# Bab 4
# Evaluasi dan Analisis

Bab ini berisi evaluasi terhadap hasil eksperimen dalam penelitian ini beserta analisis terhadap evaluasi tersebut. Evaluasi jawaban terhadap sistem tanya jawab dalam penelitian ini menggunakan sistem penilaian dari CLEF[60], dengan mencocokkan jawaban yang diperoleh dengan jawaban yang sudah disediakan untuk topik tanya jawab CLEF 2006 (kunci jawaban CLEF 2006).

Secara singkat sistem penilaian itu adalah sebagai berikut. Jawaban sistem tanya jawab mendapat nilai *R* (*Right*) bila jawaban dan dokumen pendukung tempatnya berasal sesuai dengan kunci jawaban (jawaban dan nama dokumen sama). Bila jawaban sesuai dengan kunci tetapi dokumen pendukungnya berbeda dengan jawaban yang ada maka jawaban tersebut mendapat nilai *U* (*Unsupported*). Selain itu jawaban mendapat nilai *W* (*Wrong*).

Evaluasi jawaban tidak dilakukan dengan pemeriksaan manual melainkan dilakukan secara otomatis oleh sistem tanya jawab. Oleh karena itu nilai *X* (*ineXact*) tidak diikutsertakan dalam evaluasi ini karena memerlukan justifikasi manual.

Kumpulan pertanyaan yang digunakan dalam penelitian ini diambil dari pertanyaan CLEF 2006 untuk subtopik tanya-jawab bilingual Indonesia-Inggris; berisi total 200 pertanyaan terdiri dari 150 pertanyaan fakta (*factoid questions*), 40 pertanyaan definisi (*definition questions*), dan 10 pertanyaan daftar (*list questions*). Sistem tanya jawab yang dikembangkan oleh penulis tidak didisain untuk menangani pertanyaan daftar, tetapi hanya pertanyaan fakta dan definisi.

## 4.1 Evaluasi Sistem Tanya Jawab

Evaluasi sistem tanya jawab ini akan dibagi menjadi dua bagian; yaitu evaluasi untuk pertanyaan fakta dan evaluasi untuk pertanyaan definisi.

60 Landasan teori mengenai evaluasi sistem tanya jawab diberikan di Bab 2, Subbab 2.9, "Evaluasi Sistem Tanya Jawab"

### 4.1.1 Evaluasi Jawaban Fakta

Pertanyaan fakta yang digunakan dalam penelitian ini berjumlah 150 dari total 200 pertanyaan. Sebagai rangkuman, pada Modul Ekstraksi Jawaban Entitas Bernama yang menangani pertanyaan fakta ini, jawaban akhir dipilih di antara para kandidat dengan 4 penilaian (yang dilambangkan dengan 4 huruf), yaitu:

1. **Skor *G***, dengan memberikan kueri berupa teks pertanyaan ke Google, nilai *G* ini didapatkan dengan menghitung banyaknya kemunculan setiap kandidat jawaban di 50 cuplikan *website* teratas.
2. **Skor *R***, merupakan nilai dokumen (juga nilai *passage*) ketika melakukan perolehan dokumen menggunakan Lemur Toolkit. Dokumen yang berada di urutan teratas mendapatkan nilai yang lebih baik dibandingkan dengan dokumen yang berada di bawah. Bila ada dua *passage* berasal dari satu dokumen yang sama maka mereka akan mendapatkan nilai *R* yang sama
3. **Skor *W***, merupakan *Average Distance Weight* (ADW), yaitu jarak rata-rata antara posisi kandidat jawaban dengan posisi setiap kata-kata dari teks pertanyaan yang juga ada di *passage*.
4. **Skor *T***, merupakan bobot yang dihitung dengan menggunakan TF-IDF kandidat jawaban pada 20 dokumen teratas dalam perolehan dokumen (1 dokumen untuk masing-masing dari 20 *passages* yang didapat dari Modul Perolehan *Passage*).

Empat nilai tersebut dilebur menjadi satu nilai dengan kombinasi linear aG + bR + cW + dT, di mana a, b, c, d adalah elemen dari {0, 1, 2}. Seperti yang telah dijelaskan dalam bab sebelumnya[61], terdapat 81 kombinasi rumus penilaian jawaban untuk diujicobakan. Lima hasil terbaik di antaranya ditampilkan pada evaluasi di Tabel 4.1. Evaluasi selengkapnya disertakan dalam Lampiran B.

**Tabel 4.1 Lima Hasil Terbaik Untuk Pertanyaan Fakta**

| No. | Rumus | Right | Unsupported | Wrong |
|---|---|---|---|---|
| 1. | 2G + 1R + 1W + 0T | 23 | 21 | 106 |
| 2. | 2G + 1R + 1W + 1T | 23 | 20 | 107 |
| 3. | 2G + 2R + 1W + 0T | 22 | 21 | 107 |
| 4. | 2G + 2R + 2W + 1T | 22 | 19 | 109 |
| 5. | 2G + 2R + 2W + 0T | 22 | 19 | 109 |

61 Bab 3, Subbab 3.6, "Nilai Akhir Kandidat Jawaban"

Hasil terbaik diperoleh oleh versi sistem dengan pemilihan kandidat jawaban menggunakan rumus 2G + 1R + 1W + 0T dengan perolehan 23 jawaban benar (*Right*) dan 21 *Unsupported*.

Melalui pemeriksaan manual (seperti evaluasi *lenient* yang digunakan oleh [Kim 2000]) ditemukan bahwa di antara 21 jawaban *Unsupported* tersebut terdapat jawaban yang sebetulnya benar dan didukung oleh cuplikan dokumennya namun dianggap salah karena berbeda dokumen dengan kunci jawaban dari CLEF. Berikut adalah dua contoh jawaban yang sebetulnya benar tetapi harus dievaluasi sebagai jawaban *Unsupported* (bagian cuplikan dokumen yang digarisbawahi menunjukkan informasi yang mendukung jawaban, dan teks seperti "*GH950518-000166*" atau "*LA050194-0034*" adalah nama dokumennya), ditampilkan pada Gambar 4.1 dan 4.2 berikut ini.

