

Cross-domain and cross-language super sense tagging



Anders Johannsen

Hèctor Martínez Alonso

Anders Søgaard

University of
Copenhagen

Layers of semantic annotation

NE	B	I	I	I	O	O	O
LYONS	1st	1st	1st	1st	(1st)	-	2nd
SUPER	n.person	?	?	n.quantity	v.motion	-	n.act
WS	pope#1	?	?	two#1	head#3	-	procession#2
	Pope	John	Paul	II	headed	the	procession



English noun supersenses

act	acts or actions
animal	animals
artifact	man-made objects
attribute	attributes of people and objects
body	body parts
cognition	cognitive processes and contents
communication	communicative processes and contents
event	natural events
feeling	feelings and emotions
food	foods and drinks

... 12 senses skipped ...

state	stable states of affairs
substance	substances
time	time and temporal relations
Tops	abstract terms for unique beginners

26 noun supersenses

15 verb supersenses

Senses of *tag* (noun)

* a label attached to something to indicate its owner, nature, price, etc.

* a label associated with something for the purpose of identification

COMMUNICATION

* a small piece of cloth or paper

ARTIFACT

* a game in which one child chases the others; the one who is caught becomes the next chaser

* the act of touching a player in a game

ACT

ANIMAL
ARTIFACT

Semantic distinctions
in supersenses

The cranes left
at the onset of the building crisis

RELATION
ARTIFACT

He burned the bridge
along with the rest of the ship

CREATION

They fabricated the data
for the synthetic experiments

Overview

1.Super-sense tagging on English Twitter

2.Tagging on Danish

2.1.Across domains

2.2.Across languages

3.A note about active learning

More or less supervised super-sense tagging of Twitter

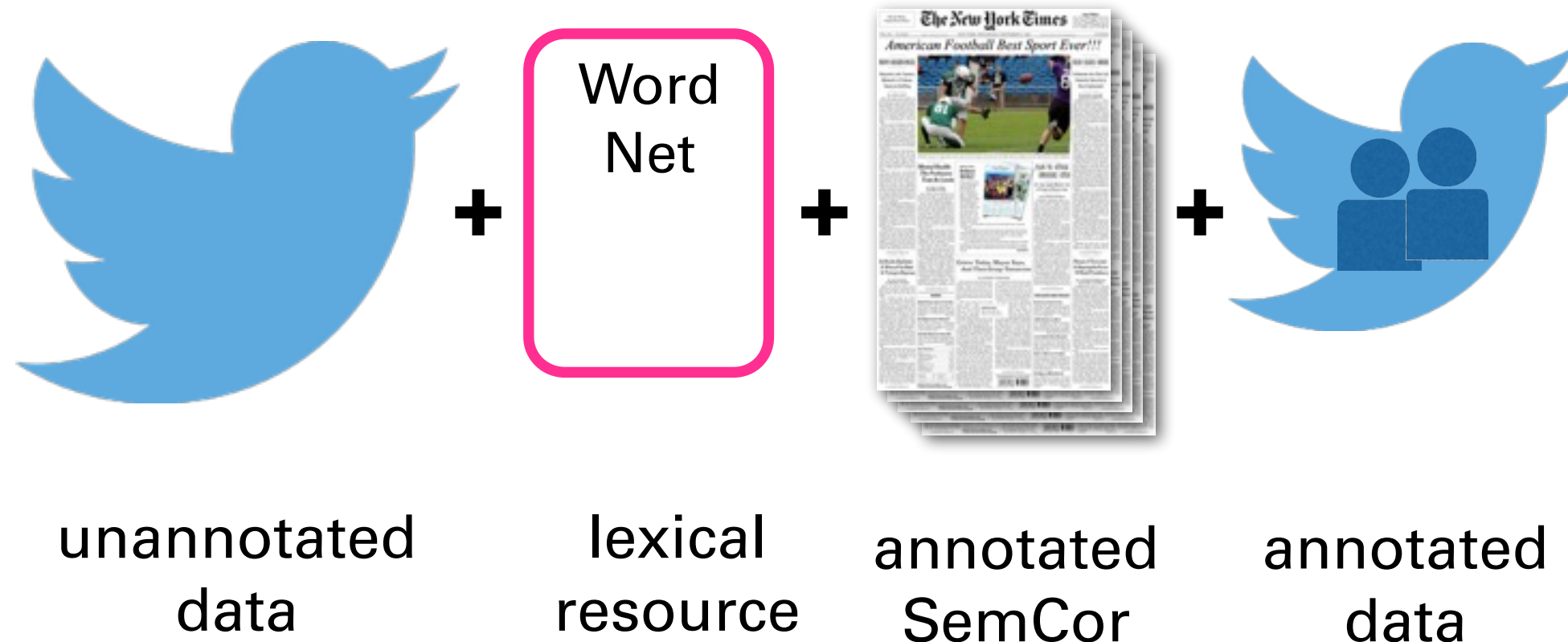
Anders Johannsen, Dirk Hovy, Hector Martinez, and Anders Søgaard (2014)



