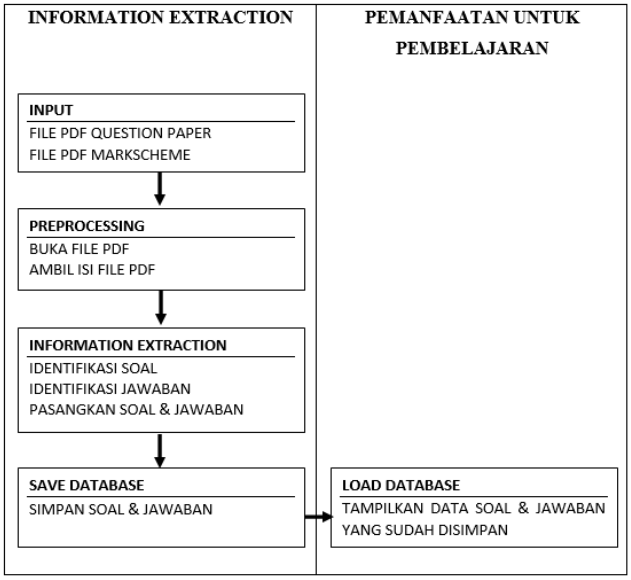
Pada Gambar 3 dan 4 bisa dilihat contoh Question Paper dan Mark Scheme dengan mata pelajaran ekonomi. Jenis soal pada gambar ini adalah soal pilihan ganda.

Page 2	Mark Scheme: Teachers' version GCE AS/A LEVEL – May/June 2011		Syllabu 9708
Question Number	Key	Question Number	Key
1	D	16	D
2	D	17	C
3	A	18	B
4	C	19	B
5	D	20	C
6	C	21	A
7	C	22	D
8	B	23	A

Gambar. 4. Contoh Question Paper

D. Arsitektur Sistem

Bagian utama terletak pada Information Extraction di mana sistem akan menerima inputan berupa dua buah file pdf yaitu soal dan jawaban. Yang pertama dilakukan sistem adalah mengambil soal dan jawaban, kemudian memasangkan antara soal dengan jawaban yang terkait. Setelah berhasil medapatkan keduanya, maka data tersebur kemudian akan disimpan kedalam database

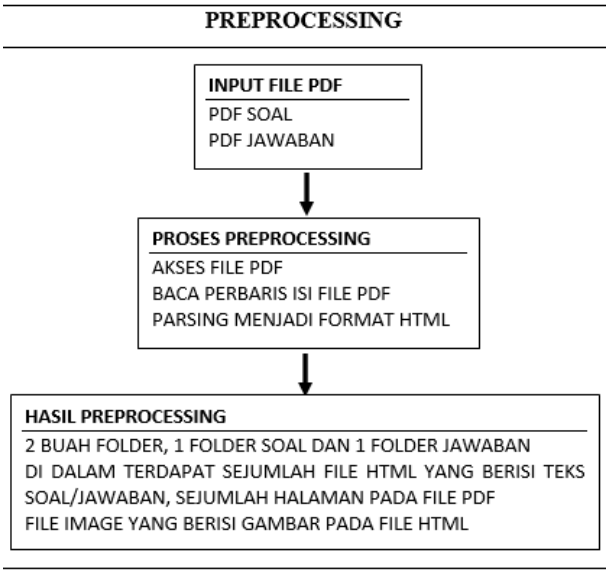


Gambar. 5. Arsitektur Sistem

Fase pemanfaatan untuk pembelajaran merupakan fase menampikan data soal dan jawaban yang sudah disimpan di dalam database. Pasangan soal dan jawaban tersebut dapat digunakan untuk semacam latihan soal. Detail untuk fase penting bisa dilihat sebagai berikut:

1. Fase Preprocessing

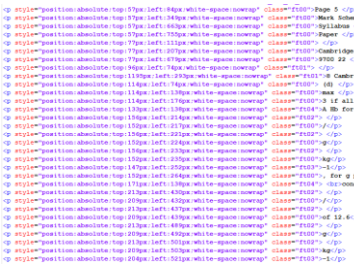
Pertama-tama pdf akan dirubah menjadi html menggunakan library PDF-To-HTML , Gufy , dan Proppler. Masing-masing halaman pdf akan dipisah menjadi 1 halaman html sendiri, dan 1 gambar background sendiri, di dalam folder output. Untuk detail proses preprocessing bisa dilihat pada Gambar 6.



