Question Answering using Sentence Parsing and Semantic Network Matching

Sven Hartrumpf
http://pi7.fernuni-hagen.de/hartrumpf

Intelligent Information and Communication Systems
Computer Science Department
University of Hagen (FernUniversität in Hagen)
58084 Hagen, Germany

2004-09-16
CLEF 2004, Bath, UK

Introduction

InSicht: question answering (QA) system implemented for German

Key characteristics:

1. Deep syntactico-semantic analysis of questions and documents (with a parser)
2. Independence from other document collections (like WWW documents)
   ⟶ avoids unsupported answers
3. Answer generation from semantic representations of documents (no direct extraction)

Related system for German: ⟶ Neumann and Xu (2003).
Relies on shallow, but robust methods.
InSicht: builds on deep parsing

Related system for English: ⟶ Harabagiu et al. (2001).
Applies a theorem prover and a large knowledge base to validate candidate answers

Overview

Document Processing

Each article is stored in an SGML file conforming to the CES
(Corpus Encoding Standard, (Ide et al., 1996))

Elimination of duplicate articles

Table 1: Statistics from Document Preprocessing

subcorpus	articles without duplicates	sentences	words	average sentence length	duplicate articles	
					identical bytes	identical words
FR	122541	2472353	45332424	18.3	22	17152
SDA	140214	1930126	35119427	18.2	333	568
SP	13826	495414	9591113	19.4	0	153
all	276581	4897893	90042964	18.4	355	17873

Syntactico-semantic parser WOCADI (WOrd ClAss based DIsambiguating): transforms articles into semantic networks (MultiNet formalism, (Helbig, 2001; Helbig and Gnörlich, 2002))

Each sentence is represented by one semantic network

Semantic networks are simplified and normalized
⟶ allows more efficient search

Table 2: Statistics from Document Parsing

subcorpus	parse results	full parse (%)	chunk parse (%)	no parse (%)
FR	2469689	44.3	21.7	34.0
SDA	1930111	55.8	19.0	25.2
SP	485079	42.7	19.3	38.0
all	4884879	48.7	20.4	30.9

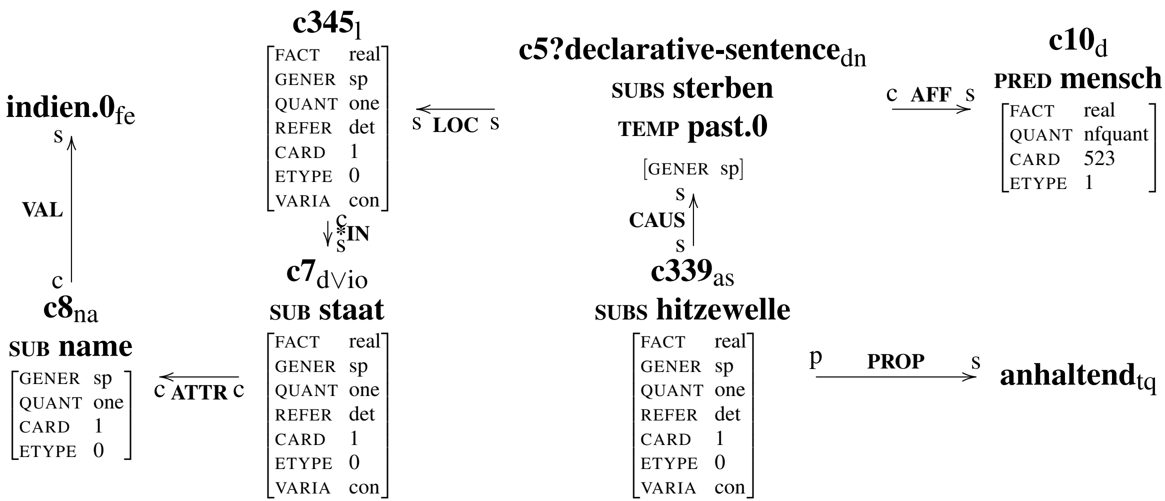

Figure 1: MultiNet generated for document sentence SDA.950618.0048.377:
In Indien starben [. . .] 523 Menschen infolge der [. . .] anhaltenden Hitzewelle.
(*‘523 people died in India due to the continuing heat wave.’*)

Question Processing

Question is parsed by the WOCADI parser
⟶ semantic network, (question) focus, sentence type