Supervised domain adaptation

Unsupervised domain adaption

Weakly supervised



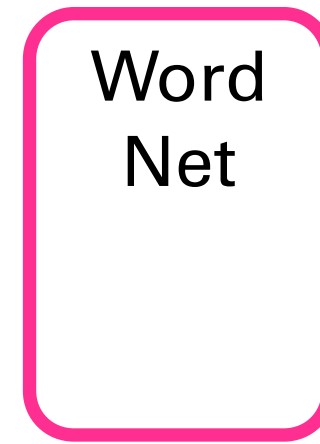


unannotated
data



distributed word
representations

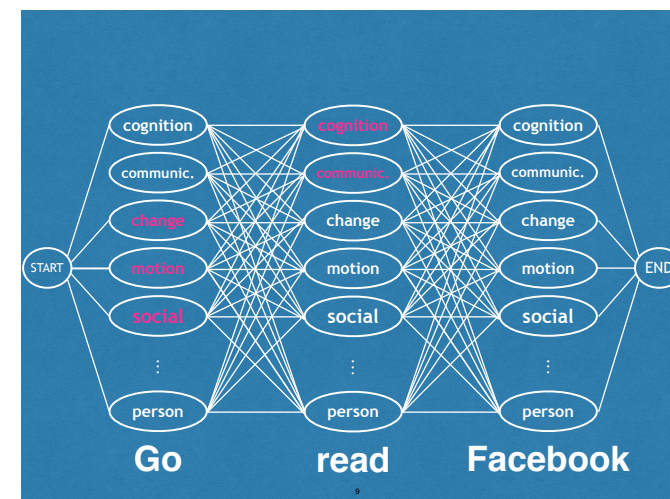
words



lexical
resource



type constraints
in decoding



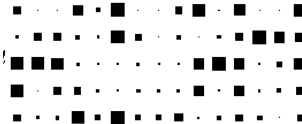

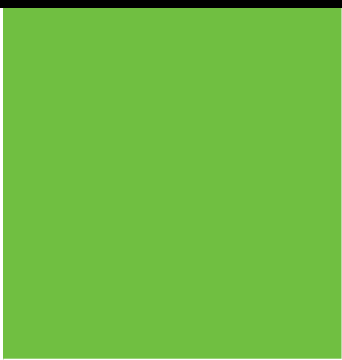
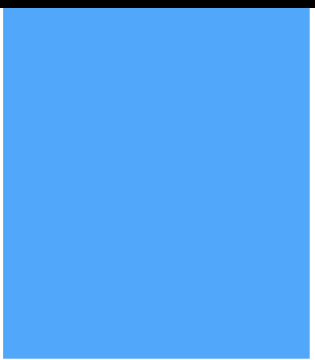
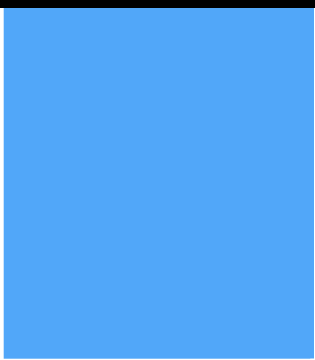

