

Question answering using Syntactic Analysis



Gosse Bouma Jori Mur Gertjan van Noord
Lonneke van der Plas Jörg Tiedemann
www.let.rug.nl/~gosse/Imix

Information Science
Rijksuniversiteit Groningen

Nijmegen 06/05 - p.1/42

Open Domain Question Answering

- Given a question,
- Find the **exact answer** in a text collection.
- QA for English: TREC
- QA for European languages:
 - Mono- and multilingual: CLEF
 - Dutch Text Collection: Algemeen Dagblad & NRC Handelsblad (1994 & 1995)
 - Test sets: 200 (clef 04, 05) - 450 questions (clef 03)

Nijmegen 06/05 - p.3/42

Overview

- Open Domain Question Answering
- Using Syntactic Analysis for QA
 - Named Entity Classification
 - Dependency Patterns and Equivalences
- Using Dependency Relations
 - Answer Extraction and Ranking
 - Off-line Extraction
- Evaluation
- Conclusions & Further Research

Nijmegen 06/05 - p.2/42

Examples

- 1 Wanneer vond de Duitse hereniging plaats?
- 2 Wie is bestuursvoorzitter van Fiat?
- 3 Wie stelde een embargo in tegen Irak?
- 4 In welk land ligt Bosnië?
- 5 Wat is de afkorting voor gekke-koeienziekte?
- 10 Welke bevolkingsgroepen voerden oorlog in Rwanda?
- 20 Hoeveel inwoners heeft Berlijn?
- 21 Hoeveel kost een XTC pil?
- 60 Hoe lang was de Berlijnse Muur?

Nijmegen 06/05 - p.4/42

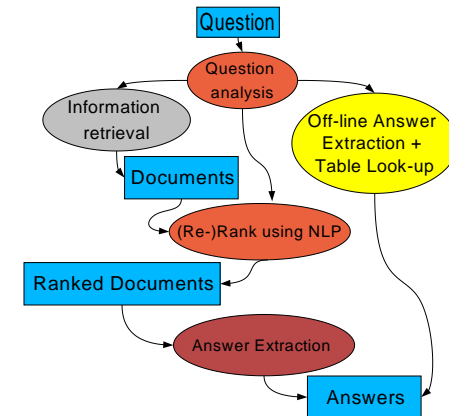
CLEF 2003 examples

Hoe heet de nationale luchtvaartmaatschappij van Zwitserland ?

<DOCID>NH19941215-0098</DOCID>
<P>BRUSSEL, 15 DEC. De Belgische regering zal op afzienbare termijn besprekingen houden met de Zwitserse luchtvaartmaatschappij **Swissair** over de mogelijkheid van een verbintenis tussen Swissair en de Belgische nationale luchtvaartmaatschappij Sabena.</P>

Nijmegen 06/05 - p.5/42

Our QA System (Joost)



Nijmegen 06/05 - p.7/42

CLEF 2003 examples

Wie is de oprichter van de Orde van de Zonnetempel ?

<P>HAAGS FILMHUIS Kuifje en de **zonnetempel** zo. wo. 14u ...</P>
<P>De leiders van de **Orde van de Zonnetempel**, de Canadees Luc Jouret (46) en de Fransman Joseph Dimembro (70),... </P>
<P>...nadat in de Canadese stad Québec het huis was afgebrand van **Luc Jouret**, een 46-jarige homeopaat die **oprichter** is **van de Orde van de Zonnetempel**.</P>

Nijmegen 06/05 - p.6/42

Using Dependency Relations for QA

- Syntactic Structure can provide useful information
 - For Question Analysis
 - For Answer Extraction & Ranking
 - For Off-line extraction
- **Bottlenecks**
 - Robust, wide-coverage, parsing?
 - How to use syntactic information?

Nijmegen 06/05 - p.8/42

Syntactic Analysis for Q Classification

- **Waar** staat de oudste vuurtoren van Nederland? (LOC)
 - *Where is the oldest lighthouse of the Netherlands located?*
- **Waar** staat de afkorting EMU voor? (ABBR)
 - *what does EMU stand for?*
- **Waar** leven cicaden van? (OBJECT)
 - *What do cicada live on?*
- **Waar** is rioolwater rijk aan? (OBJECT)
 - *What does water from the sewer contain abundantly?*