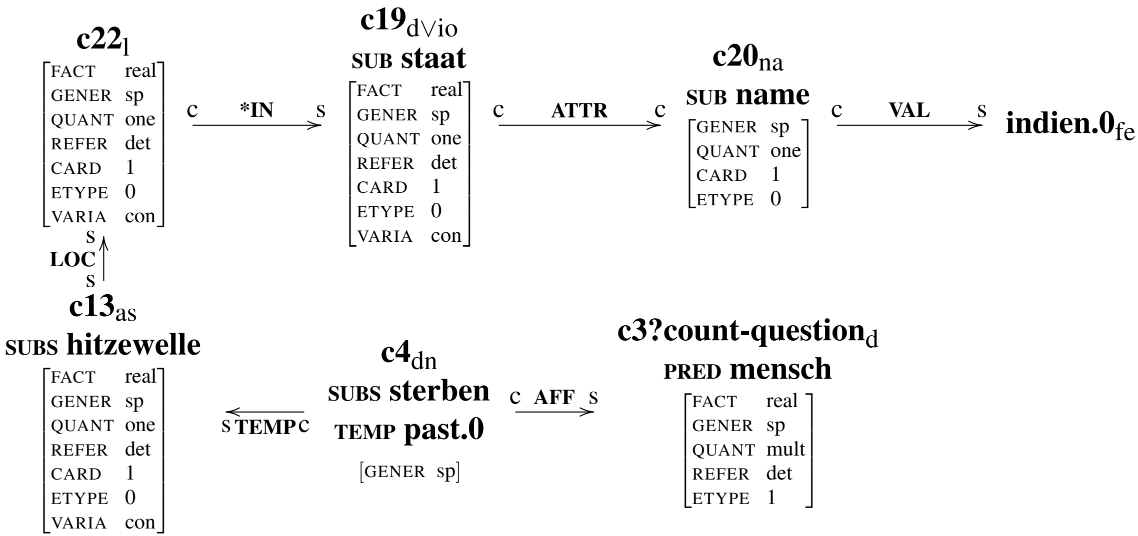

Figure 2: MultiNet generated for question 164:
Wie viele Menschen starben während der Hitzewelle in Indien?
(*'How many people died during the heat wave in India?'*)

Query Expansion

Query expansion generates equivalent (or similar) semantic networks
$\longrightarrow$ find answers that are not explicitly contained in a document but only implied

1. Equivalence rules (or paraphrase rules) for MultiNet:
   work on semantic networks, not on surface strings (important because of freer word order)

2. Rule schemas (for maintenance reasons):
   e.g. one schema generates 190 connections of the type:
   Spanien, *Spanier*, *spanisch*
   (*'Spain'*, *'Spaniard'*, *'Spanish'*)

3. Implicational rules for lexemes (used in backward chaining):
   e.g. entailment between *ermorden.1.1* (*'kill'*) and *sterben.1.1* (*'die'*)

4. Lexico-semantic relations (synonymy, hyponymy, etc.):
   from the lexicon (HaGenLex, (Hartrumpf et al., 2003)),
   from GermaNet

Query expansion results per question from QA@CLEF 2004:
6.5 additional semantic networks,
215 using lexico-semantic relations