**Pertanyaan**: Siapa nama wanita pertama yang mendaki Gunung Everest tanpa masker oksigen?

**Jawaban**: Alison Hargreaves

**Cuplikan Dokumen Sistem Tanya Jawab**: (GH950518-000166)

WITH love from the top of the world, British climber Alison Hargreaves yesterday sent a message to her family. Still savouring her triumph at becoming the first woman to climb Everest unaided and without oxygen, she said in a fax to her children Tom, six, Katie, four, and husband Jim Ballard, at Spean Bridge, Inverness-shire: "Hi, team: "I am now at Base Camp -- very tired -- but happy after climbing the world's highest mountain. "We have problems with our satellite tel/fax -- but this has obviously got through to you! "Tomorrow a jeep

**Cuplikan Dokumen Kunci Jawaban**: (GH950516-000155)

Alison Hargreaves, 32, mother of two young children and one of the few professional mountaineers in Britain, has just climbed her way into mountaineering history by becoming only the second solo climber, and the first woman, to have made the climb up the north ridge of the world's highest mountain with no artificial oxygen or sherpas to carry her gear

**Gambar 4.1 Contoh Jawaban *Right* Yang Dievaluasi Sebagai *Unsupported* (1)**

**Pertanyaan**: Pada tahun berapakah terjadinya bencana di Chernobyl?

**Jawaban**: 1986

**Cuplikan Dokumen Sistem Tanya Jawab**: (LA050194-0034)

and Amy Fisher: Evidence that nearly 10,000 people died in the 1986 Chernobyl disaster; that untapped oil reserves in Somalia motivated the U.S. intervention; that the much-hyped DARE program has had little effect in discouraging drug use among students; that the United Nations Children's Fund reported 9 out of 10 young people murdered in industrialized countries are slain in the U.S. Required reading for broadcasters, journalists and well-informed citizens

**Cuplikan Dokumen Kunci Jawaban**: (LA042994-0048)

The Chernobyl nuclear-power plant is in Ukraine, but the reactor that exploded during the night of April 26, 1986, is only 10 miles from the Belarusian border

**Gambar 4.2 Contoh Jawaban *Right* Yang Dievaluasi Sebagai *Unsupported* (2)**

Bila jawaban-jawaban seperti di atas dianggap benar, maka evaluasi yang didapat dari rumus 2G + 1R + 1W + 0T adalah 40 *Right* dan 4 *Unsupported*. Berdasarkan evaluasi tersebut maka akurasi (prosentase jawaban *Right* terhadap keseluruhan pertanyaan) untuk pertanyaan fakta yang diperoleh dalam penelitian ini adalah 26,67% (benar 40 dari 150).

## 4.1.2 Evaluasi Jawaban Definisi

Pertanyaan definisi yang digunakan dalam penelitian ini berjumlah 40 pertanyaan dari total 200 pertanyaan. Jawaban definisi didapat dari frasa atau kata majemuk yang mengandung keterangan dari sebuah nama orang, istilah, atau nama penting lainnya. Misalnya seperti yang sudah dijelaskan pada Subbab 3.7 pada bab sebelumnya, pertanyaan "*Siapakah Guglielmo Marconi?*" memperoleh jawaban "*Italian electrical engineer*" dengan mengekstraknya dari *passage* "*Guglielmo Marconi, Italian electrical engineer who invented wireless telegraphy and ...*".

Dengan menggunakan metode evaluasi yang sama dengan pertanyaan fakta, hasil yang diperoleh adalah 5 *Right*, 1 *Unsupported*, dan 34 *Wrong*.

Pemeriksaan manual lebih lanjut (evaluasi *lenient*) menemukan tambahan 2 jawaban yang benar. Situasinya sama seperti pertanyaan fakta, yaitu ada jawaban yang

sebetulnya benar tetapi harus dievaluasi sebagai *Unsupported* atau *Wrong* karena tidak sesuai dengan kunci jawaban dari CLEF. Dua jawaban yang dimaksud ditunjukkan pada Tabel 4.2. Bagian cuplikan dokumen yang digarisbawahi menunjukkan frasa atau kata majemuk di mana jawaban definisi diambil.

**Tabel 4.2 Dua Jawaban Definisi yang Sebetulnya Benar**

| No. | Pertanyaan | Jawaban Sistem Tanya Jawab | Cuplikan Dokumen |
|---|---|---|---|
| 1. | Apakah Euro-Disney itu? | theme park outside Paris | (LA083094-0242) * Theme parks: Shares in Euro Disney SCA, the operator of the Euro Disney theme park outside Paris, plunged once again Monday |
| 2. | Siapakah Nick Leeson itu? | FORMER Barings trader | (GH951230-000065) FORMER Barings trader Nick Leeson will not pursue an appeal against a six year jail sentence for cheating Singapore's financial futures exchange. Leeson |

Jawaban nomor 1 pada Tabel 4.2 tersebut sebelumnya dievaluasi sebagai *Unsupported* karena menurut kunci jawaban CLEF cuplikan dokumennya berasal dari dokumen "*LA010194-0041*" sementara sistem tanya jawab mengekstrak jawaban tersebut dari dokumen "*LA083094-0242*".

Sementara itu jawaban nomor 2 pada tabel yang sama (Tabel 4.2) sebelumnya dievaluasi sebagai *Wrong* karena menurut kunci jawaban CLEF jawabannya seharusnya adalah "*Nick Leeson, the 28-year-old trader who brought down Barings, Britain's oldest merchant bank, last February.*"

Bila dua jawaban tersebut dianggap benar maka hasil evaluasi untuk pertanyaan definisi adalah 7 *Right*.

Dengan menggunakan evaluasi *lenient* maka total nilai untuk jawaban fakta dan definisi adalah 47 *Right* (40 + 7 *Right*) dan 4 *Unsupported*. Sementara itu dengan menggunakan evaluasi yang patuh (*strict*) dengan kunci jawaban CLEF total nilai yang didapat adalah 28 *Right* (23 + 5) dan 22 *Unsupported* (21 + 1)

Hasil evaluasi ini dirangkum dalam Tabel 4.3 berikut ini (kolom Akurasi menghitung prosentase jawaban *Right* terhadap total 200 pertanyaan). Seperti konsep evaluasi pada [Kim 2000], jenis evaluasi *strict* mengikuti/patuh terhadap kunci jawaban

CLEF, sementara evaluasi *lenient* lebih toleran dengan menganggap jawaban *Unsupported* juga sebagai jawaban *Right*.

