

Joost at CLEF 2007

Gosse Bouma, Geert Kloosterman, Jori Mur, Gertjan van Noord, Lonneke van der Plas, Jörg Tiedemann

Information Science
University of Groningen

Imix 13/09/07

Outline

- 1 Question Answering Task
- 2 Innovations
 - Wikipedia
 - Query Expansion
 - Anaphora Resolution
 - Multilingual QA
- 3 Results and Future Work

Question Answering

Who is the murderer of John Lennon?

10 may - 1955 – **Mark David Chapman**, murderer of John Lennon

How often was he hit?

he → John Lennon

Lennon was hit **four times** and died at 11.15 pm.

Where was he murdered?

he → John Lennon

John Lennon **On All Music Guide**

Question Answering

Who is the murderer of John Lennon?

10 may - 1955 – **Mark David Chapman**, murderer of John Lennon

How often was he hit?

he → John Lennon

Lennon was hit **four times** and died at 11.15 pm.

Where was he murdered?

he → John Lennon

John Lennon **On All Music Guide**

Question Answering

Who is the murderer of John Lennon?

10 may - 1955 – **Mark David Chapman**, murderer of John Lennon

How often was he hit?

he → John Lennon

Lennon was hit **four times** and died at 11.15 pm.

Where was he murdered?

he → John Lennon

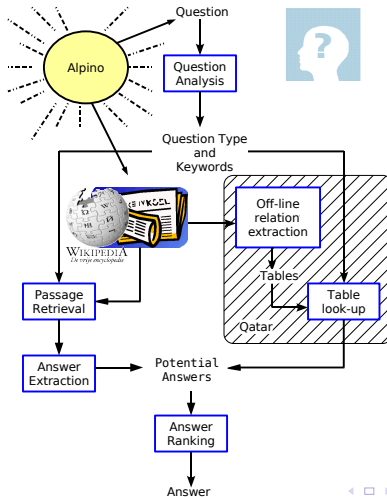
John Lennon **On All Music Guide**

QA at CLEF 2007

Question types

Factoid	Waar wordt Ndonga gesproken?
Definition	Wat is het Semantisch Web?
List	Welke daimyo's vormden het Oostelijk leger bij de slag bij Sekigahara
Temporally restricted	Hoeveel mobiele telefoons waren er in Nederland in gebruik in juni 1995?

- Monolingual and multilingual QA
- **New** in 2007 Task
 - **Wikipedia** added to document collection
 - **Follow-up Questions**



Wikipedia

The screenshot shows a Mozilla Firefox browser window displaying the Dutch Wikipedia article for Albert Plesman. The browser's address bar shows the URL 'del.icio.us'. The page title is 'Albert Plesman - Wikipedia - Mozilla Firefox'. The browser's menu bar includes 'File', 'Edit', 'View', 'History', 'Bookmarks', 'Tools', and 'Help'. The page content includes the Wikipedia logo, a navigation sidebar with links like 'Hoofdpagina', 'Artikelindex', and 'Categorieën', and the main article text. The article text describes Plesman's life, his role in the KLM, and his involvement in the aviation industry during the World Wars. A bust of Albert Plesman is shown on the right side of the page.

Albert Plesman - Wikipedia - Mozilla Firefox

File Edit View History Bookmarks Tools Help del.icio.us

Aanmelden en inschrijven

artikel overleg bewerk geschiedenis

Met uw steun houden wij Wikipedia online!

Albert Plesman

Albert Plesman (Den Haag, 7 september 1889 – aldaar, 31 december 1953) was een Nederlands luchtvaartpionier en medeoprichter van de KLM.

Hij werd geboren als zoon van een eierhandelaar uit Den Haag. In 1915 was hij gelegerd in Soesterberg alwaar hij als beroepsofficier bij de gemobiliseerde Nederlandse luchtmacht, toen de 'militaire luchtvaartafdeling' genoemd, in 1918 zijn militaire vliegbrevet behaalde. Na de Eerste Wereldoorlog mede-organisator van de ELTA, de Eerste Luchtvaart Tentoonstelling Amsterdam, dat van 1 augustus tot 15 september 1919 gehouden werd en waar maar liefst 800.000 bezoekers op af kwamen. Voor deze gelegenheid waren expositiehallen gebouwd, welke na het evenement in gebruik werden genomen door Anthony Fokker, voor zijn nieuw op te richten Fokker vliegtuigfabriek te Amsterdam-noord.

Al deze activiteiten leidden tevens tot de oprichting van de KLM, waarvan Plesman directeur werd, en wat hij tot een bloeiend bedrijf maakte. Na de Tweede Wereldoorlog werd Plesman benoemd tot president-directeur van de KLM. Na het herstel van de door de oorlog geleden schade werd het bedrijf onder zijn leiding een luchtvaartmaatschappij van grote allure. Plesman pleitte tevergeefs voor verplaatsing van Schiphol naar een locatie bij Burgerven.