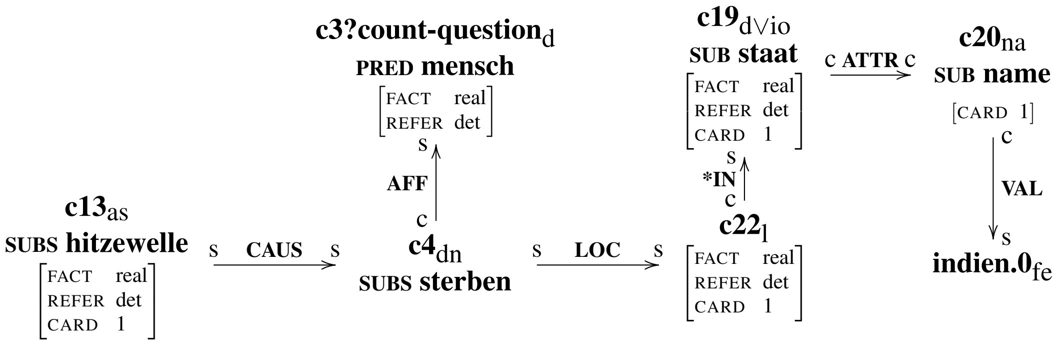

Figure 3: One result from query expansion for question 164 from Figure 2

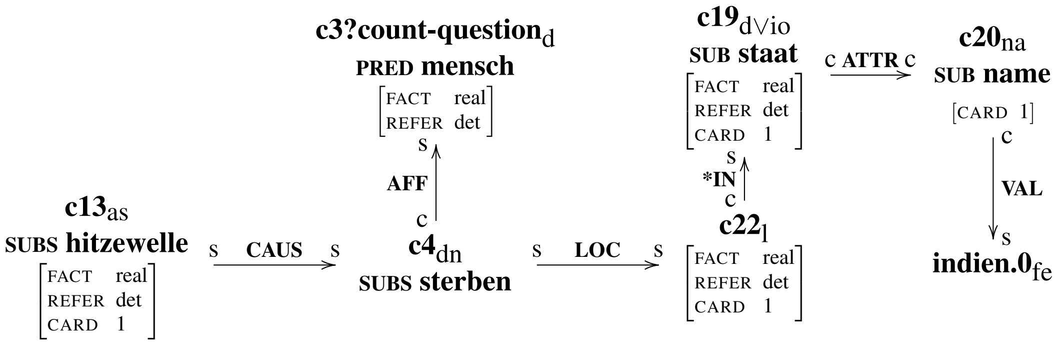

Figure 3: One result from query expansion for question 164 from Figure 2

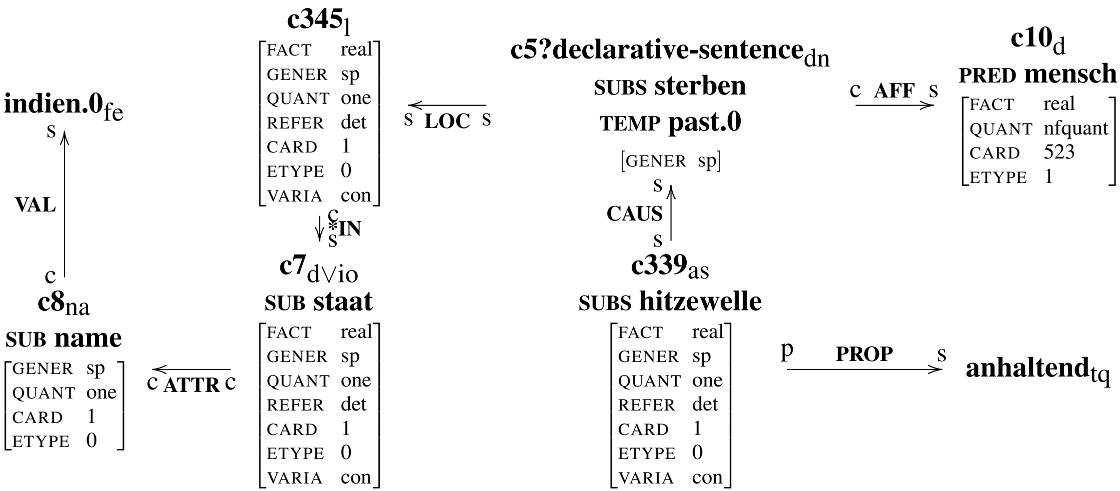

Figure 4: MultiNet for document sentence (repeated from Figure 1)

Search for Matching Semantic Networks

Idea: find a document sentence containing an answer by semantic network matching

Semantic network for the question is split:

1. the *queried network*
   (roughly corresponding to the phrase headed by the interrogative pronoun or determiner)

2. the *match network*
   (the semantic network without the queried network)

Concept ID index server for speedup

Semantic networks are simplified and normalized to achieve acceptable answer times:

1. Inner nodes of a semantic network that correspond to instances (cN) are combined with their concept nodes
   $\longrightarrow$ a lexicographically sorted list of MultiNet edges as a canonical form
   $\longrightarrow$ allows efficient matching with many question networks in parallel

