Enhancing Medical Named Entity Recognition

with Features Derived from Unsupervised Methods

Maria Skeppstedt, mariask@dsv.su.se

Creating the annotated corpus for training a named entity recognition model is expensive, particularly in specialised domains, such as medicine, which require expert annotators. Moreover, a model trained on text from one medical sub-domain often shows a drop in performance when applied on texts from another sub-domain, and annotated text from this other sub-domain might be required.

When incorporating features from unsupervised methods, to what extent is it possible to:

- Reduce the amount of annotated data needed to achieve a fixed level of performance?
- Reduce the amount of additional annotated data needed for adapting a model to a new sub-domain?

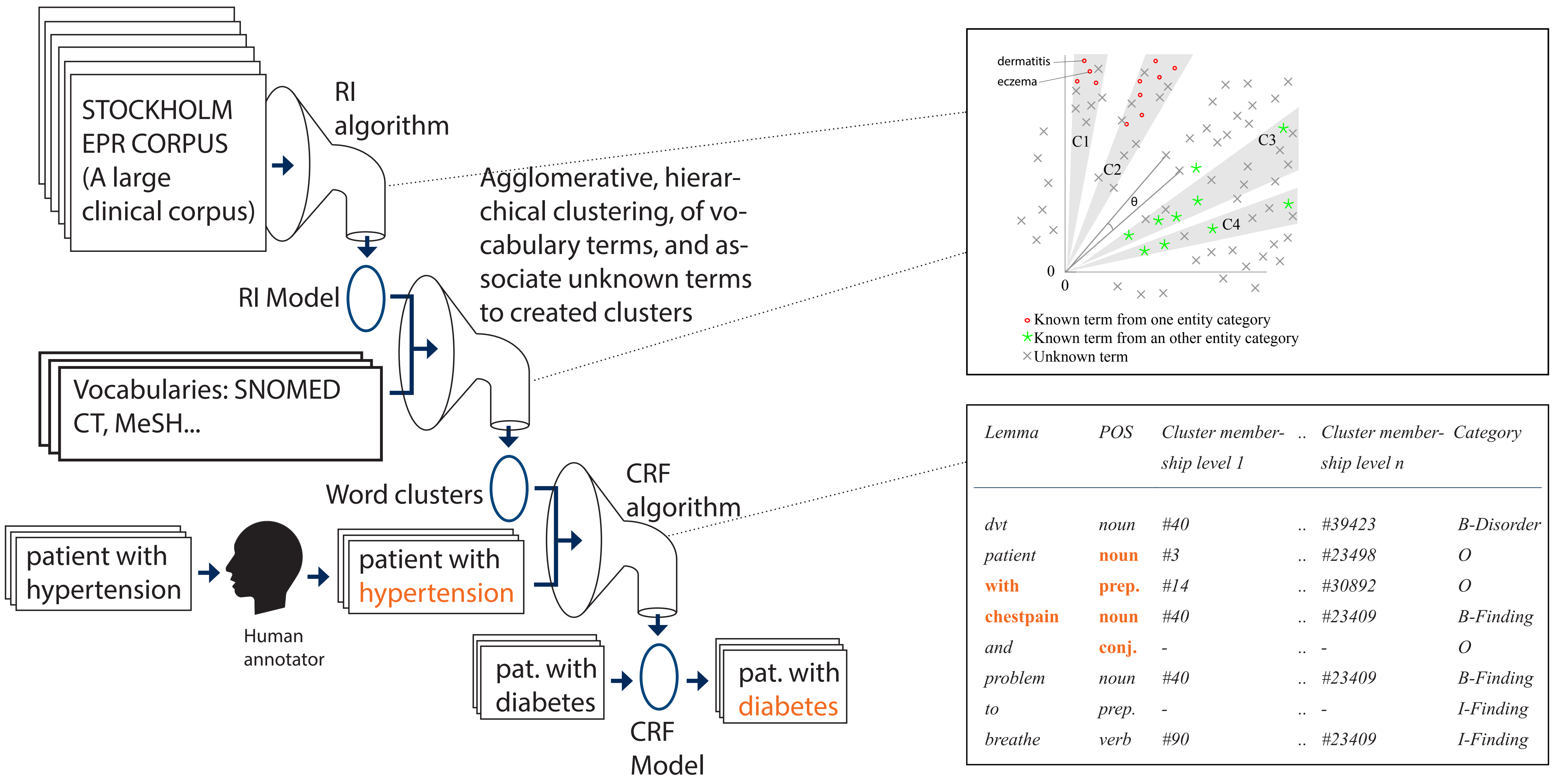
Data sets

The entity types disorder, finding, pharmaceutical and body structure annotated in texts from three medical subdomains:

- Internal medicine ER (i.m.)
- Cardiac ICU (Cardiac)
- Orthopaedic ER (Orthop.)

Internal medicine ER data is divided into training data and evaluation data.