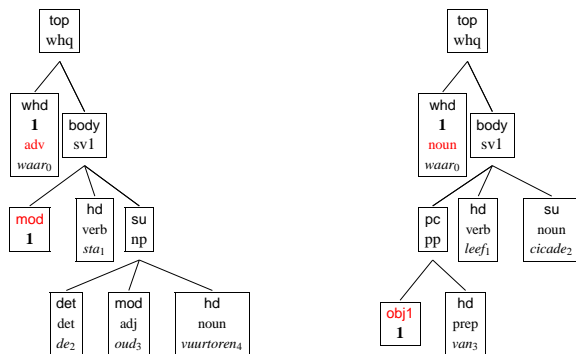
Nijmegen 06/05 – p.9/42

Answer Extraction and Ranking

- IR engines return large amounts of text
 - \approx 10K words (500 sentences) per question
- Off-line Answer Extraction strategies **search the corpus exhaustively**,
- All of this **needs to be parsed** before Dependency Relations can be used

Nijmegen 06/05 – p.11/42

Classification of *Waar* questions



Nijmegen 06/05 – p.10/42

But Full Parsing is impossible

*Over the last five years, I don't know how often I have read a clause regretting the impossibility of full parsing and expressing the conviction that one has to make do with less desirable approximations. ...**“Full parsers are slow, ..yield so many parses, ..are brittle.”**... When this mantra was first invented it might have expressed a truth, but by now the reality is different...*

Annie Zaenen, *ElsNews* 13.1

Nijmegen 06/05 – p.12/42

Wide-coverage Parsing for QA

- Existing Dutch wide-coverage parser (Alpino)
 - produces (CGN) dependency structures
- Integrated Named Entity Classifier
- Used to parse full CLEF corpus
- *Tuned for processing of questions*

Nijmegen 06/05 – p.13/42

Parser

- Optimized version of Left-corner Parser
- If no full parse available, return sequence of **partial parses**
- Parse Forest: compact representation of all parses
- Extract *best* parse from the parse forest
- Uses **maximum-entropy disambiguation** model

Nijmegen 06/05 – p.15/42

Alpino Grammar

- Wide-coverage HPSG > 500 rules
- Specific rules *and* general constraints (object-oriented)
- Large lexicon (> 100K words)
- Many **heuristics** for unknown words and word sequences
- Output: dependency structures

Nijmegen 06/05 – p.14/42

Accuracy

Corpus	Sentences	Accuracy (%)
Eindhoven cdbi	7000	85.9
Trouw 2001	500	88.0
CLEF 03	450	95.3
CLEF 04	200	95.8
CLEF 05	200	97.6

(Malouf and van Noord 04, IJCNLP workshop
Beyond Shallow Analyses)

Nijmegen 06/05 – p.16/42

Named Entity Classification

- Question Class predicts Answer Type
 - *Wie heeft de Berlijnse Muur gebouwd?*
(**Person**)
 - ...partijleider **Walter Ulbricht**,
die de Berlijnse Muur liet
bouwen...
- Classes: **persons, locations, organizations.**

Nijmegen 06/05 – p.17/42

Processing the Text Collection

- NRC 1994, 1995; AD 1994,1995
- 4,1 mln sentences, 78 mln words
- Parsed with Alpino, stored dependency trees
 - 4,1 mln XML files
 - **25 Gb**
 - **500 CPU days**

Nijmegen 06/05 – p.19/42

NE Classification in Alpino

- Named Entity **recognition** already exists
- Named Entity **classification** uses
 - **dictionary** of names (150K, mostly from TwNC)
 - **maximum entropy classifier** (trained on CONLL data)
- **88.5% accuracy** (50% dict, 50% MaxEnt)

Nijmegen 06/05 – p.18/42

Multi-purpose Corpus

- **Corpus-linguistics**
 - phrasal prepositions, support verb constructions, determinerless PPs, focus adverbs inside PPs, PP-fronting (van der Beek, Villada, Bouma,..)
- **Acquisition of Ontological knowledge**
 - ISA-relations, distributionally similar words, terminology (van der Plas, Fahmi)
- **Error Mining** (van Noord ACL 04)
- **Psycholinguistic Research** (Nijmegen),
- **Other QA research** (Amsterdam)

Nijmegen 06/05 – p.20/42

How to use dependency relations?

- Wie kreeg asiel?

$$Q = \left\{ \begin{array}{l} \langle \text{krijg/K}, \text{obj1}, \text{asiel/A} \rangle, \\ \langle \text{krijg/K}, \text{su}, \text{Su/S} \rangle \end{array} \right\}$$

- Mengistu kreeg asiel in Zimbabwe

$$A = \left\{ \begin{array}{l} \langle \text{krijg/2}, \text{su}, \text{mengistu/1} \rangle, \\ \langle \text{krijg/2}, \text{obj1}, \text{asiel/3} \rangle, \\ \langle \text{krijg/2}, \text{mod}, \text{in/4} \rangle, \\ \langle \text{in/4}, \text{obj1}, \text{zimbabwe/5} \rangle \end{array} \right\}$$

