Understanding Medical Data: Text Analysis and Coding with Semfinder Expert System

Jörg Niggemann, Michael Owsijewitsch, Hans Rudolf Straub
Semfinder AG, Kreuzlingen, Switzerland www.semfinder.com straub@semfinder.com

Swisstext 2016
1st Swiss Text Analytics Conference

Challenge Unstructured information represents the largest and most relevant source of information in hospitals

The Result Software solution in daily use for coding:
- 500+ hospitals in Germany, Switzerland and Spain
- Integration in IBM Insurance Service Hub for analysis of 2 Mio diagnoses / month in Germany

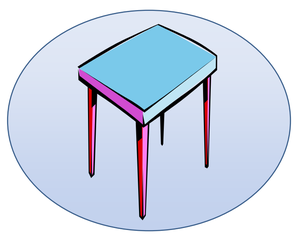
Sources of complexity in medical free texts
- → Overlappings (streptococcal pneumonia, postoperative)
- → Ambiguities (heads in the shoulders and abdomen)
- → Negations (diabetes, non-insulin-dependent, with complications)
- → Non-information (diabetes, not otherwise specified)
- → Implications (radius → bone, forearm)
- → Omissions (fracture of humerus and radius)
- → Composite Diagnoses with mutual dependencies

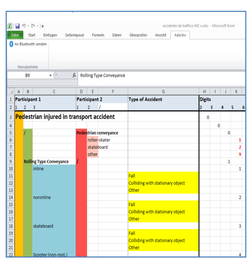
Semantics: words ≠ concepts

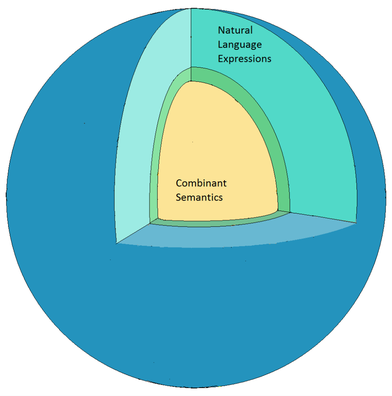
Typical for semantics:
→ Overlapping of concepts
→ Implicit concepts
→ Resolvable Ambiguities

Implications:
Fracture → Bone
Fracture → Diagnosis
Fracture → Injury

Ambiguities:
English: Table → German: Tisch
English: Table → German: Tabelle







Semantics is the core

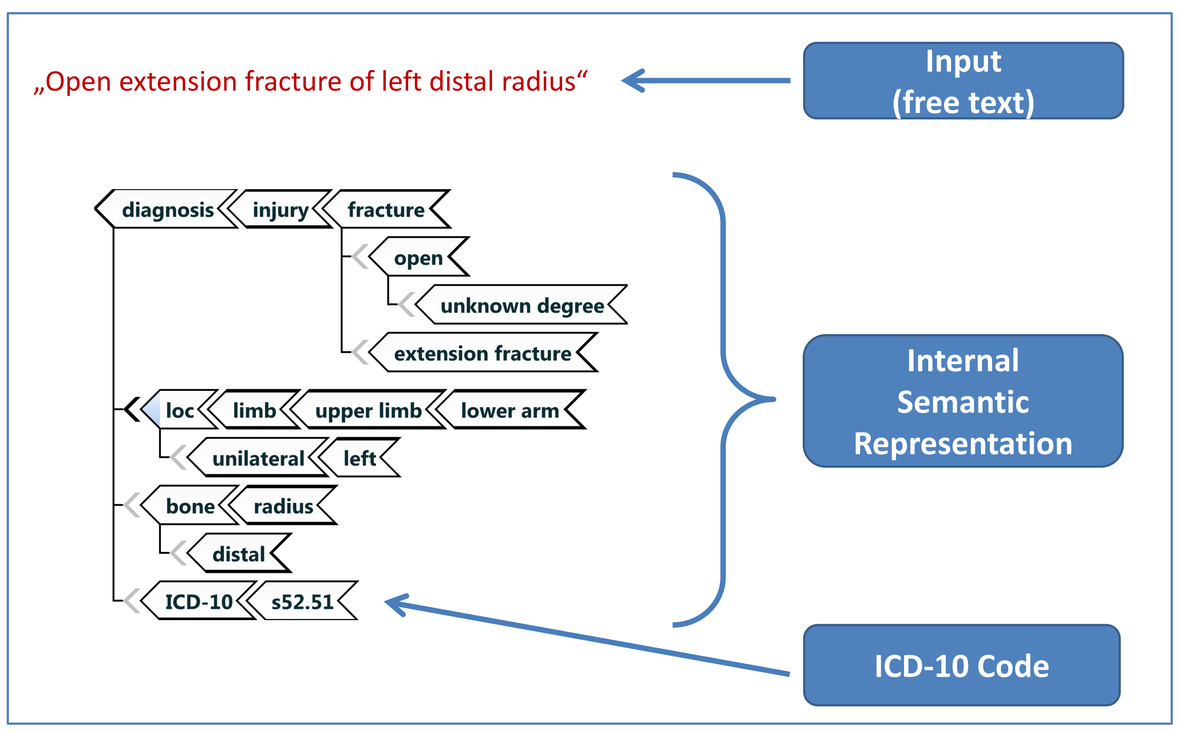
Language is the outside:
- what you hear and see

