KOM341
Temu Kembali Informasi

KULIAH #10
- Question Answering System

QA vs Search Engine

- Search engine atau IR
  - Query berbentuk keyword
  - Hasil jawaban berupa dokumen
- QAS
  - Query berbentuk bahasa alami (pertanyaan) → NLP
  - Hasil jawaban berupa entitas yang tepat
  - Kombinasi IR dan NLP

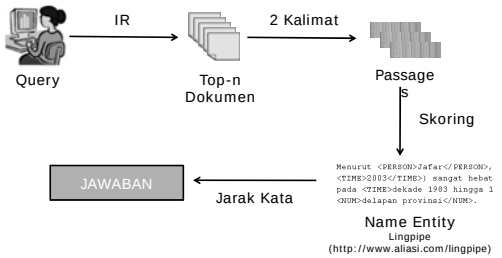
Jenis Pertanyaan

- Orang: pekerjaan, peranan, posisi
  - Siapa Susilo Bambang Yudhoyono?
- Organisasi: nama, kegiatannya dsb.
  - Apa itu UNICEF?
- Lokasi
  - Dimana ibukota Indonesia?
- Waktu dari suatu kejadian (date)
  - Kapan Pangeran Diponegoro wafat?
- Ukuran (kuantitas. jarak, lamanya - measure of distance)
  - Berapa jumlah penduduk Bogor?
- Cara suatu hal bisa terjadi (how)
  - Bagaimana Pangeran Diponegoro wafat?
- Pertanyaan bisa tidak mempunyai jawaban

Evaluasi

- Dilakukan oleh manusia
- Jawaban dinilai dari segi
  - Reponsiveness (pasangan jawaban-dokID)
  - Ketepatan (untuk setiap jawaban)
- Pemberian penilaian
  - Wrong (W) : jawaban tidak benar
  - Unsupported (U): jawaban benar tapi dokumen tidak mendukung
  - Inexact (X): jawaban dan dokumen benar tapi terlalu panjang
  - Right (R) : jawaban dan dokumen benar

Metode QAS, Li & Croft (2001)



Contoh Proses Memperoleh Jawaban

- Query :
  What is the capital of <LOC>Somalia</LOC>?
- Passage :
  Here there is no coordination. <PERSON>Steffan de Mistura</PERSON> – <ORG>UNICEF</ORG> representative in the Somali Capital, <LOC>Mogadishu</LOC> and head of the anti-cholera team – said far more <LOC>Mogadishu</LOC>, anti refugees are crowded together here without proper housing or sanitarian than during the <LOC>Somalia</LOC> crisis. And many are already sick and exhausted by the long trek from <LOC>Rwanda<LOC>.
- Jarak antara capital dengan Mogadishu = 1, Rwanda = 38.
- String yang diambil adalah yang mempunyai jarak yang terkecil dan yang paling banyak mempunyai kata sesuai query.

Pengembangan QAS

- Riloff & Thelen (2000), Ikhsani (2006), Romaida (2008)
- Anggraeni (2007)
- Ballesteros & Xiaoyan-Li (2007), Dewa Ayu Tenara Kardinia Cidhy (2009)

Riloff & Thelen (2000), Ikhsani (2006), Romaida (2008)

- Menemukan jawaban pada dokumen yang menggunakan bahasa baku
- Jawaban berupa kalimat pada dokumen yang memiliki nilai skor tertinggi.
- Rule-based Question Answering System
- Jenis pertanyaan: Apa, Siapa, Kapan, Mengapa, Mana

Tingkatan pemberian nilai

- clue : +3
- good_clue : +4
- confident : +6
- slam_dunk : +20

- WordMatch:
  kondisi dimana setiap token pada kalimat query sama dengan setiap token pada kalimat dokumen, dan diberi nilai CLUE

Rules: KAPAN?

Score(S) += WordMatch(Q,S)

If contains(S,{saat, ketika, kala, semenjak, sejak, waktu, setelah, sebelum, sesudah,selama, pada}) and contains(S, TIME) then
    Score(S) += good_clue

If contains(Q,TIME) Or containsS(S, TIME) then
    Score(S) += good_clue

If contains(S,{saat, ketika, kala, semenjak, sejak, waktu, setelah, sebelum}) then
    Score(S) +=clue

Rules: MANA?