Voor zijn grote verdiensten voor de internationale burgerluchtvaart ontving hij in 1959 postuum de eerste Edward Warner-medaille.

Albert Plesman heeft ook de naam Randstad bedacht. Volgens overlevering gebeurde dat toen hij op zoek was naar een locatie voor een nationale luchthaven. Vliegend boven het westen van Nederland zag hij 'een ring van steden aan de rand van een grote groene ruimte'. De ring van steden noemde hij later de Randstad, en het groen daartussenin werd bekend als het Groene Hart.



Buste van Albert Plesman op Schiphol

Wikipedia formats

- HTML
- Internal Wikipedia-format
- XML (provided by University of Amsterdam)

```
<head>
  <title>Albert Plesman</title>
  <meta name="wx_pagename" content="Albert_Plesman"/>
</head>
<body>
  <div id="wx_article">
    <wx:section level="1" title="Albert Plesman">
      <h1 class="pagetitle">Albert Plesman</h1>
      <p><b>Albert Plesman</b>
      (<a href="/wiki/Den_Haag">Den Haag</a>,
      . . . . .
```

Processing Wikipedia

- Extracted **text** of each page (using XSLT)
- Segmented into **paragraphs** and **sentences**
- approx. 250K pages, 4.7M sentences, 50M words
- Parsed using **Alpino**

7-1-1|Albert Plesman

7-2-1|Albert Plesman (Den Haag , 7 september 1889 -- ..

7-3-1|Hij werd geboren als zoon van een eierhandelaar ..

7-3-2|In 1915 was hij gelegerd in Soesterberg alwaar ...

7-3-3|Na de Eerste Wereldoorlog mede-organisator van ...

...

7-6-1|Albert Plesman heeft ook de naam Randstad bedacht

Information Retrieval for Wikipedia

- IR retrieves relevant **paragraphs** given a question (query)
- Wikipedia contains many section headings, and short (1 sentence) paragraphs
 - 4.7M sentences
 - 2M paragraphs
 - Many section headings
 - Many 1 sentence paragraphs
- **Paragraphs** were merged until they had a length of **at least 200 words**

Question Answering for Wikipedia

- We **used the existing code** developed for the newspaper corpus for QA
- Answers are found by pattern matching over **dependency parse trees only**
 - *Isaac Newton werd geboren op 25 december 1642.*
- Page-structure of Wikipedia was ignored...

1643

Geboren

januari

4 – Isaac Newton, Brits Natuurkundige

Linguistically Informed Information Retrieval

- Document collection is indexed on words, root forms, syntactic relations, named entity types, and combinations of these

Het embargo tegen Irak werd ingesteld in 1990

root: het embargo tegen Irak word stel_in in
1990

root/head: het/embargo tegen/embargo embargo/word
stel_in/word in/in_stel 1990/in

Query Expansion

- Query Expansion can improve recall
 - Add related words to words in the query
 - Verenigde Staten → VS, USA, Amerika
 - Al Qaida → Bin Laden, Afghanistan
- Method 1: **Blind Relevance Feedback**
 - Query → select 10 keywords from first 5 documents returned
 - New Query = Original Query + selected keywords
- Method 2: **Add (Near) Synonyms**
 - Wikipedia **redirects** (spelling variants),
 - (near) **synonyms** found in automatically aligned Europarl-corpus (van der Plas & Tiedemann, 2006)
 - **ISA-relations** mined from appositions (*composer Aaron Copeland*)

Effect of Query Expansion

Wie stelde een embargo in tegen Irak?

word: (stelde embargo^{4.12} Irak^{4.12} blokkade^{0.5} iraaakse^{0.5} iraaiks^{0.5}
uitvoerverbod^{0.5} buurland^{0.5} Iraaks^{0.5} irak^{0.5} Iraakse^{0.5} kwestie^{0.5})

Wie is de bestuursvoorzitter van Fiat?

word: (bestuursvoorzitter^{4.12} Fiat^{4.12} auto_concern^{0.5} FIAT^{0.5}
autofabrikant^{0.5} auto^{0.5} aandeel^{0.5} merk^{0.5} concern^{0.5})

Follow Up Questions and Anaphora

personal and possessive pronouns

- When was **Napoleon** born?
- Which title was introduced by *him*?
- Who were *his* parents?

impersonal pronouns

- What is the **KNMI**?
- When was *it* founded?

deictic pronouns

- What is an **ecological footprint**?
- When was *this* introduced?

Follow Up Questions and Anaphora

personal and possessive pronouns

- When was **Napoleon** born?
- Which title was introduced by *him*?
- Who were *his* parents?

impersonal pronouns

- What is the **KNMI**?
- When was *it* founded?

deictic pronouns

- What is an **ecological footprint**?
- When was *this* introduced?