**Tabel 4.3 Hasil Evaluasi Sistem Tanya Jawab**

| Jenis Evaluasi | Jawaban Fakta | Jawaban Definisi | Total | Akurasi (%) |
|---|---|---|---|---|
| *Lenient* | 40 R + 4 U | 7 R | 47 R + 4 U | 23,50 |
| *Strict* | 23 R + 21 U | 5 R + 1 U | 28 R + 22 U | 14,00 |

# 4.2 Mengapa Skornya Begitu Rendah?

Melihat hasil evaluasi sistem tanya jawab ini, maka pertanyaan selanjutnya adalah mengapa hasilnya hanya mencapai angka-angka tersebut? Menurut evaluasi *lenient*, sistem tanya jawab ini hanya mampu menjawab benar 23,50% dari semua pertanyaan. Dengan evaluasi *strict* justru lebih sedikit lagi, hanya 14,00%.

Sebagai perbandingan, Subbab 4.5 "Perbandingan Dengan Peserta CLEF" menunjukkan bahwa pencapaian sistem tanya jawab ini tidak begitu jauh dari rata-rata pencapaian penelitian serupa lainnya.

Sekarang mari kita analisis dua penyebab umum yang berkontribusi terhadap rendahnya akurasi yang dicapai oleh sistem tanya jawab ini.

## 4.2.1 Bagian Analisis Pertanyaan

Modul Analisis Pertanyaan bertugas untuk mengenali kategori jawaban yang diperlukan dari sebuah pertanyaan. Namun ternyata tidak semua pertanyaan berhasil diklasifikasikan dengan baik. Ada pertanyaan yang tidak bisa diklasifikasikan sama sekali kategorinya, dan ada yang berhasil dikenali tetapi salah.

Suatu pertanyaan tidak bisa diklasifikasi karena beberapa alasan; kategori pertanyaannya tidak termasuk yang bisa ditangani oleh sistem tanya jawab ini (seperti yang dijabarkan di Lampiran A) sehingga tidak terjaring oleh automata (pola-pola kalimat dalam *regular expression*) yang telah dibuat sebelumnya, atau memang automatanya tidak cukup baik untuk mengenali pertanyaan-pertanyaan yang kompleks. Berikut pada Tabel 4.4 adalah beberapa contoh pertanyaan yang tidak bisa diklasifikasikan (kolom No. menunjukkan nomor urut pertanyaan).

**Tabel 4.4 Pertanyaan yang Tidak Bisa Diklasifikasikan**

| No. | Pertanyaan | Kunci Jawaban CLEF |
|---|---|---|
| 9. | Simfoni manakah yang disusun oleh Beethoven pada 1824? | *Ninth Symphony* |
| 23. | Apa yang dituduhkan pada O.J. Simpson? | *murder* |
| 52. | Planet mana yang ditabrak oleh bintang berekor Pembuat-Sepatu-Levy? | *Jupiter* |
| 115. | Pada pertunjukan udara manakah sebuah F-86 Mk 6 jatuh pada 1993? | *El Toro Air Show* |
| 166. | Apakah jenis bom yang digunakan oleh IRA dalam serangannya di Heathrow pada 1994? | *mortar* |

Kelima pertanyaan tersebut, pada Tabel 4.4, memerlukan jawaban yang berasal dari kategori yang belum bisa ditangani oleh sistem tanya jawab ini.

Selain itu karena keterbatasan automata yang dibuat, ada pertanyaan-pertanyaan lain yang berhasil diklasifikasikan tetapi jatuh ke kategori yang salah. Berikut pada Tabel 4.5 adalah beberapa contoh pertanyaan yang dimaksud.

**Tabel 4.5 Pertanyaan yang Diklasifikasikan secara salah**

| No. | Pertanyaan | Kategori (oleh sistem tanya jawab) | Seharusnya |
|---|---|---|---|
| 37. | Berapa jumlah peneliti yang terlibat pada penelitian Top Quark selama tujuh belas tahun? | Ukuran – Waktu | Numerik |
| 72. | Di organisasi manakah Peter Anderson menjadi penasehat alkohol? | Lokasi | Organisasi |
| 162. | Pada film dari Kevin Reynolds yang manakah Kevin Costner berperan? | Lokasi | Judul Film |
| 194. | Siapakah yang menciptakan sistem operasi OS/2? | Orang | Organisasi – Perusahaan |

Statistik kesalahan-kesalahan analisis pertanyaan di atas dirangkum pada Tabel 4.6 berikut ini.

**Tabel 4.6 Kesalahan Pada Analisis Pertanyaan**

| Kesalahan | Fakta | Definisi | Daftar | Total |
|---|---|---|---|---|
| Tidak bisa diklasifikasikan | 25 | 1 | 0 | 26 |
| Salah kategori | 14 | 0 | 2 | 16 |
| **Total** | 39 | 1 | 2 | 42 |

Artinya, sebanyak 42 dari 200 pertanyaan tidak akan bisa dijawab dengan benar oleh sistem tanya jawab ini. Selain itu, karena pertanyaan bertipe daftar juga tidak bisa ditangani oleh sistem tanya jawab ini, maka seluruhnya ada 50 pertanyaan yang tidak bisa ditangani (terdapat 10 pertanyaan daftar, 2 di antaranya sudah tercantum di Tabel 4.6).

Itu artinya hanya tersisa 150 pertanyaan fakta dan definisi (111 fakta dan 49 definisi) yang bisa dijawab dengan benar oleh sistem tanya jawab ini.

### 4.2.2 Penerjemahan

Penerjemahan dengan mesin selalu menjadi permasalahan dalam sistem perolehan informasi lintas bahasa [Grossman 2004]. Modul Penerjemahan dalam sistem tanya jawab ini, yang menggunakan mesin penerjemah Indonesia-Inggris ToggleText, juga mengalami kendala dalam mendapatkan pertanyaan Bahasa Inggris terjemahan yang baik.

Berikut adalah beberapa contoh penerjemahan pertanyaan dalam sistem tanya jawab ini (Tabel 4.7).