2. Semantic details from some layers in MultiNet are omitted

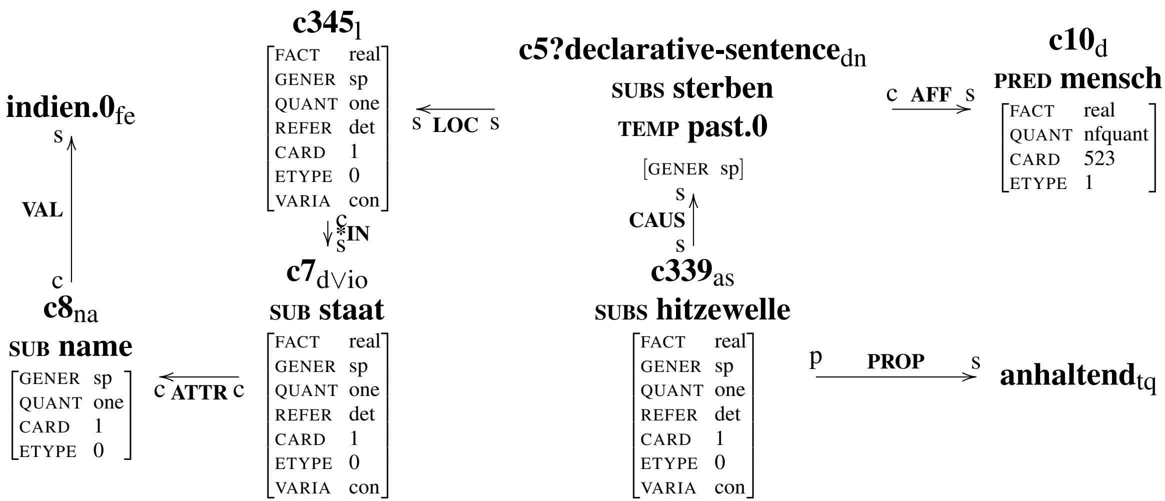

Figure 5: MultiNet for document sentence (repeated from Figure 1)

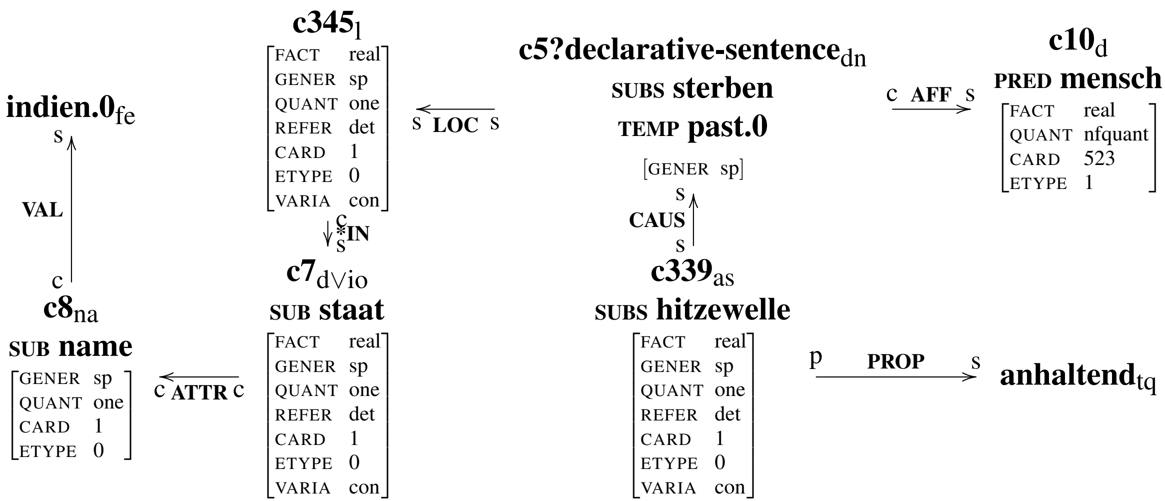

Figure 5: MultiNet for document sentence (repeated from Figure 1)

(*in ”c1*in” ”c1staat.1.1”)
(aff ”c1sterben.1.1” ”c1mensch.1.1”)
(attr ”c1staat.1.1” ”c1name.1.1”)
(caus ”c1hitzewelle.1.1” ”c1sterben.1.1”)
(loc ”c1sterben.1.1” ”c1*in”)
(prop ”c1hitzewelle.1.1” ”anhaltend.1.1”)
(temp ”c1sterben.1.1” ”past.0”)
(val ”c1name.1.1” ”indien.0”)

Figure 6: Simplified and normalized semantic network for the MultiNet of Figure 5 (without layer features)

Answer Generation

Generation rules

Input:

1. simplified semantic network of the question (the *queried network* part)
2. sentence type of the question
3. matching semantic network from the document

Output: a German phrase as a candidate answer or failure

Answer Generation

Generation rules

Input:

1. simplified semantic network of the question (the *queried network* part)
2. sentence type of the question
3. matching semantic network from the document

Output: a German phrase as a candidate answer or failure

Answer Selection

Result of the preceding step:
pairs of generated answer string and supporting sentence ID

Choice from candidate answers:
preference for longer answers and preference for more frequent answers

Evaluation on the QA@CLEF 2004 Test Set

One goal: Identify areas of improvement
by annotating each question leading to a suboptimal answer with a *problem class*

InSicht achieved 80 (submitted run: 67) correct and 7 (subm. run: 2) inexact answers for 197 questions
⟶ leaves 110 questions (with incorrect empty answer) to be annotated