Follow Up Questions and Anaphora

personal and possessive pronouns

- When was **Napoleon** born?
- Which title was introduced by *him*?
- Who were *his* parents?

impersonal pronouns

- What is the **KNMI**?
- When was *it* founded?

deictic pronouns

- What is an **ecological footprint**?
- When was *this* introduced?

Follow Up Questions and Anaphora

definite NPs

- Since when is **Cuba** ruled by Fidel Castro?
- When was the flag of *the country* designed?

deictic NPs

- Who lead the Russian Empire during **the Russian-Turkish War of 1787-1792**?
- Who won *this war*?

Follow Up Questions and Anaphora

definite NPs

- Since when is **Cuba** ruled by Fidel Castro?
- When was the flag of *the country* designed?

deictic NPs

- Who lead the Russian Empire during **the Russian-Turkish War of 1787-1792**?
- Who won *this war*?

Anaphora Resolution

- Antecedent has to be a named entity
- From first question or the answer to the first question

What is the capital of Russia?

Moscow

How many inhabitants does it have

8 million

Anaphora Resolution Results

Questions	200	
Qs with Anaphor	56	100%
Correct Antecedent	29	52%
Wrong Antecedent	15	27%
Missed	12	21%

Problematic Cases

Antecedent is not a named entity

- Wat is **mede**?
- Hoe heet **het** in India?

Locative and Temporal Anaphora

- Hoe groot is **Pitcairn**?
 - Welke talen worden **er** gesproken?
-
- Wanneer werd Contra-Aquincum gesticht?
 - **294**
 - Welke keizer was **destijds** aan de macht?

Problematic Cases

Antecedent is not a named entity

- Wat is **mede**?
- Hoe heet **het** in India?

Locative and Temporal Anaphora

- Hoe groot is **Pitcairn**?
 - Welke talen worden **er** gesproken?
-
- Wanneer werd Contra-Aquincum gesticht?
 - **294**
 - Welke keizer was **destijds** aan de macht?

Problematic Cases

Bridging?

- In welke gemeente ligt **Helvoirt**?
- Hoe heet **het jaarlijkse evenement** rond Hemelvaartsdag?

- Wanneer werd **de Efteling** geopend ?
- **Welke nieuwe attractie** werd geopend in 1993 ?

Multilingual QA

When was Napoleon Bonaparte born?

Babelfish
& Wikipedia
& geonames.org

Question class
Answer

Wanneer was Napoleon Bonaparte geboren?

born_date(Napoleon Bonaparte)
Napoleon Bonaparte (1769-1821),
Frans militair , dictator (1799-1804)
en keizer (1804-1815)

Question Classification for MLQA

Translations for which Question Classification Fails

Hoe lang gaat de kosmische opdracht verder " Voyager " ?

Wat PB Simpson werd beschuldigd van ?

Met hoeveel " lander " zijn er in Duitsland ?

Welke persoon goedkeurde niet de toekenning die door het
Instituut Goethe wordt toegekend ?

- Question classification is essential for Answer Extraction

Question Classification for MLQA

- Do question classification on English source question
 - Using on-line system (**QUEST**)
- **Map** English question classes to Joost question classes
- **Use both** the question class of the English question and the Dutch translation (if any).

Question classification results

Testset	Qs	Joost		union	
		MRR	1st	MRR	1st
2003	377	0.329	0.292	0.347	0.310
2004	200	0.406	0.350	0.429	0.375
2006	200	0.225	0.195	0.213	0.185

Joost at CLEF2007

Run	Acc (%)	R	X	U	W
Dutch-mono	24.5	49	11	4	136
Dutch-mono + QE	25.5	51	10	4	135
Univ Ams	8.0	15	1	23	161
En-Du	13.0	26	8	7	159
En-Du + QE	13.5	27	7	5	161

R = Right, X = IneXact, U = unsupported, W = wrong

Joost at CLEF2007

Q type	# q's	Acc (%)	R	X	U	W
Factoids	156	25.6	40	5	4	107
List	16	6.3	1	0	5	10
Definition	28	35.7	10	0	0	18
Temp. Restricted	41	19.5	8	3	3	27
NIL	20	0.0	0	0	0	20

R = Right, X = IneXact, U = unsupported, W = wrong

Some Observations

- Impact of **Wikipedia** is significant:
 - 150 answer snippets from Wikipedia, 30 from newspapers
- Effect of Query Expansion
 - 27 questions receive a **different** answer
 - Small improvement in accuracy
- **20** questions with **NIL** as answers
 - Mistakes in anaphora resolution
 - Question Classification failed

Follow Up Questions per Language

Target	FQs	Target	FQs
EN	133	RO	78
NL	122	FR	76
DE	84	PT	50
IT	84	ES	30

Future Work

- Relation Extraction from Wikipedia XML

Patterns for Dates of Birth

Dependency Triples	54.963
Dep Triples + Wikipedia 1st sentences	98.697
RegEx over text of year pages	6.007
XQuery (look at 1st par of person pages)	33.123
XQuery (Wikipedia templates)	1.103