Diplomarbeitspräsentationen der Fakultät für Informatik

Coreference Resolution in Clinical Practice Guidelines Focusing on Hypernym/Hyponym Relations

FAKULTÄT FÜR !NFORMATIK

Masterstudium:
Wirtschaftsinformatik

Marco Romauch
marco.romauch@gmail.com

Technische Universität Wien
Institute für Softwaretechnik und Interactive Systeme
Arbeitsbereich: Information and Software Engineering Group
Betreuerinnen: Ao.Univ.-Prof. Mag. Dr. Silvia Miksch
Projektass. Mag. Dr. Katharina Kaiser

Motivation

Automated processing of medical information can considerably support and improve the clinical decision making process.

VISION

The majority of medical information is only available in natural language text such as clinical practice guidelines (CPG) that represent the current valid knowledge in a certain medical field.

PROBLEM

The correct automated interpretation of **COREFERENCE RELATIONS** is an important step to achieve that goal.

SOLUTION

Development of automated Methods that aim to help correctly "understand"the content of a natural language medical text.

TASK

Use natural language processing (NLP) techniques as an approach to provide computer interpretable CPG documents.

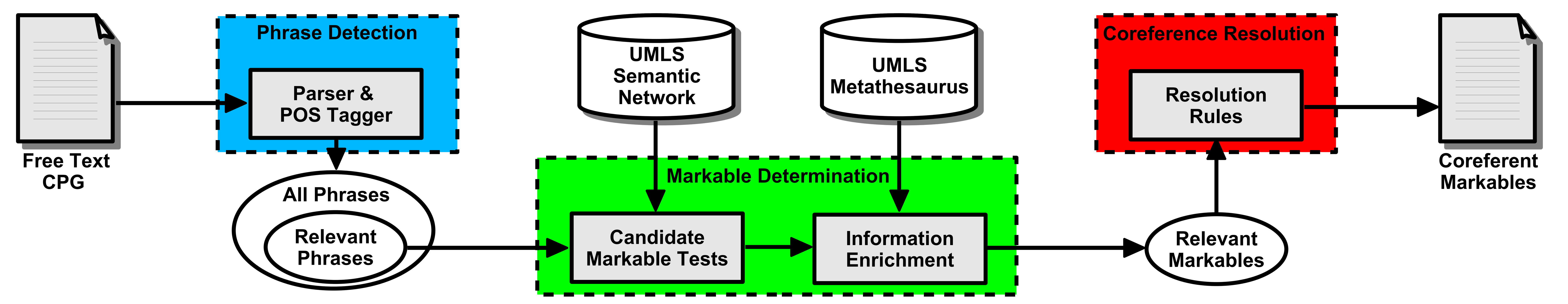
Theoretical Background

A **COREFERENCE RELATION** is a linguistic structure that holds between two or more expressions in a text whereas all are related to the same entity.

Frequent Coreference Types in CPGs

- **Acronym Definition Coreference**
  "For patients with stage I and II **non-small cell lung cancer** (**NSCLC**), surgery to remove the NSCLC is the treatment of choice."
- **Acronym Coreference**
  "Evaluation with preoperative computed tomography (**CT**) …"
  "Many series have reported the utility **of CT** in detection of …"
- **Hypernym/Hyponym Coreference**
  "To reduce the incidence and mortality rate of **cervix cancer**, effective screening and preventive strategy must be actively pursued, in addition to early detection of **the disease**."
  "… for patients with **chemotherapy-associated anemia** ..."
  "… depending upon the severity **of anemia** or..."

Schematically Illustration of the Coreference Resolution Algorithm



Coreference Resolution Steps

Step 1: Phrase Detection

In the first processing step the input text gets tokenized and parsed in order to identify all existing phrases. Furthermore, the algorithm uses the functionality provided by the MMTx in order to map all identified (bio)medical terms to the best matching UMLS concept or set of concepts.

Step 2: Markable Determination

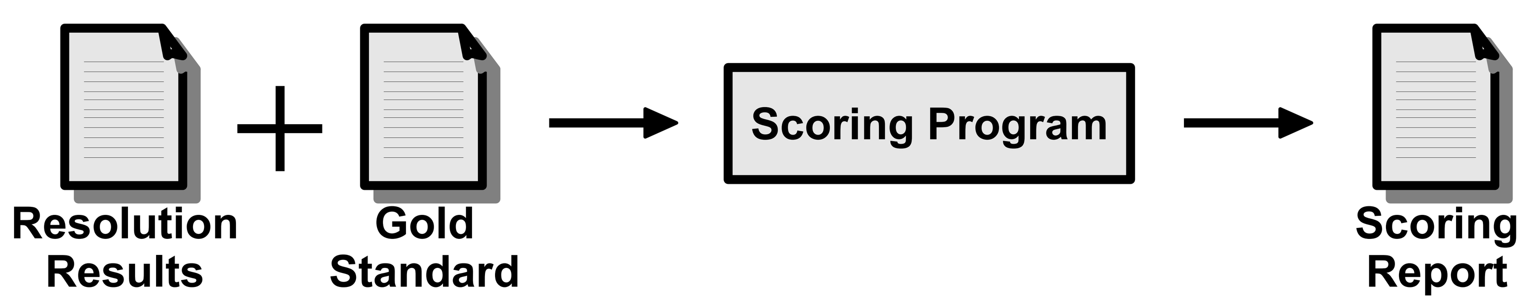
This stage aims to determine all (bio)medical phrases that are possibly part of a coreferent relationship. In order to identify these so called markables, the algorithm firstly selects all relevant phrases (noun and prepositional phrases) created by the MMTx. In a second step morphological and semantic information derived from the UMLS Metathesaurus is incorporated in order to determine the relevancy of a markable candidate. Relevant markables are finally enriched with additional semantic information that is also gathered from the Metathesaurus.

Step 3: Coreference Resolution

In the final step a set of predefined coreference resolution rules is applied to every possible pair of markables in order to to denote a coreference resolution. The necessary semantic, syntactic and morphological information to perform this task is provided by the MMTx respectively the UMLS.

Performance Evaluation & Results

After a training phase the performance of the algorithm is evaluated by comparing its resolution results with respect to a so called gold standard template. A scoring program computes two metrics (recall and precision) that show the resolution capabilities of the algorithm.



The performance of the coreference resolution algorithm was tested on 3 guidelines from the Scottish Intercollegiate Guidelines Network (SIGN).

	POS	ACT	COR	INC	MIS	PAR	REC (%)	PRE (%)
Acronym Definition	6	3	3	0	3	0	50,00%	100,00%
Acronym	6	3	2	1	4	0	33,33%	66,67%
Hypernym/Hyponym	254	324	221	103	33	0	87,01%	68,21%
Overall	266	330	226	104	40	0	84,96%	68,49%

Our coreference resolution algorithm reaches an overall performance of 84,96% in recall and 68,49% in precision.
The reduction in recall can be explained by a higher complexity of the test CPGs compared with the documents used in the training stage. Furthermore, the gold standard was sometimes not as accurate as desired. The missing precision score mostly depends on the complexity of the UMLS Metathesaurus and the enormous number of relations that are defined between its concepts that are sometimes not comprehensible for human experts.

Nevertheless, these are promising results that form an important basis for further automated processing of CPG documents.