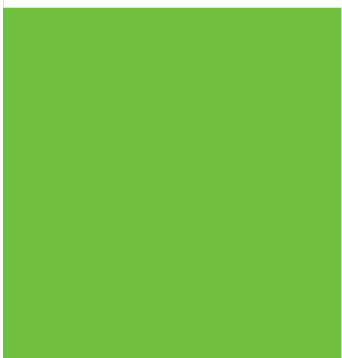
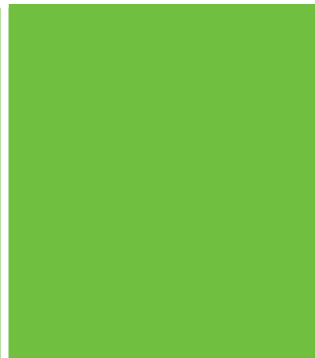
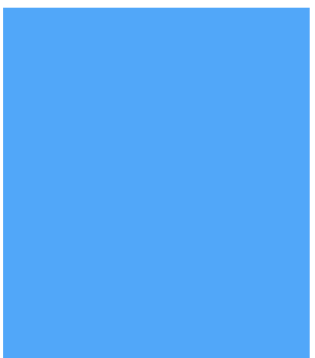
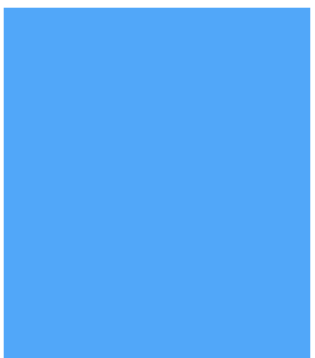
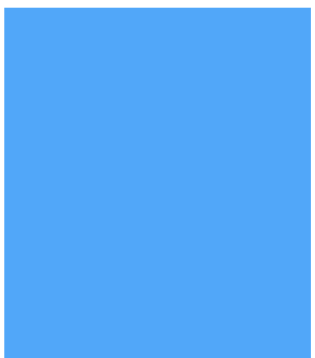

Experiments on Danish

1. Can the same methodology be applied for Danish?
2. Can we *directly* use the labeled English resources?

Danish annotations

Domain	Sentences	Tokens
Blog	100	1.744
Magazine	200	4.095
Forum	200	4.302
News wire	600	11.081
Parlament	200	6.442
Chat	200	4.302
Total	1.500	31.966

Cross-language features

	X-language 	Target 	Word shape	Universal POS	Lexical forms	MFS supersense
English						
Danish						

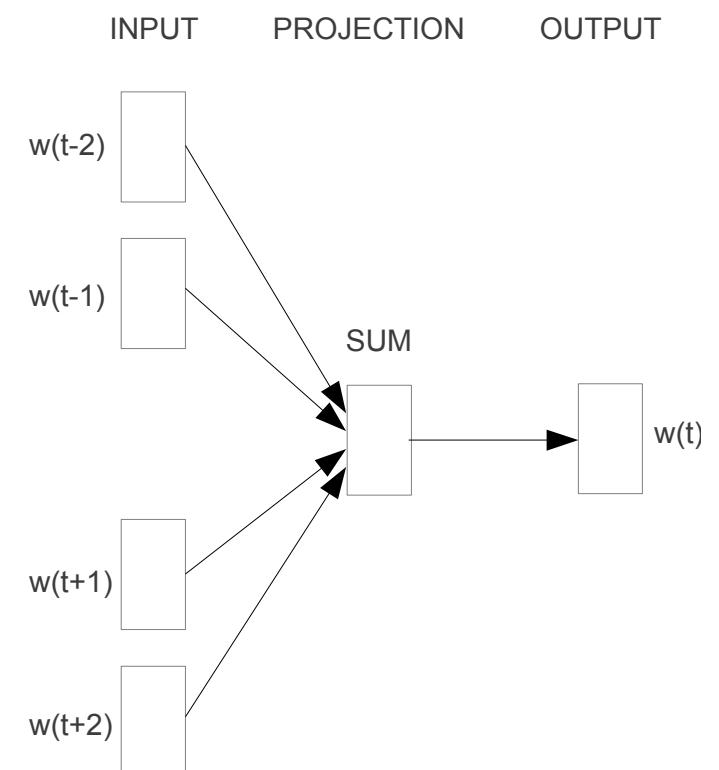
Cross-language word representations

#1 Generate mixed language training data

EN	The	queen	has	only	one	job	-	to	lay	eggs
DA	Den	dronning	har	kun	en	arbejde	-	at	lægge	æg
Mixed	Den	queen	has	kun	one	arbejde	-	to	lay	æg

#2 Estimate continuous bag-of words model

Den queen has kun one arbejde - to lay æg



Experimental setups

#1 Danish to Danish

Train on Danish newswire.
Test across six Danish domains.