Score(S) += WordMatch(Q,S)

If contains(S,LOCATION) and contains(S,{dalam, dari, pada}) then
    Score(S) += slam_dunk

If contains(S,{dalam, dari, pada}) then
    Score(S) += clue

If contains(S,LOCATION) then
    Score(S) += good_clue

Rules: MENGAPA?

Score(S) += WordMatch(Q,S)

If contains(S,{karena, sebab, akibat, maka, agar, supaya}) then
    Score(S) += clue

Rules: SIAPA?

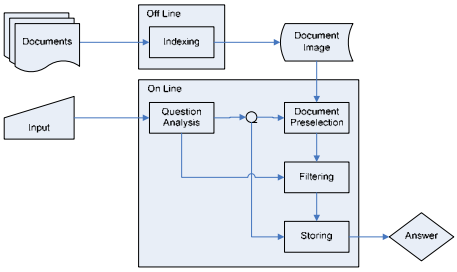
score(S) += WordMatch(Q,S)

If contains(Q,HUMAN) then
score(S) += confident

Ballesteros & Xiaoyan-Li (2007), Dewa Ayu Tenara K C (2009)

- QAS yang sesungguhnya
- Jawaban berupa entitas yang tepat
- Menggunakan metode Li & Croft (2001)
- Name entity tagging pada dokumen (Pangudi, 2009)
- Pemilihan passage menggunakan pembobotan heuristic
- Tipe pertanyaan : Siapa, Kapan, Dimana, Berapa

Metode



Ekstraksi jawaban

- Identifikasi named entity pada dokumen
- Dokumen dipartisi menjadi passages yang terdiri dari dua kalimat yang saling berdampingan
- Dilakukan pembobotan terhadap passages
  - Jumlah kata pada query yang terdapat pada passages
  - Apakah kata-kata yang cocok terdapat pada kalimat yang sama
  - Ukuran dari passages terbaik
  - Jarak kandidat jawaban dengan passages terbaik.

Pembobotan passages

- Hanya perhitungkan passages yang memiliki NE yang sesuai dengan query.
- Ambil count_m = jumlah kata pada query yang terdapat pada passages, dan count_q = jumlah kata pada query
- Jika jumlah kata yang cocok kurang dari threshold (t), score = 0, selain itu, score = count_m
- Cara menentukan threshold (t)
  - Jika jumlah kata pada query < 4, t = count_q
  - Jika antara 4-8, t = count_q/2.0+1.0
  - Jika >8, t = count_q/3.0+2.0.

Pembobotan passages

- Jika seluruh mathching words terletak pada satu kalimat
  - Sm = 1, selainnya Sm = 0
- Jika matching words terletak pada urutan yang sama
  - Ord = 1, selainnya Ord = 0
- W = the best matching window. Ukuran passages yang mengandung matching word terbanyak
  - score = score + (count_m / W)
- Heuristic Score
  - score = count_m + 0.5*Sm + 0.5*Ord + count_m/W

Contoh hasil QAS

No	Query	Jawaban	Keterangan
1	Siapa Asisten Sekretaris Daerah (Assekda) Bidang Kesejahteraan Rakyat Provinsi DIY?	Bambang Purnomo	*Right*
2	Siapa Bambang Purnomo?	Asisten Sekretaris Daerah Assekda Bidang Kesejahteraan Rakyat	*Right*
3	Siapa Juru Bicara Departemen Luar Negeri Republik Indonesia?	Marty Natalegawa	*Right*
4	Siapa Marty Natalegawa?	Departemen Luar Negeri Republik Indonesia	*Wrong*
5	Siapa menteri pertanian?	Bungaran Saragih	*Right*
6	Siapa Bungaran Saragih?	null	-
7	Berapa harga beras dalam negri antara bulan Juni-Juli?	Rp 4000	*Unsupported*
8	Berapa luas areal sagu malaysia?	51.3%,1 128 juta ha, 2 201 juta ha	*Wrong*

Open Ephyra

- Ephyra is a modular and extensible framework for question answering.
- OpenEphyra is the freely-available, open-source version of the Ephyra QA system
- Four stages: question analysis, query generation, search and answer extraction, and selection.
- http://www.cs.cmu.edu/~nico/ephyra
- Based on Java
- Indexer:
  - Google knowledge miner
  - Indri knoledge miner (http://sourceforge.net/projects/lemur/)

Open Ephyra Architectures