Nijmegen 06/05 – p.21/42

Syntactic Variation

- de bondscoach van Noorwegen, Egil Olsen \Leftrightarrow
Egil Olsen, de bondscoach van Noorwegen
- president van Rusland, Jeltsin \Leftrightarrow
Jeltsin is president van Rusland
- Australië's staatshoofd \Leftrightarrow
staatshoofd van Australië
- Moskou heeft 9 miljoen inwoners \Leftrightarrow
de 9 miljoen inwoners van Moskou
- Swissair en AA hebben vluchten naar Kroatië \Leftrightarrow
Swissair heeft vluchten naar Kroatië

Nijmegen 06/05 – p.23/42

Syntactic Variation

- Wie verleende asiel aan Mengistu?
- Zimbabwe verleende asiel aan Mengistu
- Aan Mengistu werd asiel verleend door Zimbabwe

$$\text{equiv}(\{\langle V, \text{su}, \text{Su/S} \rangle\}, \left\{ \begin{array}{l} \langle \text{word/W}, \text{vc}, V \rangle, \\ \langle V, \text{mod}, \text{door/D} \rangle, \\ \langle \text{door/D}, \text{obj1}, \text{Su/S} \rangle, \end{array} \right\}).$$

Nijmegen 06/05 – p.22/42

Syntactic Variation

- Ulbricht liet de Berlijnse Muur bouwen \Leftrightarrow
Ulbricht, die de Berlijnse Muur liet bouwen.
- De veerboot Estonia zonk in de Oostzee \Leftrightarrow
de in de Oostzee gezonken veerboot Estonia
- De EEG werd in 1957 opgericht \Leftrightarrow
De Europese Economische Gemeenschap werd in 1957 opgericht
- ...

Nijmegen 06/05 – p.24/42

Answer Ranking by Dep Match

- Wie is de voorzitter van het Europese Parlement?
- **Klaus Hänsch**, voorzitter van het Europese Parlement , ...
- **Karin Junkers** (SPD), lid van **het Europese Parlement** en **voorzitter** van de vereniging van sociaal-democratische vrouwen in Europa,

$$\text{match}\left(\left\{\begin{array}{l} \langle \text{voorzitter}/V, \text{mod}, \text{van}/W \rangle, \\ \langle \text{van}/W, \text{obj1}, \text{parlement}/X \rangle \end{array}\right\}, Q\right)$$

Nijmegen 06/05 – p.26/42

Answer Ranking

$$\begin{aligned} \text{a-score}(Q, A) = & \alpha.\text{type-score}(Q, A) & + \\ & \beta.\text{name-score}(Q, A) & + \\ & \gamma.\text{deprel-score}(Q, A) & + \\ & \delta.\text{frequency}(Q, A) & + \\ & \epsilon.\text{IR}(Q, A) \end{aligned}$$

Nijmegen 06/05 – p.27/42

Answer Ranking by Dep Match

- Wanneer vond de Duitse hereniging plaats?
- `date(hereniging)`
- Sinds de Duitse hereniging in oktober 1990 ...
- Al in 1962 voorspelde hij de Duitse hereniging ...
- Prefer answers where **date** is a dependent of hereniging

Nijmegen 06/05 – p.28/42

General WH-questions

- Answering wh-questions requires ontological knowledge:
 - Welke vulkaan barstte uit in juni 1991?
 - Answer **isa** vulkaan
 - Welke voetballer ontving "De Gouden Bal" in 1995?
 - Answer **isa** voetballer
 - Welke oliemaatschappij is eigenaar van de Brent Spar?
 - Answer **isa** oliemaatschappij

Nijmegen 06/05 – p.28/42

Definition Questions

- Provide a characteristic property of a person or organization
 - Wie is Bettino Craxi?
Italiaanse oud-premier
 - Wat is Sabena?
Belgische luchtvaartmaatschappij
 - Wie is Willy Claes?
Belgische secretaris-generaal van de NAVO
 - Wie is Diego Armando Maradona?
ongewenste zorgenkind
- 50 announced for CLEF 05

Nijmegen 06/05 – p.29/42

Using ISA relations

- Wh-questions
 - Welke veerboot zonk ten zuidoosten van het eiland Utö?
 - Tallinn, Estonia, Raimo Tiilikainen,...
 - Prefer answer in ISA relation to veerboot
- Definitions
 - Wat is BMW?
 - BMW ISA autofabrikant,
 - find source sentence, include modifying adjectives and PPs