#2 Cross-language

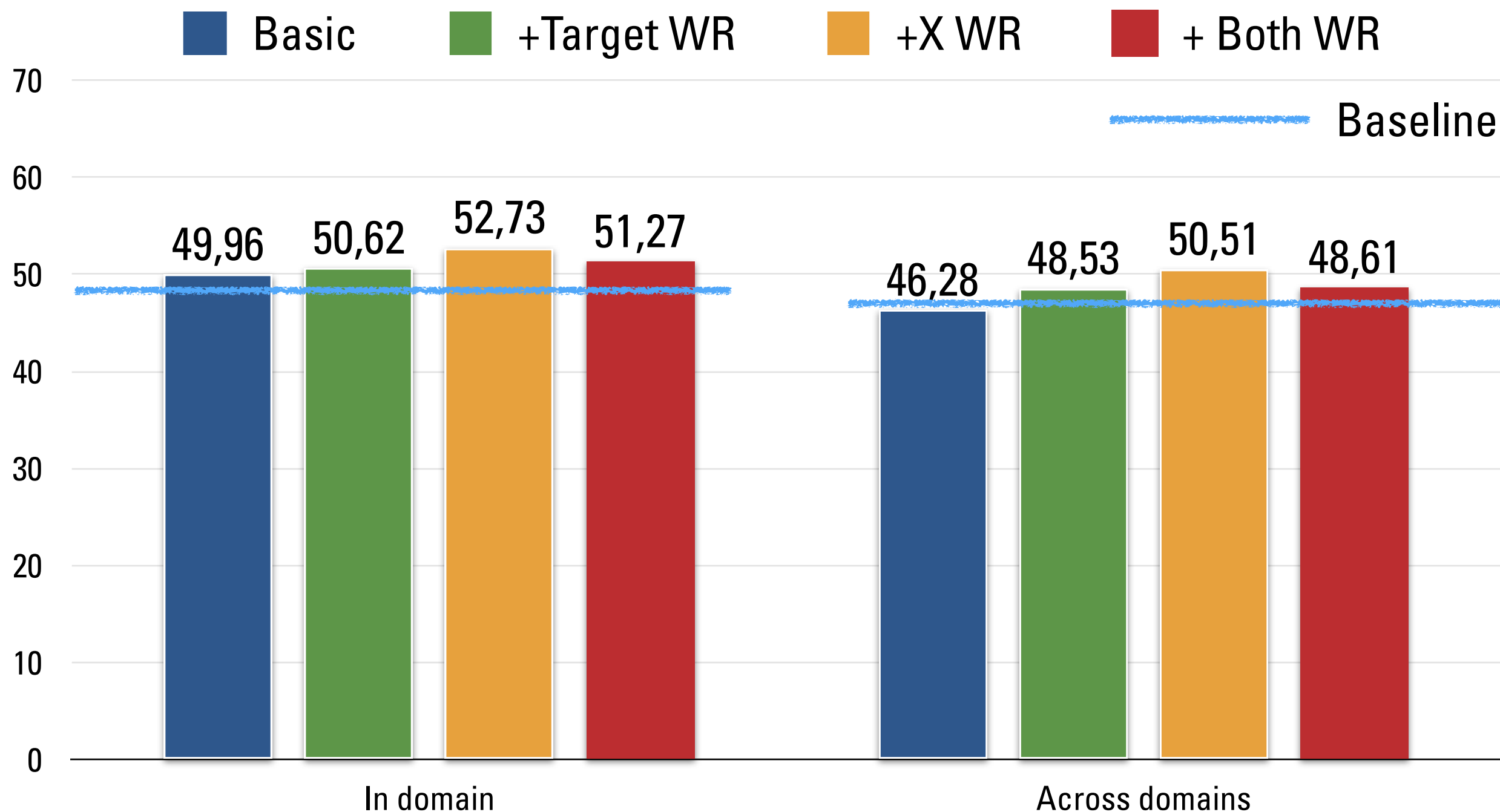
Train on English SEMCOR.
Test across six Danish domains