Gambar. 6. Proses Preprocessing

2. Fase Pembentukan Rule

Pada penelitian ini, rule dihasilkan dengan mengamati pola yang terdapat pada file HTML hasil ekstraksi menggunakan library pdf-to-html. Semua informasi yang akan diambil terletak di dalam tag html. Yang memberatkan dalam fase ini adalah terdapat beraneka ragam pattern dalam soal dan jawaban. Oleh karena itu rule harus bisa sedapat mungkin untuk mengakomodasi dari semua pola yang terdapat dalam soal. Sebagai gambaran, bentuk isi file HTML bisa dilihat pada Gambar 7. Di mana bisa dilihat data yang mau diambil kebanyakan terletak di dalam tag HTML <P>.



Gambar. 7. Contoh Isi HTML

Cara kerja dari program adalah membaca masing-masing tag html <P>, dan membaca posisi masing-masing dom <P>. Secara umum yang dilakukan pertama kali adalah membaca bagian div yang terletak sebelah kiri dengan posisi paling banyak dan isinya berupa angka, supaya kita mempunya anchor point. Kemudian dari sana kita akan membagi soal dari anchor ke anchor. Masing-masing component html akan disimpan dalam database agar dapat dengan mudah ditampilkan lagi sesuai posisinya. Gambar background juga akan dipotong-potong sesuai dengan posisi dan ukuran masing-masing soal, supaya pada soal yang memiliki gambar, soal atau jawaban dapat menampilkan gambarnya.

Berikut ini adalah rule yang digunakan didalam program:

RULE CHECK ANCHOR

1. IF setelah cleanup RETURN angka THEN (cleanup = memecah br, dan ambil posisi masing-masing br ada yang isinya berupa angka)
2. IF angkanya urut THEN titik tersebut dikembalikan sebagai anchor.

Anchor yang dimaksud dalam hal ini adalah titik awal untuk pengambilan soal, yaitu nomor soal. Hal ini dilakukan dengan membaca isi didalam tag HTML jika ditemukan angka numerik di awal tag, maka anggap sebagai nomor hingga sampai ketemu nomor selanjutnya yang urut.

RULE PENGAMBILAN SOAL BERDASARKAN:

1. **CHECK** dari <p> ke <p> selanjutnya:
   IF batas kiri sama dengan anchor **THEN**
   jika YA ke point 2,
   ELSE ke point 4
2. IF batas atas/kiri masih limit dan angka yang ditentukan (menghindari pengambilan nomer halaman) **THEN**
   jika iya ke point 3
3. Masukkan komponen html dan text ke dalam array
4. **IF** content dalam batas kiri ada yang mengandung angka, dan sesuai urutan (jika sebelumnya 1, maka yang dicari 2) **THEN**
   jika YA ke point 5
5. Masukkan semua data array html ke dalam class tersendiri, Selesai pengecekan soal didapat jumlah soal

Gambar. 8. Contoh Isi HTML

Pada Gambar 8 akan menunjukan ilustrasi bagaimana mengambil soal. Bisa dilihat kalau bagian A, B dan C merupakan awal dari soal. Maka dalam hal ini digunakan Rule Check Anchor dan Rule Pemgambilan soal. Pada bagian A ditemukan data numerik yaitu angka 11, kemudian pembacaan perbaris diteruskan hingga sampai pada bagian B. Ternyata pada awal bagian B ditemukan angka 12 maka sudah memenuhi kriteria rule, maka awal bagian A sampai ke awal bagian B dianggap sebagai 1 soal tersendiri. Koordinat Top dan Left pada tag HTML yaitu (94,74) hingga (254,74) juga disimpan umtuk melakukan crop pada image background.

RULE tentukan anchor jawaban :

1. Cari batas kiri masing-masing jawaban dengan cara yang sama dengan soal
2. Simpan semua jumlah content yang mengandung angka di masing-masing batas kiri
3. **IF** dari masing-masing jumlah data di batas kiri ada yang sama dengan soal dan isi angkanya urut **THEN**
   batas kiri tersebut diambil sebagai patokan soal,
   ELSE loncat ke point 4
4. **IF** ada 2 kelompok batas kiri yang jumlahnya sama dengan jumlah soal, **THEN** 2 batas kiri tersebut akan dimerge (mengatasi terjadinya kunci jawaban yang batas kirinya variatif). jika tidak loncat ke point 5
5. **IF** angka semua angka pada batas kiri dalam range tertentu, **AND** dimerge jika jumlanya sesuai dengan jumlah soal **THEN** batas kiri tersebut dijadikan anchor, **RLSE** ke point 6
6. Dari halaman kedua ambil semua jawaban dengan pemisahan spasi cocokkan dengan jumlah soal (hal ini karena ada beberapa soal yang 2kolom)

untuk point 1-5 pengambilan jawaban sama dengan pengambilan soal dari anchor ke anchor lainnya

IV. UJICOBA

Dari total pengujians sebanyak 100 pasangan soal dan jawaban, yang diberhasil di ekstrak adalah sebanyak 46 soal. Maka bisa dilihat tingkat keberhasilan pengambilan soal adalah sebanyak 46%. Angka ini masih dibawah angka harapan pada hipotesa yaitu sebanyak 65%.

