Aleksandar Savkov

EXTRACTING INFORMATION FROM GP NOTES TEXT

PhD Poster

HOW? create a scalable system for identification and classification of medical observations in patient records

WHY? provide access to unstructured free text data from the vast resources of the CPRD for the purposes of epidemiology research

Data Processing Flow Plan

Natural Language Processing	Classification	Clustering	Observation Classes
Processing steps • tokenization with Dr.Tokenizer • part-of-speech tagging with RASP • chunking with YamCha	**Identifying Observations** • assume base noun phrase chunks to be observation candidates • use information from all chunk types to model features • use machine learning classifier to identify observations	**Classifying Observations** • cluster observations • use specific patterns as features for each class • find the best matches of each class among the clusters	present observation absent observation possible observation conditional observation family history of... history of...
FREE TEXT	CHUNKS	OBSERVATIONS	

Natural Language Processing

NATURAL LANGUAGE PROCESSING

GOAL create reusable annotation guidelines and produce an annotated text corpus of medical records

MOTIVATION identify text units that could be medical observations and produce linguistic patterns to aid classification

Clinical Practice Research Datalink

- the CPRD (formerly the GPRD) is an electronic patient records database that was started in the early 90s
- each record contains structured data and free text regarding a visit or an event that the general practitioner (GP) has logged
- data needs to be manually anonymised before being released to third parties

Free Text Issues

- telegraphic style of expression
- professional slang: words and acronyms
- irregular and inconsistent abbreviations
- idiosyncratic use of symbols
- missing and confusing punctuation
- non-grammatical expressions
- spelling mistakes and typos

Chunks

- *base noun phrase chunks* include a head noun and a number of its modifiers on its left (adjectives, definite articles, etc.)
- *adjectival and adverbial phrase chunks* include adjectival phrases that do not modify a noun directly and adverbs that modify certain intransitive verbs
- *the main verb* in a clause
- *on-examination marker* used by GPs to separate patient complaints from observations
- *quantitative expressions* except the ones referring to time
- *locative expressions* mainly referring to body parts
- *temporal expressions* referring to dates, periods, and repetitions

Observations

Observations are statements that GPs record in their notes. *Examples:*

- *presence*: high fever
- *absence*: no coughing
- *possibility*: possible cancer
- *condition*: difficulty breathing on getting up
- *history*: history of respiratory diseases
- *family history*: mother had Parkinson's

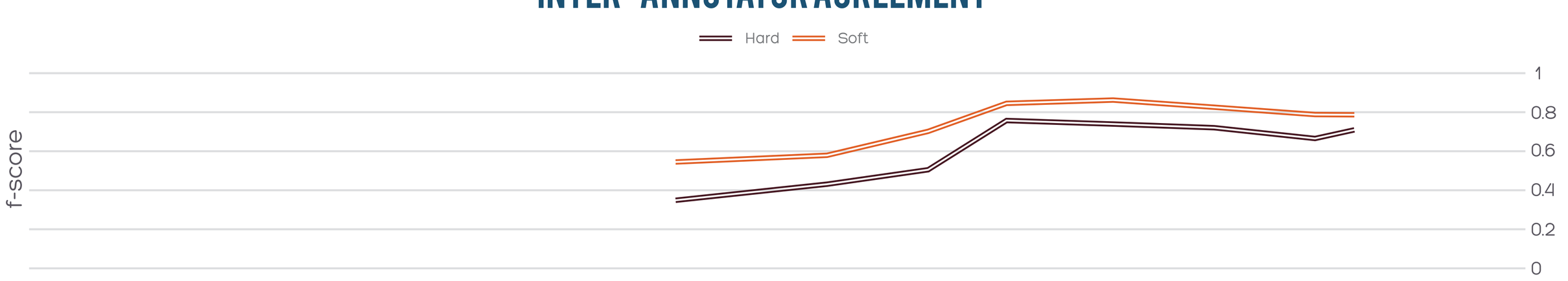
Agreement Calculation

Inter-annotator agreement is calculated using the standard MUC7 evaluation scheme:

- *soft comparison*: *Precision* and *Recall* are calculated based on matching tags
- *hard comparison*: *Precision* and *Recall* are calculated based on matching tags and tag borders

Agreement

INTER-ANNOTATOR AGREEMENT



Annotation

ANNOTATION TIMELINE