[SIAN GOODING, EKATERINA KOCHMAR, ALAN BLACKWELL, ADVAITH SAKAR]

Comparative judgements are more consistent than binary classification for labelling word complexity

kage

ncu ited

le

MOTIVATION

muja

or

pc

ling

eni

nt

luct

-hit

inte

S7

afft

۲

neti

Lexical simplification (LS) systems replace complex words with simpler alternatives.

They aimed to **<u>ameliorate</u>** the situation.

They aimed to **improve** the situation.

<u>**Complex word identification**</u> (CWI) is a sub-task of LS concerned with the detection of words in need of simplification. Current datasets for this task have low levels of annotator agreement:

Data	IAA Statistic	Interpretation					
2012	$\kappa = 0.386, 0.398$	minimal agreement					
2016	$\alpha = 0.244$	inconclusive					
2018	1% unanimous	idiosyncratic					

Current drawbacks:

- Homogeneity of the annotator group is usually not controlled for
- Represented as a binary, not continuous task

spificant atmosphere rededicated STUDY

➢ 30 annotators

- ➤ 20 sentences
- ~25 minutes per participant



Annotator group:

Same first language (English)

Cambridge ALTA

UNIVERSITY OF

uste

nfi

tect

orv

 m_{i}

nda

coll

m

adamantly

ycc in

meth

ung

riposi

delegati

overextended

- Same level of educational background
- Similar age range (21-30)

Annotators were presented with professionally written news sentences from Yimam et al. (2017) dataset:

- > 10 sentences presented per interface
- Chosen to contain a range of word complexities based on the number of annotations from Yimam et al. (2017):

$hard \in [10, 20]$	politicizing (14)					
$medium \in [6, 9]$	warily (9)					
$low \in [1, 5]$	<i>trip</i> (2)					

RESULTS

Kappa Coefficient

Comparative Judgement 0.6775

Binary Judgement 0.3937 (minimal)

cease-fire HYPOTHESIS

Do comparative judgments for CWI lead to higher inter-annotator agreement and higher quality labelled data than binary?

Additional Questions

- Does controlling for the homogeneity of the group of annotators contribute to higher agreement?
- Can comparative judgments be made in a significantly shorter period of time than binary judgments for word complexity?

INTERFACE

Within-subjects design used; annotators were asked to label word complexity in continuous vs binary fashion:

Task 1 INSTRUCTIONS

Assume the following sentences are meant for non-native language learners, children, or people with disabilities.

Please mark words that you think would be hard to understand by clicking them.

Select at least 3 words per section.

Reg	ime	for	ces	lau	unc	hed	an	offensi	ve	on	Rast	an	at	the
weel	kend	bu	tm	et w	ith	shar	p re	esistanc	e fr	om	rebe	ls	see	eking
					1									
the	oust	er	of	the	re	gime	of	Presid	ent	Ba	shar	al-	As	sad.
									,					
											Click for	r nex	t sen	tence

Binary Judgement Interface

 Alpha Coefficient
 0.6821
 0.4960

 Avg Time (s)
 28.77
 38.69

(moderate)

According to Cohen (1968), our Kappa results indicate *moderate* agreement for comparative judgements and *minimal* for the binary annotation task supporting our hypothesis

Binary judgement :

- 62 distinct words from 10 sentences marked as complex by annotators
- Higher agreement than previously reported studies: α = 0.496 vs α = 0.244 in Paetzold and Specia (2016)

Comparative judgement:

- → Higher agreement than previously reported studies: $\kappa = 0.6775$ vs $\kappa = 0.398$ in Specia et al. (2012)
- 9.92s less time per sentence on average than binary judgements

at CONCLUSIONS

sali

 \mathbf{P}

sorv

resented ^{ar}

spectrophotometer dehydration

assumption

am

- This study demonstrates the advantage of annotating datasets using *comparative judgments rather than binary classifications*, both for *efficiency* and *accuracy*.
- Our results also show *higher agreement coefficients* for both binary and relative judgment tasks when compared to previously collected datasets.
- Our work supports the case that the concept of word

On

um

citie

rtem

militants

Task 2 INSTRUCTIONS

Assume the following sentences are meant for non-native language learners, children, or people with disabilities.

Please rank the highlighted words according to their complexity by typing them in the below boxes, if equal complexity place words next to each other.

Most Comple	ex [Ouster								
		Regime								
		offensive		resistance						
Least Comple	ex 🛛									
	[
		Regime for	ces	a launched	an	offensive	on	Rastan	at	the
		weekend but	m	et with shar	p re	esistance fr	om	rebels	see	king
ck for next			-				_		_	
entence		the ouster	of	the regime	ot	President	Bas	shar al	-Ass	sad.
	L									

Continuous Judgement Interface

complexity, and thus the level of agreement, is *aligned between individuals that share a common background*.

Future steps for this research include:

perpetrated

nythalagias

- more thorough investigation of effects of annotator group homogeneity on the inter-annotator agreement
 more detailed larger study of the efficiency of the comparative judgments as opposed to binary judgments
- Our results are applicable to other natural language tasks where specific user experiences like simplicity can be modelled as an ordering so that they may be optimized or personalized.

[shg36, ek358, afb21] @cam.ac.uk, advaith@microsoft.com

snipers

CONTACT INFORMATION

Intensifieo

Sian Gooding, Ekaterina Kochmar, Alan Blackwell, Advaith Sakar