**Tabel 4.7 Contoh Penerjemahan**

| No. | Pertanyaan Bahasa Indonesia | Pertanyaan Terjemahan |
|---|---|---|
| 9. | Simfoni manakah yang disusun oleh Beethoven pada 1824? | symphonies whatever that was compiled by Beethoven in 1824? |
| 89. | Siapakah pengarang yang menulis opera "Pelayan Wanita dari Pskov"? | Who the writer that wrote the opera of the "female Attendant from Pskov"? |
| 141. | Di kota manakah letak Taman Air Dunia Laut? | in what city the location of the Garden of Sea of world Water? |
| 161. | Berapa kali Zinedine Zidane memenangkan Pertandingan Amerika Terbuka? | How Many Zinedine Zidane times won the American Match was open? |

Pada pertanyaan nomor 9, kata "*disusun*" yang artinya "*membuat lagu*" diterjemahkan menjadi "*compiled*", padahal kita sebenarnya membutuhkan kata "*composed*".

Penerjemahan menjadi masalah bila melibatkan judul, nama tempat, atau istilah-istilah. Seperti pada pertanyaan nomor 89, "*Pelayan Wanita dari Pskov*" yang diterjemahkan menjadi "*female Attendant from Pskov*" seharusnya diterjemahkan sebagai "*The Maid of Pskov*". Begitu juga dengan "*Taman Air Dunia Laut*" yang seharusnya menjadi "*Sea World aquatic park*", dan "*Pertandingan Amerika Terbuka*" yang seharusnya menjadi "*the US Open*".

Kata-kata penting yang salah diterjemahkan, dan selanjutnya digunakan sebagai kueri untuk melakukan perolehan dokumen (atau perolehan *passage*) tentunya menurunkan relevansi hasilnya. Meskipun pengaruh langsung penerjemahan ini sulit diukur secara langsung terhadap fenomena sedikitnya pertanyaan yang berhasil dijawab dengan baik, dapat dipastikan bahwa permasalahan penerjemahan ini turut berkontribusi.

Dua subbab berikutnya, "Analisis Jawaban Definisi" dan "Analisis Jawaban Fakta" berusaha menganalisis lebih dalam mengenai permasalahan-permasalahan yang berhasil diidentifikasi dalam sistem tanya jawab ini.

## 4.3 Analisis Jawaban Definisi

Seperti yang telah dijelaskan sebelumnya, sistem tanya jawab menggunakan teknik pembentukan *consituency tree* dari kalimat-kalimat pada dokumen untuk mendapatkan jawaban. Jawaban itu sendiri diambil dari frasa kata benda atau bentuk aposisi.

Ternyata kelemahan dari teknik ini adalah ketidakmampuannya untuk mengkonfirmasi apakah frasa atau kata majemuk yang dipilih memang benar-benar menjelaskan apa yang diminta oleh pertanyaan. Teknik ini tidak mengerti konteks kalimat. Tabel berikut (Tabel 4.8) menunjukkan beberapa contoh kesalahan yang dibuat oleh sistem tanya jawab ini. Bagian cuplikan dokumen yang digarisbawahi menunjukkan frasa atau kata majemuk di mana jawaban definisi diambil.

**Tabel 4.8 Contoh Kesalahan Jawaban untuk Pertanyaan Definisi**

| No. | Pertanyaan | Jawaban Sistem Tanya Jawab | Cuplikan Dokumen |
|---|---|---|---|
| 1. | Apakah Hubble itu? | team | consider definitive. "The Hubble constant has been fraught with controversy for some time," said John P. Huchra, a Harvard University astronomer on the Hubble team. "There are also other forces |
| 2. | Apakah ECU itu? | basket of currencies | Until now most countries have favoured ECU but Bonn is against that because of links to the ECU "basket of currencies", which has steadily devalued in recent years. The |
| 3. | Apakah APEC itu? | was established 1989 to promote trade and investment in the Pacific Basin, in response to a more exclusive East Asian zone | and small: 1. Asia Pacific Economic Cooperation (APEC): APEC was established 1989 to promote trade and investment in the Pacific Basin, in response to a more exclusive East Asian zone |
| 4. | Siapakah Picts itu? | the ancient | nothing to do with it," he said. New plan for the ancient Picts |

Contoh nomor 1 pada tabel tersebut (Tabel 4.8) memberikan jawaban salah "*team*" yang didapat dari frasa "*the Hubble team*". Sementara untuk contoh nomor 3 seharusnya jawaban yang diberikan adalah "*Asia Pacific Economic Cooperation*". Contoh-contoh kesalahan tersebut mengilustrasikan bahwa sistem tanya jawab tidak bisa membedakan mana informasi yang benar-benar menerangkan apa yang ditanyakan dan mana yang tidak.

# 4.4 Analisis Jawaban Fakta

Untuk membantu mengungkap alasan mengapa hanya sedikit pertanyaan fakta yang bisa dijawab dengan baik oleh sistem tanya jawab ini, subbab ini akan memberikan analisis terhadap hasil evaluasi jawaban fakta. Analisis lebih lanjut atas evaluasi jawaban fakta disajikan dalam empat bagian berikut ini.

### 4.4.1 Kesalahan Pengenalan Entitas Bernama

Kandidat jawaban untuk setiap pertanyaan fakta dihasilkan oleh Modul Anotasi Entitas Bernama[62], yang diperoleh dari kata-kata yang mempunyai entitas bernama pada *passage.* Bila terdapat kesalahan dalam mengenali entitas bernama dari kata-kata yang terdapat pada *passage*, maka kandidat-kandidat jawaban yang dihasilkan menjadi tidak relevan. Sebagai contoh kesalahan dalam pengenalan entitas bernama yaitu pada kata "*Jordan.*" Kata "*Jordan*" dikenali sebagai nama orang padahal konteks pada kalimat merujuk pada nama negara.