#3 Mixed training

Train on English SEMCOR
and Danish newswire.
Test across six Danish domains

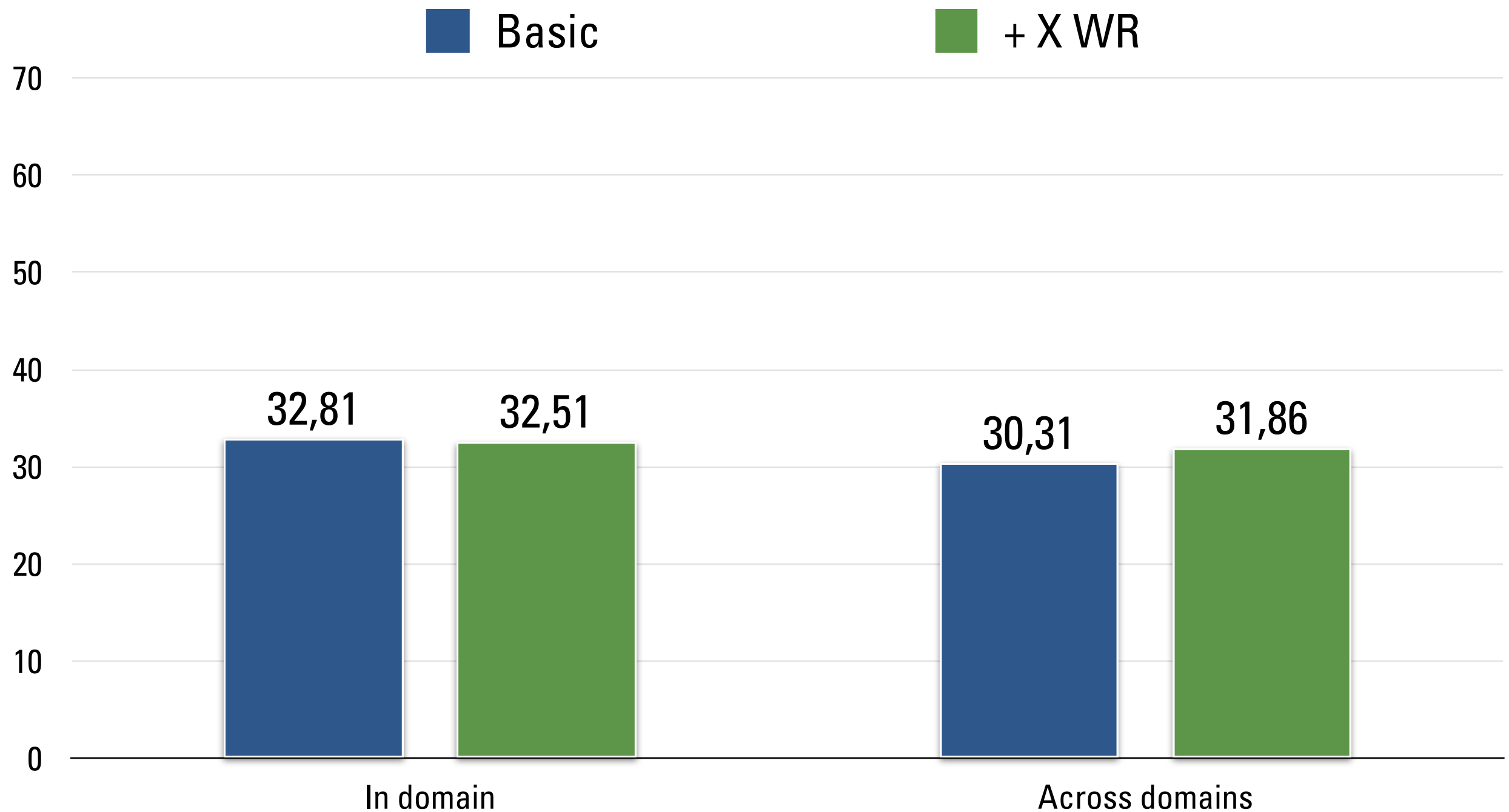
Results

#1 Danish to Danish
Train on Danish newswire.
Test across six Danish
domains.