Sample of 43 questions

Table 3: Hierarchy of problem classes and problem class frequencies

name	description	%
problem		
q.error	error on question side	
q.parse_error	question parse is not complete and correct	
q.no_parse	parse fails	0.0
q.chunk_parse	only chunk parse result	0.0
q.incorrect_parse	parser generates full parse result, but it contains errors	13.3
q.ungrammatical	question is ungrammatical	2.7
d.error	error on document side	
d.parse_error	document sentence parse is not complete and correct	
d.no_parse	parse fails	33.2
d.chunk_parse	only chunk parse result	2.0
d.incorrect_parse	parser generates full parse result, but it contains errors	7.8
d.ungrammatical	document sentence is ungrammatical	2.0
q-d.error	error in connecting question and document	
q-d.failed_generation	no answer string can be generated for a found answer	2.0
q-d.matching_error	match between semantic networks is incorrect	5.9
q-d.missing_cotext	answer is spread across several sentences	5.9
q-d.missing_inferences	inferential knowledge is missing	25.4

Three problems per question possible, but stop after first problem to avoid speculation

Conclusions and Perspectives

InSicht's achievements:

1. High precision: non-empty answers (i.e. non-NIL answers) are rarely wrong for QA@CLEF 2004: 0 (submitted run: 1)

2. Deep level of representation based on semantic networks:

   allows intelligent processes, e.g. paraphrasing on semantic level, inferences

Conclusions and Perspectives

InSicht's achievements:

1. High precision: non-empty answers (i.e. non-NIL answers) are rarely wrong for QA@CLEF 2004: 0 (submitted run: 1)
2. Deep level of representation based on semantic networks:
   allows intelligent processes, e.g. paraphrasing on semantic level, inferences

Problem areas and directions for future work:

1. Inferential knowledge
   ⟶ encode and semi-automatically acquire entailments etc.
2. Parser coverage
   ⟶ extend the lexicons and improve robustness and grammatical knowledge of the parser
3. Partial semantic networks
   ⟶ devise methods to utilize partial semantic networks for finding answers
4. Answers spread across several sentences
   ⟶ apply the parser in text mode (coreference resolution, (Hartrumpf, 2001))
5. Processing time for documents
   ⟶ develop a strategy for on-demand processing

References

Harabagiu, Sanda; Dan Moldovan; Marius Paşca; Rada Mihalcea; Mihai Surdeanu; Răzvan Bunescu; Roxana Gîrju; Vasile Rus; and Paul Morărescu (2001). The role of lexico-semantic feedback in open-domain textual question-answering. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics (ACL-2001)*, pp. 274–281. Toulouse, France. 1

Hartrumpf, Sven (2001). Coreference resolution with syntactico-semantic rules and corpus statistics. In *Proceedings of the Fifth Computational Natural Language Learning Workshop (CoNLL-2001)*, pp. 137–144. Toulouse, France. URL http://www.aclweb.org/anthology/W01-0717. 14

Hartrumpf, Sven (2003). *Hybrid Disambiguation in Natural Language Analysis*. Osnabrück, Germany: Der Andere Verlag. ISBN 3-89959-080-5.

Hartrumpf, Sven; Hermann Helbig; and Rainer Osswald (2003). The semantically based computer lexicon HaGenLex – Structure and technological environment. *Traitement automatique des langues*, 44(2):81–105. 7

Helbig, Hermann (2001). *Die semantische Struktur natürlicher Sprache: Wissensrepräsentation mit MultiNet*. Berlin: Springer. 4

Helbig, Hermann and Carsten Gnörlich (2002). Multilayered extended semantic networks as a language for meaning representation in NLP systems. In *Computational Linguistics and Intelligent Text Processing (CICLing 2002)* (edited by Gelbukh, Alexander), volume 2276 of *LNCS*, pp. 69–85. Berlin: Springer. 4

Helbig, Hermann and Sven Hartrumpf (1997). Word class functions for syntactic-semantic analysis. In *Proceedings of the 2nd International Conference on Recent Advances in Natural Language Processing (RANLP'97)*, pp. 312–317. Tzigov Chark, Bulgaria.

Ide, Nancy; Greg Priest-Dorman; and Jean Véronis (1996). *Corpus Encoding Standard*. URL http://www.cs.vassar.edu/CES/. 3

Neumann, Günter and Feiyu Xu (2003). Mining answers in German web pages. In *Proceedings of the International Conference on Web Intelligence (WI-2003)*. Halifax, Canada. 1

Osswald, Rainer (2004). Die Verwendung von GermaNet zur Pflege und Erweiterung des Computerlexikons HaGenLex. *LDV Forum*, 19(1):43–51.