Untuk mengetahui ketepatan pengenalan entitas bernama, maka perlu dilakukan perbandingan antara jumlah anotasi entitas bernama yang salah dengan yang benar. Pada penelitian ini digunakan 10 pertanyaan fakta yang diambil secara acak dari 150 pertanyaan fakta. Dari setiap pertanyaan digunakan 20 *passage* yang telah mempunyai anotasi entitas bernama (setelah melalui Modul Anotasi Entitas Bernama). Pemeriksaan dilakukan secara manual di mana pengenalan entitas bernama dinyatakan benar bila kategori yang diberikan kepada entitas bernama sesuai dengan konteks kalimatnya (artinya "*Jakarta*" masuk ke kategori nama kota, bukan nama negara) dan semua kata penyusun entitas bernama dikenali secara lengkap.

Hasil pemeriksaan tersebut ditampilkan pada Tabel 4.9 berikut ini.

**Tabel 4.9 Evaluasi Pengenalan Entitas Bernama**

| Total Entitas | Salah | Benar |
|---|---|---|
| 1.835 | 309 (16,84%) | 1.526 (83,16%) |

Terdapat sejumlah dua ribu entitas bernama dari 200 *passage* tersebut, dan sekitar 83% mempunyai anotasi entitas bernama yang benar dan 16% yang salah. Berikut adalah contoh-contoh kesalahan pengenalan entitas bernama tersebut yang mengakibatkan kesalahan dalam perolehan jawaban.

Yang pertama adalah kesalahan pemberian entitas bernama. Contohnya pada kalimat yang terdapat pada *passage* berikut ini: "... *<Location>Chester</Location> David Tollakson, who had scaled <Numeric>six</Numeric> of ...*", seharusnya frase "*Chester David Tollakson*" dikenali sebagai entitas nama orang.

62 Dijelaskan di Bab 3, Subbab 3.5 "Anotasi Entitas Bernama".

Yang kedua, kesalahan pada pemberian anotasi Tanggal (DATE) yang dihasilkan oleh GATE[63] untuk kata-kata berikut ini yang terdapat pada passage: "*today*", "*tomorrow*", "*10 years ago*", "*autumn*", "*saturday*", dll. Walaupun pengenalan entitas bernama ini benar, namun kandidat-kandidat jawaban ini sangat tidak mencukupi untuk menjawab pertanyaan bertipe Tanggal. Kandidat jawaban yang diharapkan dari pertanyaan bertipe tanggal adalah entitas-entitas seperti: "*May 13 1994*" dan "*July 1995*".

Yang ketiga, kesalahan pemberian entitas bernama yang disebabkan oleh adanya urutan aturan pemberian entitas bernama. Misalnya pada potongan kalimat "*...broke off <Numeric>three</Numeric> weeks ago*", sebenarnya frase "*three weeks*" masuk ke dalam kategori Ukuran Waktu (*Measurement-Time*). Pada Modul Anotasi Entitas Bernama, kategori Numerik diproses terlebih dahulu, sehingga kata "*three*" telah dikenali lebih dahulu sebagai Entitas Bernama Numerik. Akibatnya kata "*three*" (pada "*three weeks ago*") tidak dapat dikenali sebagai Waktu.

Meskipun kesalahan pada pemberian anotasi entitas bernama tidak begitu besar, namun hasilnya dapat menyebabkan kesalahan dalan perolehan jawaban.

### 4.4.2 Meneliti Daftar Kandidat Jawaban

Berikutnya penulis mencoba melihat daftar kandidat jawaban dari setiap pertanyaan fakta yang dijustifikasi sebagai *Wrong*. Apakah sebenarnya jawaban yang benar muncul di antara kandidat jawaban, hanya saja tidak berada di urutan pertama?

Hasil terbaik untuk pertanyaan fakta, seperti yang disajikan pada Tabel 4.1 yang lalu, menghasilkan 106 jawaban *Wrong* dari total 150 pertanyaan fakta. Sebelumnya, Subbab 4.2 di bagian "Bagian Analisis Pertanyaan" telah menunjukkan bahwa terdapat 39 pertanyaan fakta yang sudah pasti tidak mungkin dapat dijawab dengan benar karena kesalahan di bagian analisis pertanyaan. Maka tersisa 67 pertanyaan fakta yang salah/*Wrong* untuk diperiksa daftar kandidat jawabannya.

Untuk setiap 67 pertanyaan tersebut, penulis mendata apakah ada jawaban benar yang muncul di urutan selain ke-1 di dalam daftar kandidat jawabannya. Tabel 4.10 berikut ini merangkum hasil tersebut. Kolom 'Jumlah' menampilkan jumlah pertanyaan yang memiliki kandidat jawaban benar—sesuai kunci jawaban CLEF—yang tidak

63 GATE, a General Architecture for Text Engineering (http://gate.ac.uk). Silahkan lihat kembali Subbab 3.5.

terletak di urutan pertama dalam daftar kandidat jawaban, melainkan jatuh di urutan 2 – 5, 6 – 10, 11 – 15, 16 – 20, dan 21 ke atas.

**Tabel 4.10 Kandidat Jawaban Di Urutan Non-Pertama**

| Urutan | Jumlah |
|---|---|
| 2 – 5 | 15 |
| 6 – 10 | 10 |
| 11 – 15 | 8 |
| 16 – 20 | 5 |
| > 20 | 0 |
| **2 – 20** | **26** |

Di Tabel 4.10 tersebut, keseluruhan jumlah pertanyaan yang memiliki jawaban benar di urutan ke-2 sampai dengan 20 adalah sebanyak 26 pertanyaan (di lima jenis rentang urutan sebelumnya ada pertanyaan-pertanyaan sama yang beririsan). Mulai dari urutan ke-21 sampai seterusnya justru tidak ada kandidat jawaban yang benar sama sekali.

Jadi kesimpulannya, terdapat 26 pertanyaan (dari 67 yang salah/*Wrong*) yang sebenarnya memiliki kans untuk dijawab dengan benar. Namun kenyataannya nilainya dikalahkan oleh kandidat jawaban yang salah. Inspeksi lebih lanjut terhadap kandidat-kandidat salah yang mendapatkan penilaian tinggi di 26 pertanyaan tersebut ini menunjukkan bahwa Skor *G* mereka jauh melampaui ketiga skor lainnya (Skor *R*, *W*, dan *T*). Artinya kandidat-kandidat jawaban yang salah muncul di banyak cuplikan *website* dari Google, sementara kandidat jawaban yang benar justru terbelakang.