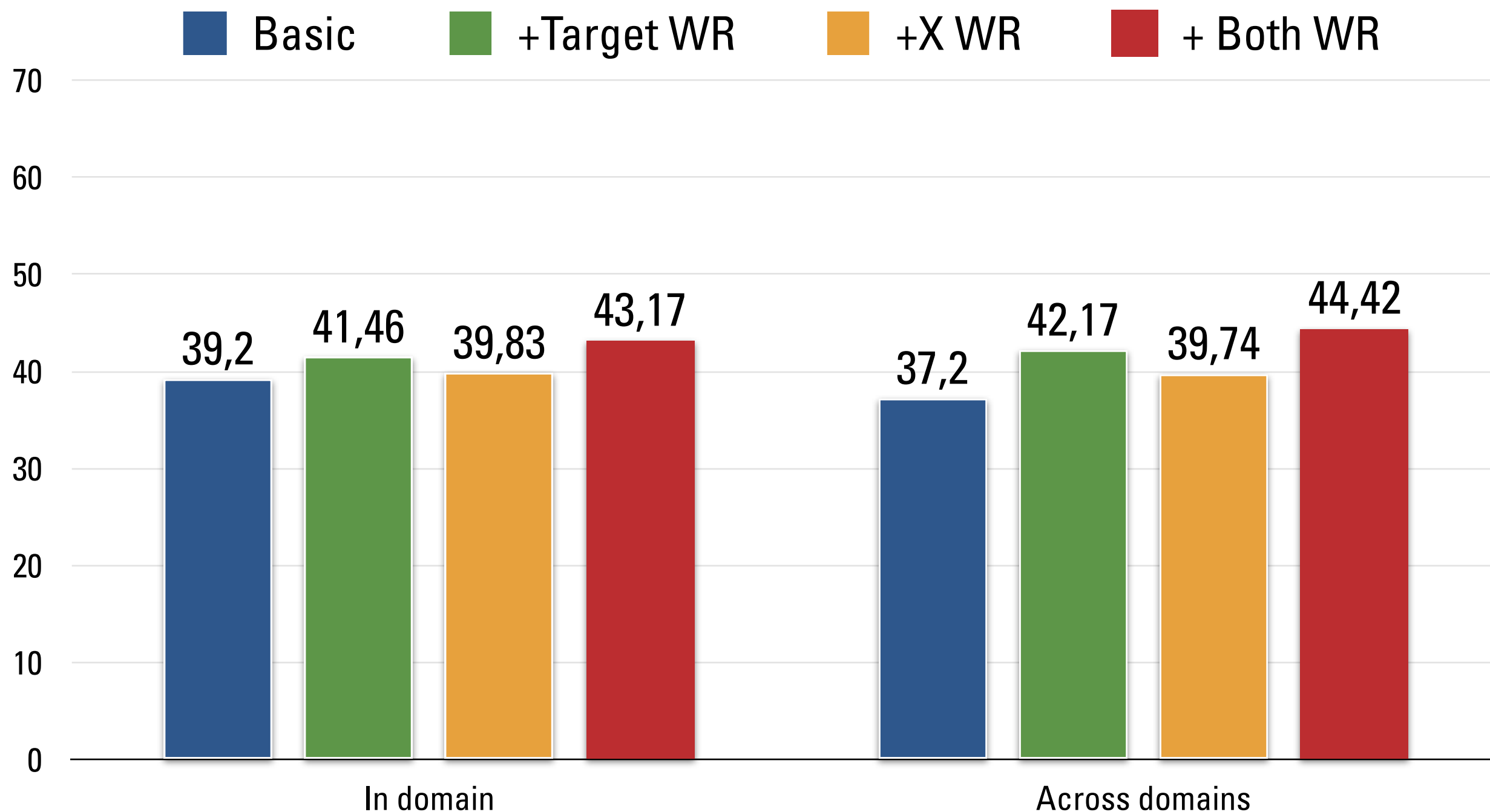
Results

#2 Cross-language
Train on English SEMCOR.
Test across six Danish
domains



Results

#3 Mixed training
Train on English SEMCOR
and Danish newswire.
Test across six Danish
domains