Semantics is the inside:
- what you think

The semantic core of Semfinder is language independent:
→ reusability of knowledge bases in different languages
→ interoperability between languages

Concept Molecules

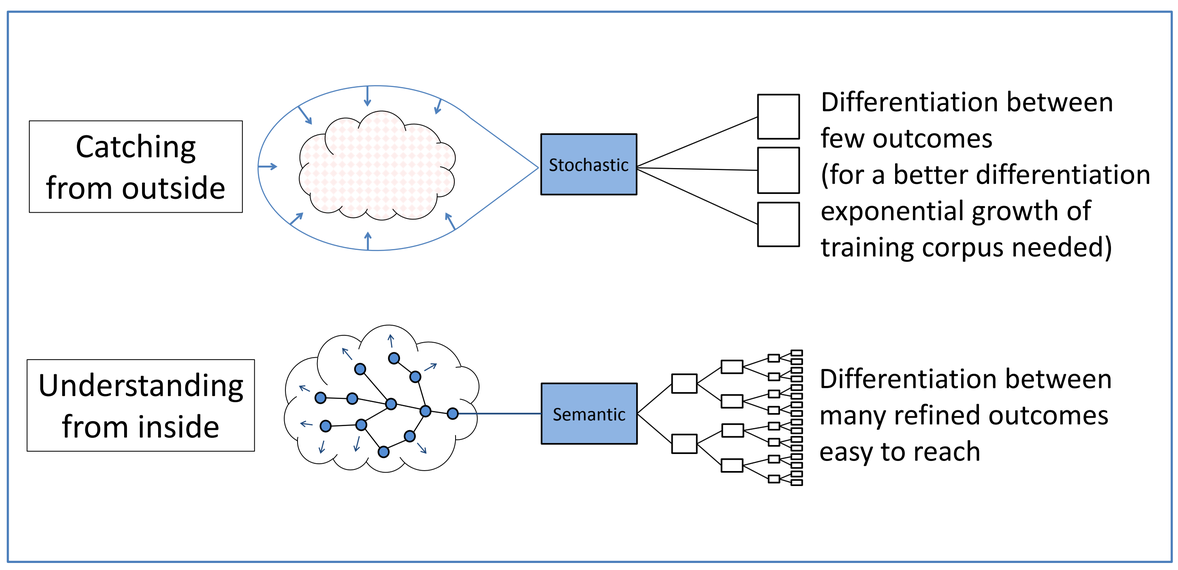
Semfinder technology works with concept molecules (CMs) which capture the semantics.
For coding, free text inputs are automatically processed to CMs. CMs are built of atomic concepts, arranged in a structure which represents the relations between the atomic concepts.
The actual coding is derived from the information in the molecule.



Noun phrase ≠ diagnosis phrase (solved by Concept Molecules)	
Linguistics (Input phrase)	Semantics (Interpretation)
Adeno-CA, Colon	diagnosis › neoplasia › carcinoma › adenocarcinoma; malignant; ICD-10 › c18.9; loc › intestine › large intestine › colon — 1 molecule, 1 diagnosis, 1 ICD-10 code
Adeno-CA, Tinnitus	diagnosis › neoplasia › carcinoma › adenocarcinoma; malignant; ICD-10 › c80.9; loc; diagnosis › tinnitus; ICD-10 › h93.1; loc — 2 molecules, 2 diagnoses, 2 ICD-10 codes
Linguistically no difference	Semantically a clear difference

Stochastics and Semantics

Semantic and stochastic methods are complementary in nature.



	Stochastics	Semantics
Learning Phase	Long (much data)	Long (much expert work)
Recall	High	Very high
Noise	Robust	Sensitive (LDE, Vocabulary)
Precision	Medium to High	Very high
Outcomes	Few, simple	Detailed, rich
Multilinguality	Expensive	Easy access
Process transparency for maintenance /fine tuning	not transparent	transparent
Further Processing (apart from coding)	Needs prior work	Ready: → Semantic Data Repository →Alerts, Proposals in Clinics →Clinical Epidemiology

Semfinder