Seperti yang disebutkan oleh [de Chalendar 2002], salah satu permasalahan dalam penggunaan bantuan *Web* untuk mencari jawaban di koleksi dokumen yang kuno (mereka juga menggunakan koleksi LA Times 1994 dan Glasgow Herald 1995) adalah tertutupnya informasi lama yang dibutuhkan oleh informasi-informasi lain yang lebih modern dan tentunya lebih banyak.

Bila 26 pertanyaan tersebut memiliki kans, mengapa sisanya—sebanyak 41 dari 67 pertanyaan—tidak memiliki satupun kandidat jawaban yang benar?

Satu hal yang dapat dipastikan kebenarannya adalah bahwa tidak ada kesalahan dalam pengenalan kandidat jawaban (di Modul Anotasi Entitas Bernama) karena

penulis telah memeriksa dan memastikan bahwa semua jawaban pertanyaan fakta—dari kunci jawaban CLEF—telah terdaftar di daftar *gazetteer*[64]. Artinya bila mereka muncul di *passage*, pasti mereka muncul di daftar kandidat jawaban.

Jadi karena mereka (kandidat jawaban yang benar) tidak muncul, berarti *passage* yang benar yang seharusnya membawa mereka juga tidak muncul di antara 20 *passage* dari Modul Perolehan *Passage*. Karena mesin perolehan informasi yang digunakan di modul tersebut, Lemur Toolkit, adalah sistem yang sangat kredibel, kesalahan pastilah terletak di kueri.

### 4.4.3 Analisis Formula *G*, *R*, *W*, dan *T*

Di bagian ini kita akan menganalisis rumus penilaian yang digunakan untuk memberi penilaian terhadap kandidat-kandidat jawaban.

Bagaimana kontribusi masing-masing skor terhadap persentase jawaban yang berhasil dijawab dengan benar? Untuk mengetahui peranan masing-masing dari nilai *G*, *R*, *W*, dan *T* dalam perolehan jawaban, perhatikan Tabel 4.11 berikut ini.

**Tabel 4.11 Peranan *G*, *R*, *W*, dan *T***

| No. | Rumus | Right | Unsupported | Wrong |
|---|---|---|---|---|
| 1. | 0G + 0R + 0W + 1T | 10 | 4 | 136 |
| 2. | 0G + 0R + 1W + 0T | 14 | 14 | 122 |
| 3. | 0G + 1R + 0W + 0T | 17 | 11 | 122 |
| 4. | 1G + 0R + 0W + 0T | 20 | 19 | 111 |

Tabel 4.11 tersebut menunjukkan 4 rumus di mana masing-masing hanya terdiri dari satu penilaian dari *G*, *R*, *W*, atau *T*. Artinya rumus nomor 1 hanya menggunakan perhitungan *T* saja (karena koefisien *G*, *R*, dan *W* bernilai 0), rumus nomor 2 hanya menggunakan perhitungan *W* saja, dan seterusnya. Tujuannya adalah untuk melihat kontribusi masing-masing dari *G*, *R*, *W*, dan *T* terhadap perolehan jawaban.

Rumus nomor 4 dari Tabel 4.11 menunjukkan bahwa nilai *G* (Google) memberikan jumlah jawaban *Right* dan *Unsupported* tertinggi di antara 4 rumus pada tabel tersebut. Artinya penggunaan Google dalam pemilihan kandidat jawaban memberikan kontribusi yang besar dalam perolehan jawaban.

64 Penjelasan tentang *gazetteer* dapat ditemukan kembali di Bab 2, Subbab 2.5, "Pengenalan Entitas Bernama".

Tabel 4.12 berikut ini (sama seperti Tabel 4.1 yang lalu) yang menunjukkan lima hasil terbaik untuk pertanyaan fakta menunjukkan satu hal yang menarik.

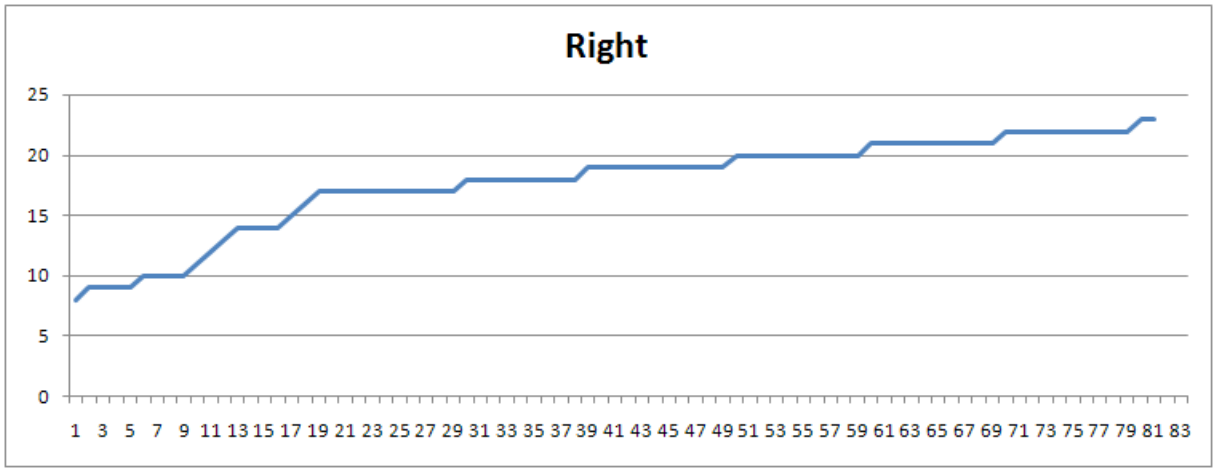
**Tabel 4.12 Lima Hasil Terbaik Untuk Pertanyaan Fakta**

| No. | Rumus | Right | Unsupported | Wrong |
|---|---|---|---|---|
| 1. | 2G + 1R + 1W + 0T | 23 | 21 | 106 |
| 2. | 2G + 1R + 1W + 1T | 23 | 20 | 107 |
| 3. | 2G + 2R + 1W + 0T | 22 | 21 | 107 |
| 4. | 2G + 2R + 2W + 1T | 22 | 19 | 109 |
| 5. | 2G + 2R + 2W + 0T | 22 | 19 | 109 |