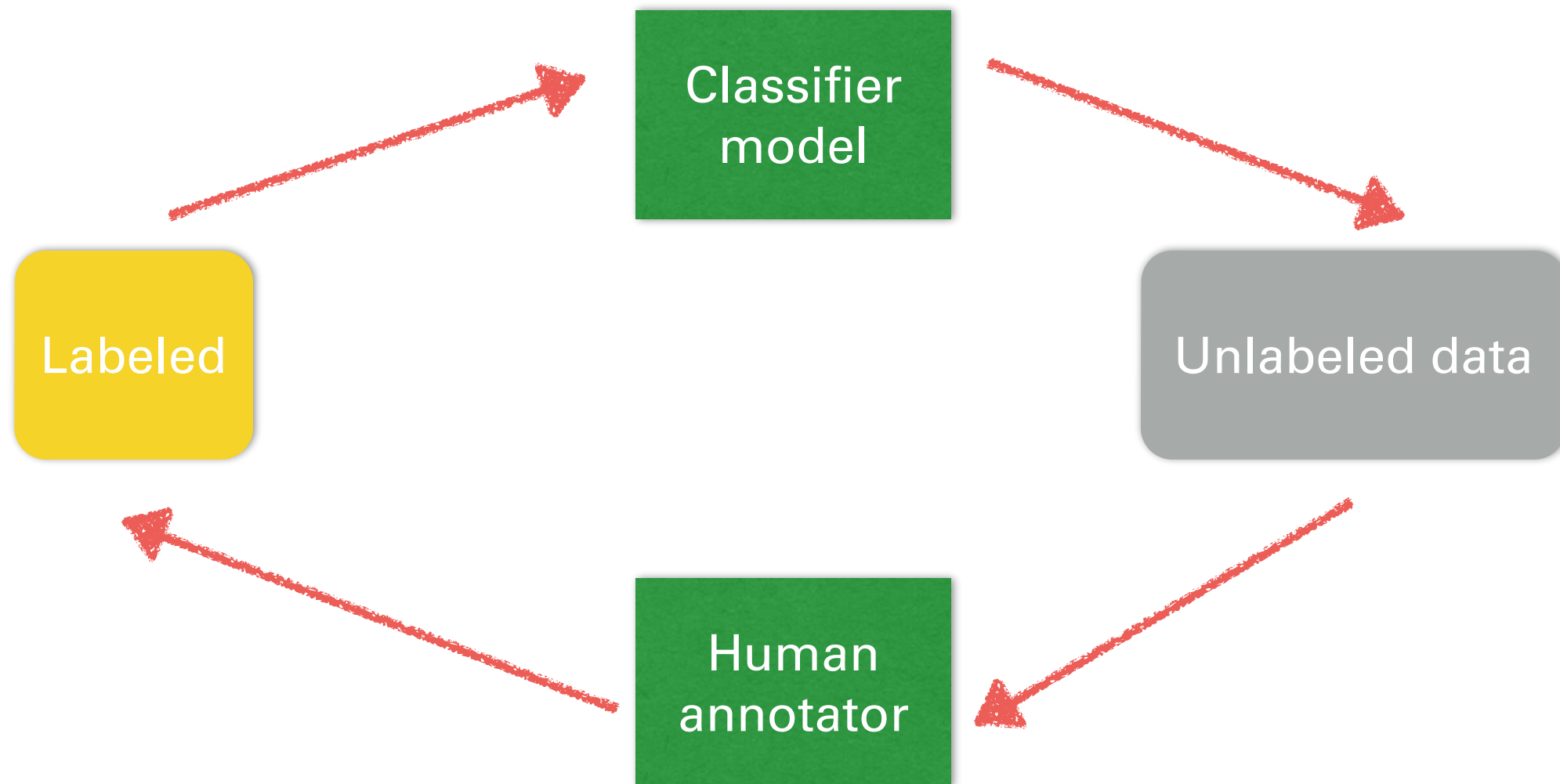


Conclusion

- Our methodology transfers to Danish:
- Constrained decoding improves $\sim 2.5\%$
- Cross-language word representations provide (some) signal for SST

Active learning note

Time permitting



Rationale

- Increase robustness of final model
- Sample more varied data
- Speed-up annotation process (after agreement has converged)

Method

- Using the SST method for Danish
- Compare two instance-selection strategies:
 - Lowest-confidence instance
 - Sampling from the classifier confidence distribution

Running!

- Labeled data: Newswire train section
- Unlabeled pool: ClarinDK
- Model: Same as above without embeddings
- Currently annotated ~100 sentences

Questions?