Nijmegen 06/05 – p.31/42

Acquisition of ISA relations

- For Named Entities only
- Extracted .5 mln apposition relations
 - **veerboot** Estonia, Anna Maria Lauro, Sally Star, Herald of Free Enterprise, Prinses Beatrix, ... (112)
 - **bondscoach** Guus Hiddink, Hans Jorritsma, Joop Alberda, Terry Venables, Berti Vogts, ... (2951)

Nijmegen 06/05 – p.30/42

Off-line Extraction

- Exhaustively search corpus for answers to specific question types
 - capital, abbreviations, number of inhabitants, year of birth, ..
- Dependency relations allow complex conditions to be stated
- Equivalences
 - reduce number of required patterns,
 - enhance recall

Nijmegen 06/05 – p.32/42

Founder-Organization

- **Minderop** richtte de Tros op toen
- Op last van generaal De Gaulle in Londen richtte verzetsheld **Jean Moulin** in mei 1943 de Conseil National de la Résistance (CNR) op.
- **Het Algemeen Ouderen Verbond** is op 1 december opgericht door de nu 75-jarige **Martin Batenburg**.
- toen **de Generale Bank** bekend maakte met de Belgische Post een "postbank" op te richten.

Nijmegen 06/05 – p.33/42

Enhancing Recall

- Function Words:
 - **Seed list** extracted from **EuroWordNet** (*voorzitter, president, leider, koning, kopman, ..., 251 words*)
 - Automatically extended with distributionally similar words (*aandeelhouder, aanvaller, componist, vader, ... 895 words*).
- **A is distributionally similar to B** if it occurs in similar syntactic contexts
 - www.let.rug.nl/~gosse/Sets

Nijmegen 06/05 – p.35/42

Function Table

voorzitter	KNOV	10	Kamminga
voorzitter	FNV	90	Stekelenburg
voorzitter	Borgharen	1	drs J. Roebroek
burgemeester	Rotterdam	88	Peper
president	Servisch	220	Milosevic
bevelhebber	VN_vredesmacht	2	Cot
president	Bundesbank	52	Tietmeyer
president	Walt Disney	1	Michael Eisner

Nijmegen 06/05 – p.34/42

Enhancing Recall

Table	-Equiv		+Equiv		Incr (%)	
	tuples	uniq	tuples	uniq	tpls	uniq
Age	15.981	13.716	22.143	18.520	38	35
Born Date	1.545	1.356	2356	1.990	54	47
Capital	1.940	406	2.146	515	10	27
Died Age	522	379	1.127	834	116	120
Founded	604	559	1.021	953	69	70
Func'n EWN+	54.016	28.543	77.028	46.589	43	63
Func'n EWN			34.191	16.530		

Nijmegen 06/05 – p.36/42

Evaluation

- CLEF 03 (372 questions), 04 (200 questions)
- **Off-line QA:** question analysis → table-lookup → answer extraction & reranking
- **On-line QA:** question analysis → IR → answer extraction & reranking
- **Scores:**
 - MRR(5): 1/(rank of 1st correct answer), only for 1st 5 answers.
 - CLEF: 1st answer correct

Evaluation of Specific Q-types

	-isa		+isa	
	# q	CLEF	# q	CLEF
WH 03	58	0.26	58	0.39
WH 04	19	0.26	19	0.47
Definition 04	23	0.17	23	0.44

Evaluation

CLEF 03	off-line		on-line		total	
	# q	mrr	# q	mrr	# q	mrr
baseline	85	0.71	287	0.36	372	0.44
+d-score	85	0.71	287	0.40	372	0.47
+equiv	89	0.77	283	0.40	372	0.49
CLEF 04						
baseline	61	0.56	139	0.42	200	0.46
+d-score	61	0.57	139	0.45	200	0.48
+equiv	71	0.73	129	0.49	200	0.58

CLEF Evaluation

	off-line		on-line		total	
	# q	CLEF	# q	CLEF	# q	CLEF
CLEF 03	87	0.74	285	0.36	372	0.45
CLEF 04	69	0.73	131	0.46	200	0.55

Official CLEF results monolingual QA	
Average	0.24
Best	0.46

Unofficial CLEF 05 Evaluation

- Data released: May 18, deadline: May 25
- 200 questions
 - 60 definition questions
- Estimated CLEF score
 - 0.54 (according to 3 of 4 judges)
 - 0.55 (according to independent judge)
 - 0.65 (definitions only)

Nijmegen 06/05 – p.41/42

Future Work

- Address **Lexical Variation**
 - More ISA-relations, Synonyms, Spelling Variation, Terminological Variants, Paraphrases, ...
- **Paragraph Level** Reasoning
 - Coreference resolution
- Use Syntactic Information in **IR**

Nijmegen 06/05 – p.42/42