Tabel 4.12 menunjukkan bahwa penggunaan TF-IDF sebagai salah satu kriteria pemilihan jawaban fakta hampir tidak berpengaruh dalam meningkatkan perolehan jawaban yang benar. Perhatikan bahwa penggunaan rumus nomor 1: 2G + 1R + 1W + 0T (di mana *T* adalah nilai TF-IDF kandidat jawaban) dan rumus nomor 2: 2G + 1R + 1W + 1T memberikan hasil yang hampir tidak berbeda, yaitu 23 jawaban *Right*. Kedua rumus tersebut menggunakan bobot (koefisien) yang sama untuk *G*, *R*, dan *W*, tetapi dengan bobot *T* yang berbeda. Fenomena ini semakin dikuatkan oleh bukti bahwa rumus 2G + 1R + 1W + 2T (dari Lampiran B), di mana bobot *T* dinaikkan menjadi 2, justru memberikan hasil yang lebih rendah yaitu 18 *Right* dan 20 *Unsupported*.

Ternyata penggunaan nilai *T* tidak diperlukan, atau dengan kata lain dapat disubstitusi secara bersama-sama oleh *G*, *R*, dan *W*. Tabel 4.12 tersebut juga menunjukkan bahwa kombinasi rumus terbaik untuk perolehan jawaban dalam penelitian ini adalah 2G + 1R + 1W.

Perhatikan Gambar 4.3 berikut ini. Grafik pada gambar tersebut menunjukkan jumlah jawaban *Right* dari ke-81 hasil evaluasi (dirangkum dari Lampiran B). Sumbu Y (vertikal) menunjukkan jumlah jawaban *Right* yang berhasil didapatkan, sedangkan sumbu X (horizontal) menunjukkan nomor hasil evaluasi dari 1 sampai dengan 81 diurutkan berdasarkan jumlah jawaban *Right*-nya dari yang paling sedikit ke yang paling banyak.




**Gambar 4.3 Grafik 81 Hasil Evaluasi**

Nilai rata-rata pencapaian ke-81 hasil evaluasi adalah sebesar 17,79 *Right*, 14,7 *Unsupported*, dan 117,51 *Wrong*. Angka-angka tersebut ternyata tidak terpaut jauh dari hasil terbaik dari semuanya (23 *Right*, 21 *Unsupported*).

Berikutnya, Tabel 4.13 di bawah ini menampilkan lima hasil terrendah di antara ke-81 hasil evaluasi tersebut.

**Tabel 4.13 Lima Hasil Terrendah Untuk Pertanyaan Fakta**

| No. | Rumus | Right | Unsupported | Wrong |
|---|---|---|---|---|
| 1. | 0G + 0R + 1W + 2T | 8 | 4 | 138 |
| 2. | 0G + 0R + 1W + 1T | 9 | 5 | 136 |
| 3. | 0G + 1R + 1R + 2T | 9 | 6 | 135 |
| 4. | 0G + 0R + 2R + 2T | 9 | 5 | 136 |
| 5. | 1G + 0R + 1W + 2T | 9 | 9 | 132 |

Kemiripan di antara lima hasil tersebut adalah hampir kecilnya Skor *G* dan *R*. Maka ini semakin menguatkan kesimpulan kita sebelumnya, bahwa Skor *G* memberikan kontribusi yang besar dalam perolehan jawaban yang benar. Sementara itu, Skor *R* memberikan kontribusi kedua terbesar, seperti yang telah ditunjukkan di Tabel 4.11 (Peranan *G*, *R*, *W*, dan *T*).

Selain itu lima hasil terrendah ini juga menguatkan kesimpulan kita yang lain, bahwa Skor *T* dapat dihilangkan sama sekali dari rumus ini, bahwa keberadaannya dapat disubstitusikan secara bersama-sama oleh Skor *G*, *R*, dan *W*. Hal ini dikarenakan lemahnya peranan Skor *T* dalam perolehan jawaban yang baik.

Dari keseluruhan 81 hasil evaluasi sistem tanya jawab ini, terdapat satu fenomena unik yang terjadi. Rumus 0G + 0R + 0W + 0T, di mana semua skor berkonstanta nol, memperoleh hasil evaluasi sebesar 18 *Right*, 11 *Unsupported*, dan 121 *Wrong*. Karena semua skor berkonstanta nol, maka dapat dikatakan tidak ada penilaian sama sekali terhadap kandidat-kandidat jawaban. Oleh karena itu sistem tanya jawab lalu memilih kandidat pertama di daftar sebagai jawaban akhir karena semua kandidat mendapatkan nilai yang sama, yaitu nol.

Dapat disimpulkan bahwa hal ini disebabkan oleh keberhasilan perolehan *passage*. Kandidat jawaban pertama dari *passage* pertama langsung menjadi jawaban yang benar. Meskipun penulis merasa skeptis—dikarenakan penerjemahan pertanyaan dari Bahasa Indonesia ke Inggris yang kurang baik sehingga menurunkan kualitas perolehan *passage*—ternyata ada beberapa kasus di mana perolehan *passage* mendapatkan hasil yang begitu baik.

### 4.4.4 Modifikasi Skor *T*

Pada bab sebelumnya telah disimpulkan bahwa Skor *T* dapat dihilangkan dari penghitungan skor. Bila komputasi Skor *T* dimodifikasi, apakah akan memberikan efek yang lebih baik atau lebih buruk pada perolehan jawaban? Bab ini berusaha menjawab pertanyaan-pertanyaan tersebut.

Sebelumnya Skor *T* dihitung dari nilai TF-IDF kandidat jawaban dengan menggunakan 20 dokumen dari 20 *passages*[65] yang digunakan pada perolehan penemuan kandidat jawaban. Selanjutnya nilai TF-IDF tersebut akan dihitung berdasarkan keseluruhan koleksi dokumen. Penilaian ini dilakukan karena pada koleksi dokumen, kata-kata yang hanya muncul di sejumlah kecil dokumen akan mendapat nilai TF-IDF yang lebih tinggi daripada kata-kata yang muncul di banyak dokumen, seperti *stopwords* [Grossman 2004]. Penelitian ini ingin mengetahui apakah kandidat jawaban berasal dari kata-kata yang bobotnya (TF-IDF) tinggi.