V. KESIMPULAN

Berdasarkan hasil penelitian ini, penulis mengambil kesimpulan dari penelitian information extraction berbasis rule untuk soal ujian, yaitu

1. Hal yang paling mempengaruhi dalam proses information extraction adalah format dari soal dan jawaban yang menjadi inputan. Tidak ada standar baku dalam penulisan soal dan jawaban. Dalam satu mata pelajaran yang sama pun, format yang digunakan bisa berbeda-beda sehingga sangat mengganggu proses information extraction.
2. Untuk tetap bisa mempertahankan bentuk soal, beberapa tag HTML yang berasal dari hasil proses preprocessing dimana inputan awal PDF diubah menjadi HTML, ikut disimpan ke dalam database, hal ini akan membantu dalam proses penampilan ulang. Bila hal ini tidak dilakukan maka akan sulit untuk menampilkan kembali bentuk awal dari soal, karena tidak ditemukan elemen lain penanda lokasi dan bentuk soal.
3. Program sudah berhasil untuk melakukan proses ekstraksi pada data berbentuk teks, gambar dan formula matematika. Namun pada waktu penyajian data masih sering terjadi pergeseran layout, terutama pada data yang mempunyai tabel. Hal ini merupakan salah satu kendala utama, yang dipengaruhi tidak adanya format yang baku dalam penulisan soal dan jawaban.
4. Pembentukan rule dilakukan secara manual dengan meneliti pola dari tag HTML hasil proses

preprocessing. Informasi yang ingin diambil berada didalam tag HTML ini. Pada penelitian ini rule yang terbentuk tidak dalam bentuk yang ideal, yaitu hardcode dan statis pada program. Ada 3 rule utama yaitu mencari anchor poin, yaitu titik start untuk pengambilan data (dengan mengidentifikasi nomor soal atau nomor jawaban), rule untuk pemgambilan soal dan rule untuk pengambilan jawaban.

5. Hasil ektraksi data pasangan soal dan jawaban bisa digunakan sebagai bahan pembelajaran. Pemanfaatan bisa disesuaikan sesuai dengan pihak yang menggunakan, baik pengajar maupun murid. Untuk pengajar bisa digunakan untuk menyusun materi dari ujian, untuk murid kumpulan soal dan jawaban bisa digunakan sebagai untuk latihan.
6. Hipotesis kurang terpenuhi yaitu hanya 46% dari angka harapan semula sekitar 65%. Tidak tercapainya angka harapan ini terutama dipengaruhi oleh format soal yang tidak standar.

DAFTAR PUSTAKA

[1] J. Tang, M. Hong, D. Zhang, B. Liang, and J. Li, “Information extraction: Methodologies and applications,” in *Emerging Technologies of Text Mining: Techniques and Applications*, 2007.

[2] S. Sarawagi, *Information Extraction*. Now Publishers, 2008.

[3] R. Gaizauskas and Y. Wilks, “Information extraction: Beyond document retrieval,” *J. Doc.*, vol. 54, no. 1, 1998, doi: 10.1108/EUM0000000007162.

[4] “A Rule-based Information Extraction System,” *Int. J. Innov. Technol. Explor. Eng.*, vol. 8, no. 9, 2019, doi: 10.35940/ijitee.i8156.078919.

[5] L. Chiticariu, Y. Li, and F. R. Reiss, “Rule-based information extraction is dead! Long live rule-based information extraction systems!,” 2013.

[6] Y. Lin, Z. Jun, M. Hongyan, Z. Zhongwei, and F. Zhanfang, “A method of extracting the semi-structured data implication rules,” in *Procedia Computer Science*, 2018, vol. 131, doi: 10.1016/j.procs.2018.04.315.

[7] N. Bhutani, Y. Suhara, W. C. Tan, A. Halevy, and H. V. Jagadish, “Open information extraction from question-answer pairs,” in *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, 2019, vol. 1, doi: 10.18653/v1/n19-1239.