Using clustering features for training a CRF model


STOCKHOLM EPR CORPUS (A large clinical corpus)
RI algorithm
RI Model
Vocabularies: SNOMED CT, MeSH...
Word clusters
CRF algorithm
patient with hypertension
Human annotator
patient with hypertension
pat. with diabetes
CRF Model
pat. with diabetes
dermatitis
eczema
C1
C2
C3
C4
θ
0
0
○ Known term from one entity category
✱ Known term from an other entity category
× Unknown term

Agglomerative, hierarchical clustering, of vocabulary terms, and associate unknown terms to created clusters

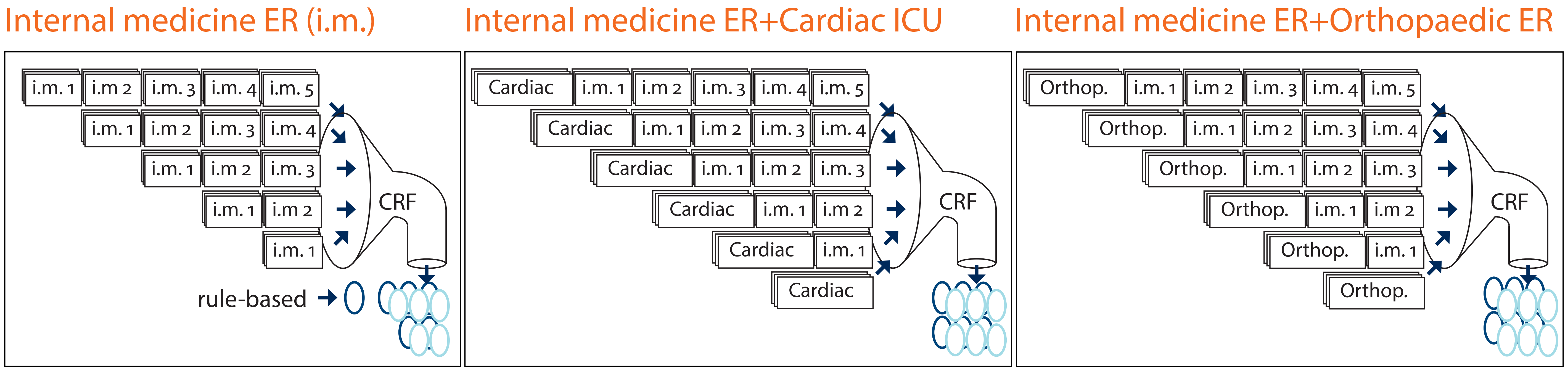
Lemma	POS	Cluster membership level 1	..	Cluster membership level n	Category
dvt	noun	#40	..	#39423	B-Disorder
patient	**noun**	#3	..	#23498	O
with	**prep.**	#14	..	#30892	O
chestpain	**noun**	#40	..	#23409	B-Finding
and	**conj.**	-	..	-	O
problem	noun	#40	..	#23409	B-Finding
to	prep.	-	..	-	I-Finding
breathe	verb	#90	..	#23409	I-Finding

Experimental setup

Internal medicine ER training data is divided into 5 partitions, and increasingly more data is used when training the model.

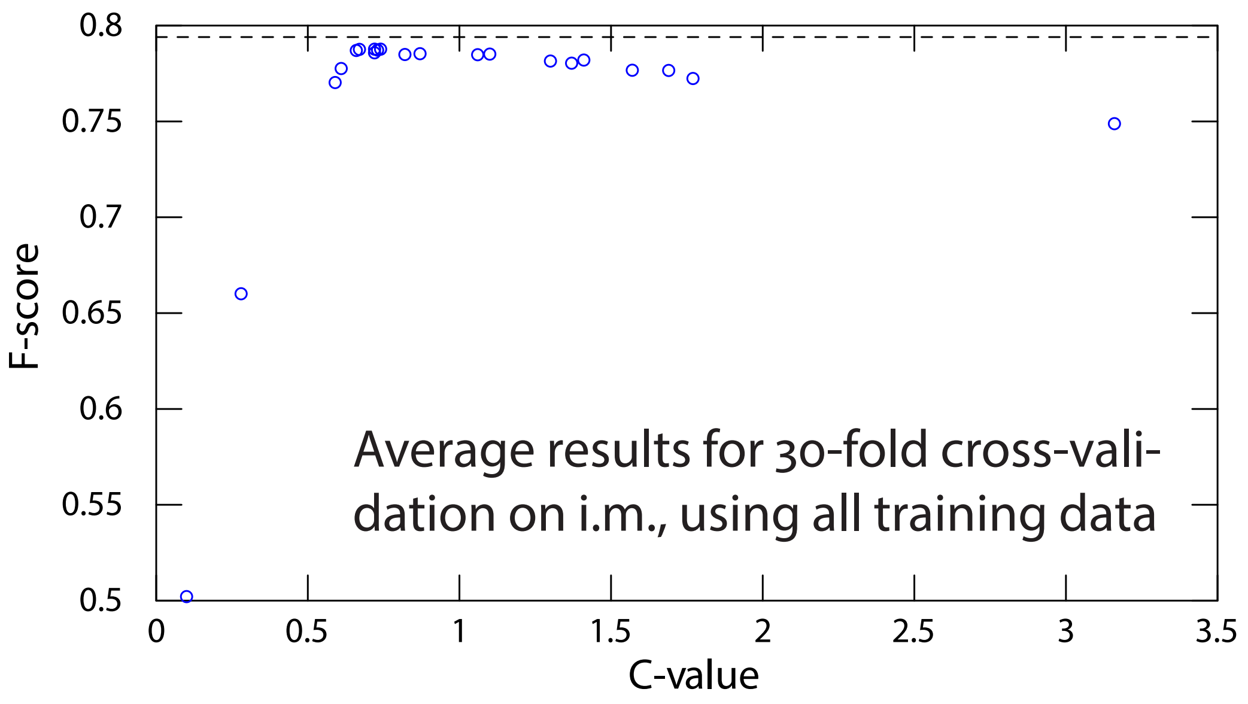
For each created model:

- One version with cluster features
- One version without cluster features


Internal medicine ER (i.m.)

Internal medicine ER+Cardiac ICU

Internal medicine ER+Orthopaedic ER



Average results for 30-fold cross-validation on i.m., using all training data

The L1-norm is used for regularisation when setting the weights of the CRF model. n-fold cross-validation is used on the training data to determine the C-value which governs the strength of the regularisation.

The created models are evaluated against the Internal medicine ER evaluation data.