

Dealing with Inconsistencies in Parallel Annotations

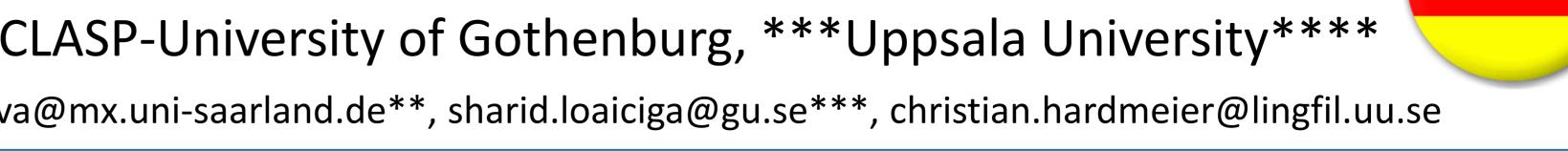




Marie-Pauline Krielke*, Ekaterina Lapshinova-Koltunski**, Sharid Loáiciga*** Christian Hardmeier****

Saarland University*, Saarland University**, CLASP-University of Gothenburg, ***Uppsala University***

mariepauline.krielke@uni-saarland.de*, e.lapshinova@mx.uni-saarland.de**, sharid.loaiciga@gu.se***, christian.hardmeier@lingfil.uu.se



AIMS AND MOTIVATION

Manually annotated parallel data are prone to inconsistencies of various kinds:

→ Inconsistencies are **interesting indicators** of:

Annotation errors linguistic contrasts translation problems Develop an automatical method for detecting annotation inconsitestencies in parallel data

for:

- Detection and correction of errors
- Detection of cohesively ambiguous situations
- Detection of language contrastive patterns in cohesion

ParCor Full (Lapshinova-Koltunski, Hardmeier, Krielke 2018):

Parallel corpus Annotated for FULL coreference chains Spoken 87%, written 13%.

http://hdl.handle.net/11372/LRT-2614

				English	German	total
			pronoun	4,650	4,269	8,919
Total no of tokens	82,379	78,350	np	2,485	2,611	5,096
Total no of chains	2,319	2,425	vp	133	132	265
Avorago chain langth	2.94	2.81	clause	335	312	647
Average chain length	2.34		mentions	7,603	7,324	14,927

TYPES OF

Linguistic

INCONSISTENCIES

language-specific constraints translation-driven transformation patterns

[" Selbst aus ganz un<u>religiöser Sicht ate</u>fft [die Homosexualität] eine sexuelle Verfehlung dar .

[Sie] ist ein armseliger , minderwertiger Ersatz für die Realität - ein kläglicher Versuch , dem Leben zu entfliehen . Als solches verdient [sie] kein Mitleid , keine Sonderbehandlung als Märtyrium einer Minderheit , und sollte nur als das betrachtet werden , was [sie] wirklich ist : eine schädliche Krankheit . "] [Das] schrieb Time Magazine 1966, als ich drei Jahre alt war .

deemed anything but a pernicious sickness . "]

It is a pathetic little second-rate substitute for reality - a pitiable flight from life . As such, [it] deserves no compassion, [it] deserves no treatment as minority martyrdom, and [it] deserves not to be

[" Even in purely non-religious terms", [homosexuality] represents a misuse of the sexual faculty .

[That] 's from Time magazine in 1966, when I was three years old.

Cognitive

Misinterpretation of cohesive relations

- By translator
- By annotaator

Plain errors

dass [es] funktioniere . Und weil man [diese 99 %] verbrennt , ist das Kostenprofil sehr viel besser . Tatsächlich verbrennt man den Müll., und man kann sogaf [den Abfall von heutigen Reaktoren als Antriebsstoff] benutzen . Anstatt [sich] [darüber] den Kopf zu zerbrechen , verbrennen Sie [es] einfach . Eine tolle Sache . Das Uran wird graduell verbraucht, ein bisschen wie eine Kerze . [Sie] sehen , dass [es] eine Art Säule ist , oft als " Wandernde Welle Reaktor " bezeichnet . [Das] löst wirklich das Treibstoffproblem . Hier ist ein Bild eines Ortes in Kentucky . [Das]