Bila perhitungan dari Tabel 4.11 (Peranan *G*, *R*, *W*, dan *T*) untuk skor 0G + 0R + 0W + 1T dilakukan sekali lagi, dengan Skor *T* berasal dari TF-IDF terhadap seluruh koleksi dokumen, evaluasi yang didapat adalah 10 *Right*, 5 *Unsupported*, dan 135 *Wrong*. Ternyata Skor *T* yang baru hanya berpengaruh sedikit terhadap perolehan jawaban, yaitu tambahan 1 buah *Unsupported*.

65 Dibahas di Bab 3, Subbab 3.6, "Pembobotan Kata dengan TF-IDF"

Untuk lebih memastikan apakah Skor *T* yang baru akan meningkatkan kinerja perolehan jawaban, maka perhitungan ke-81 kombinasi penilaian (seperti yang dilampirkan di Lampiran B) dilakukan lagi dengan menggunakan komputasi Skor *T* yang baru. Lima hasil terbaik dari perhitungan ulang ini ditampilkan dalam Tabel 4.14 berikut ini.

**Tabel 4.14 Lima Hasil Terbaik Untuk Pertanyaan Fakta**

| No. | Rumus | Right | Unsupported | Wrong |
|---|---|---|---|---|
| 1. | 2G + 1R + 1W + 1T | 21 | 20 | 109 |
| 2. | 1G + 1R + 2W + 1T | 21 | 20 | 109 |
| 3. | 2G + 2R + 1W+ 2T | 21 | 14 | 115 |
| 4. | 1G + 1R + 1W + 0T | 20 | 20 | 110 |
| 5. | 2G + 2R + 2W + 0T | 20 | 20 | 110 |

Ternyata dengan penggunaan perhitungan Skor *T* yang baru hasil terbaiknya lebih rendah dibandingkan dengan hasil terbaik dengan menggunakan Skor *T* yang lama (seperti yang tercantum di Tabel 4.12). Hasil terbaik dengan skor lama adalah 23 *Right* dan 21 *Unsupported* sedangkan hasil terbaik dengan skor *T* yang baru adalah 21 *Right* dan 20 *Unsupported*.

Ini artinya modifikasi Skor *T* dengan menghitung nilai TF-IDF dari seluruh koleksi tidak memberikan hasil yang lebih baik dibandingkan dengan menggunakan 20 dokumen. Maka Skor *T* yang digunakan di sistem tanya jawab ini adalah versi yang lama. Berikutnya hasil yang dicapai oleh sistem tanya jawab ini akan dibandingkan dengan hasil yang dicapai oleh sistem tanya jawab di CLEF 2006.

# 4.5 Perbandingan Dengan Peserta CLEF

Subbab ini mencoba membandingkan hasil yang diperoleh sistem tanya jawab dalam penelitian ini dengan hasil dari peserta-peserta topik tanya jawab bilingual CLEF 2006 yang menggunakan koleksi dokumen dalam Bahasa Inggris. Tujuannya adalah untuk melihat sejauh mana hasil yang dicapai oleh penulis dibandingkan dengan pencapaian penelitian-penelitian lain.

Pada CLEF 2006 terdapat total 13 sistem tanya jawab yang berpartisipasi (satu peserta bisa mengikutsertakan maksimal dua sistem tanya jawab). Evaluasi dari 8

sistem tanya jawab CLEF dengan hasil terbaik disajikan pada Tabel 4.15 berikut ini [Magnini 2006].

**Tabel 4.15 Evaluasi 7 Hasil Terbaik Topik Tanya Jawab CLEF**

| No. | Peserta | Akurasi Jawaban Fakta (%) | Akurasi Jawaban Definisi (%) | Akurasi Total (%) |
|---|---|---|---|---|
| 1. | UTJP (Polandia-Inggris) | 88,00 | 80,00 | 86,32 |
| 2. | LIR Group (Perancis-Inggris) | 26,00 | 22,50 | 25,26 |
| 3. | LIR Group (Perancis-Inggris) | 22,00 | 25,00 | 22,63 |
| 4. | University of Alicante (Spanyol-Inggris) | 19,33 | 22,50 | 20,00 |
| 5. | University of Limerick (Perancis-Inggris) | 21,33 | 10,00 | 18,95 |
| 6. | DFKI Lab (Jerman-Inggris) | 17,33 | 20,00 | 17,89 |
| 7. | University of Alicante (Spanyol-Inggris) | 12,00 | 27,50 | 15,26 |
| 8. | University of Alexandru Ioan Cuza (Romania-Inggris) | 15,33 | 5,00 | 13,16 |

Hasil tertinggi dicapai oleh peserta nomor 1 dengan subtopik Polandia-Inggris dengan perolehan 86,32% jawaban benar. Namun Magnini et. al. [Magnini 2006] juga menyatakan bahwa hasil untuk peserta ini masih dalam konfirmasi karena hasilnya jauh berbeda dengan peserta-peserta lain.

Sistem tanya jawab yang dikembangkan oleh penulis dalam penelitian ini akan berada di urutan ke-3 bila menggunakan evaluasi *lenient* (akurasi 23,50%) dan berada di urutan ke-8 bila menggunakan evaluasi *strict* terhadap kunci jawaban CLEF 2006 (akurasi 14,00%).

Karena semua peserta CLEF dievaluasi secara manual oleh penyelenggara CLEF, kemudian kunci jawabannya dirilis untuk publik, dan kunci jawaban tersebut digunakan untuk evaluasi *strict* dalam penelitian penulis ini, maka kita dapat menganggap bahwa semua peserta CLEF menggunakan evaluasi *strict*. Oleh karena itu yang lebih setara dibandingkan dengan peserta CLEF adalah hasil evaluasi *strict* dari sistem tanya jawab ini.

Dengan rata-rata akurasi ke-13 sistem tanya jawab pada CLEF 2006 sebesar 19,84% [Magnini 2006] maka sistem tanya jawab penulis dengan akurasi 14,00% masih berada di bawah nilai rata-rata tersebut.

Bab berikutnya, Bab 5 akan memberikan kesimpulan atas penelitian ini, saran-saran untuk pengembangan selanjutnya, dan penutup makalah ini.