verbrennt man den Müll , und man kann sogar [den Abfall von heutigen Reaktoren als Antriebsstoff] benutzen Anstatt [sich] [darüber] den Kopf zu zerbrechen , verbrennen Sie [es] einfach . Eine tolle Sache . [Das Uran] wird graduell verbraucht , ein bisschen wie eine Kerze . [Sie] sehen , dass [es] eine Art Säule ist , oft als " Wandernde Welle Reaktor " bezeichnet . [Das] löst wirklich das Treibstoffproblem . Hier ist ein Bild eines Ortes in Kentucky . [Das]

because you 're burning [that 99 percent] , you have greatly improved cost profile . You actually burn up the waste , and you can actually use as fuct [all the leftover waste from today is [reactore]] . So , instead of worrying about [them] , you just take [that] . [tt] 's a great thing . [tt] breathes this uranium as [tt] goes along . So [tt] 's kind of like a candle . You can see [it] 's a log there , often referred to as a traveling wave reactor . In terms of fuel , [this] really solves the problem

tnat now you can simulate and see tnat , yes , with the right material 's approach , [tnis] looks like [it] would work . And ,

pecause you're burning [that 99 percent], you have greatly improved cost profile. You actually burn up the waste, and you can actually use as fuel [all the leftover waste from today 's [reactors]] . So , instead of worrying about [them] , you

(1) marked mentions outside of chains, (2) antecedents of chains not marked as first elements of chains, (3) some other error types. But not for potentially missing elements!

automatic inconsistency checks for:

METHOD

... and RESULTS

Sentence based word alignment

Provides mentions with chain ID in ST and TT

Question: Which chain in ST corresponds to chain in TT?

CHAIN MATCH highest score

And **she** reminded me that my favorite color was blue. Und **sie** erinnerte mich daran , dass meine Lieblingsfarbe blau sei.

CHAIN MATCH

ENchain: set_5 corresponds to DEchain set_7

ENchain set 5 with mentions: {0: 'she'}

...AND EXAMPLES

DEmentionsDEchain set_7 with mentions: {2: 'sie'}

Scoring function computes for each English chain a score of

alignment to

each German

chain

Two pairs, same score

OVERLAP

My brother wanted a red balloon, and I wanted a pink balloon. Mein Bruder wollte einen roten , und ich wollte **einen rosafarbenen**.

OVERLAP

Mentions in ENchain: set_220 found in these DEchains ['set_4', 'set_279']

ENchain set 220 with mentions: {0: 'a pink balloon'}

DEchain: set_4 with mentions: {1: 'einen rosafarbenen', 2: 'einen roten'}

DEchain: set_279 with mentions: {1: 'einen rosafarbenen'}

So what 's the bottom line here? GermanChains Chains that

appear only in either German or English

*Subset of ParCorFull

NotInEnglish EnglishChainsNotIn German

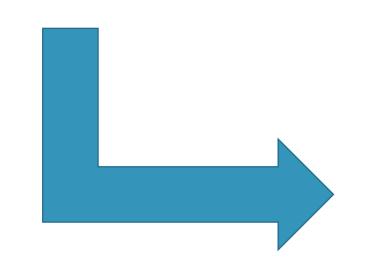
Was lernen wir daraus?

==>GermanChainsNotInEnglish > DEchain: set_86 with mentions: {0: 'daraus'}

And until you have children, you do n't know what it 's like.

Und bevor du nicht selbst Kinder hast , wirst du nicht verstehen , was ich meine .

==>EnglishChainsNotInGerman -> ENchain: set_279 with mentions: {2: 'it'}



	Source	Target
X ChainsNotIn Y	197	76
Chains	1,102	1,114
mentions	3,108	3,248
annotated words	9,009	7,345

FUTURE WORK

- detailed analysis and systematic description of output cases
- visualization in annotation tool



parallel VS. comparable

DISCUSSION

annotation

