

Technical guidelines for annotating non-standard language corpora in WebAnno

Tomaž Erjavec
Jožef Stefan Institute

UniZg, 4. 12. 2015

Overview

1. Goals and method
2. Tokenisation
3. Sentence segmentation
4. Normalisation
5. Combinations and table

Goals

„Nemrem verovat“

1. Difficult to search in corpus: „vjerovati“
2. Low quality tagging and lemmatisation

Therefore:

- Normalisation (standardisation) of word forms:
„Ne mogu vjerovati“

Additional goal: what is a word, what is a sentence?

„utakmica *Dinamo-Partizan*“
„moj *deb . brat me opet gnjavi*“

- Automatic tokenisation sometimes makes errors in assigning boundaries between tokens
- Similarly, there are errors in segmentation into sentences
- We need a „gold“ corpus, so we can improve tokenisation and sentence segmentation programs.

Manual annotation

- We have chosen tweets of different non-standardness levels
- The tweets are already automatically tokenised and sentence segmented
- Manual annotation:
 - **Technical guidelines**
 - Linguistic guidelines
 - Practical
 - Annotation...

Overview

Annotation Layer	Use
Corrections	Tokenisation corrections
Sentences	Sentence segmentation
Normalisations	Normalised word

Special chars	Use
\	No space after word
\$0	Deletion
\$.	End of sentence

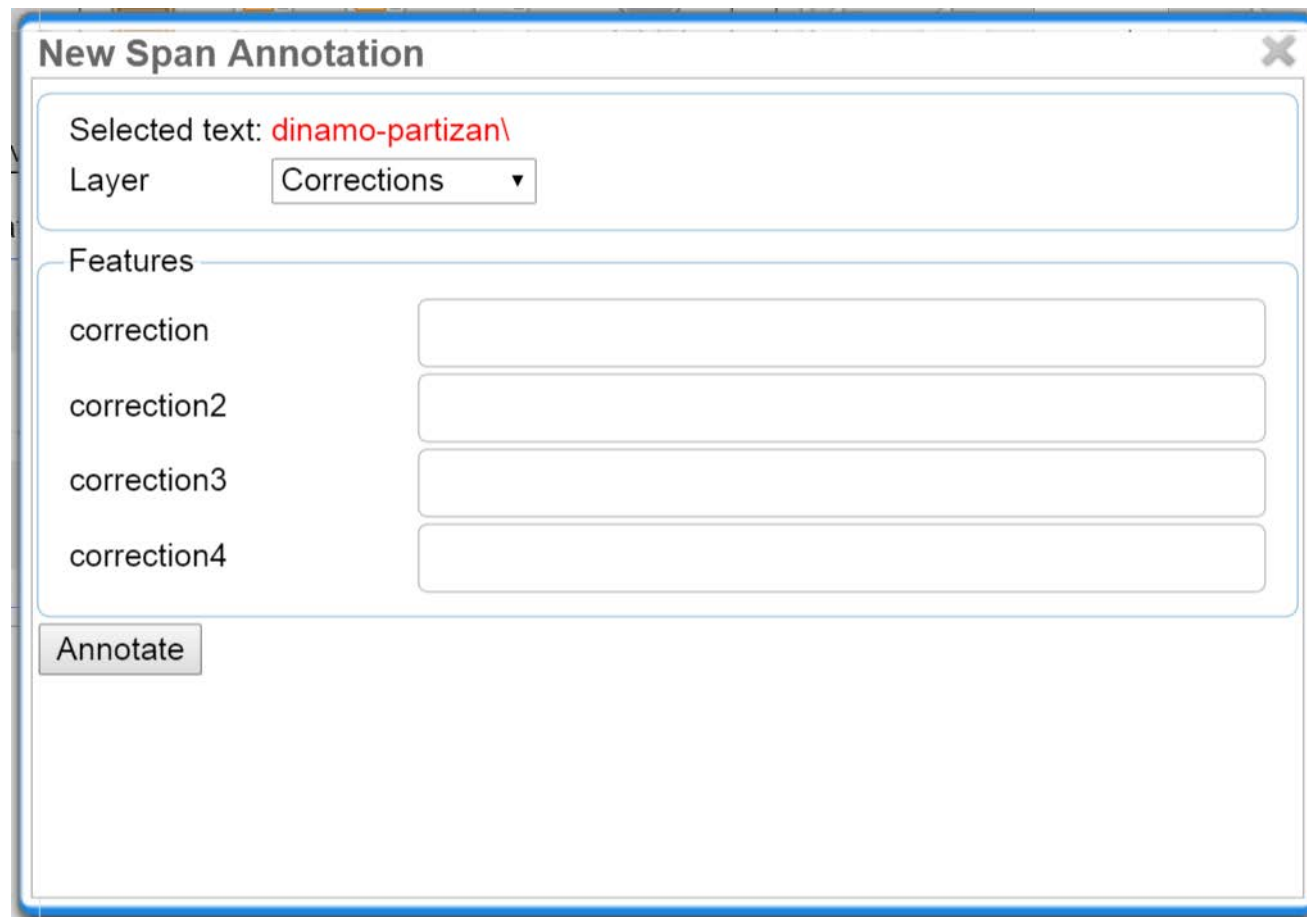
Spacing: \

- Tokens in WebAnno end in backslash if there is not space after it in the source tweet
- This char helps us to reconstruct how the original tweet was written, including spacing
- This character is added only for information; we do not write it in our annotations

- *Da, trebalo bi.* → *Da\ , trebalo bi \.*
- *Javim ti se danas/sutra.* → *Javim ti se danas\ /\ sutra .*
- *više njih? :\ brrr...* → *više njih\ ? :\\$0 brrr\ ...*

Tokenisation correction

- Annotation layer: *Corrections* (4 values)



New Span Annotation

Selected text: **dinamo-partizan**

Layer: Corrections

Features

correction	<input type="text"/>
correction2	<input type="text"/>
correction3	<input type="text"/>
correction4	<input type="text"/>

Annotate

Tokenisation: splitting tokens

Ne mogu gledati ovaj dinamo-partizan\ ...

New Span Annotation ✕

Selected text: namo-parti

Layer: Corrections ▾

Features

correction	dinamo
correction2	-
correction3	partizan
correction4	

Annotate

dinamo | - | partizan

Ne mogu gledati ovaj dinamo-partizan\ ...

Tokenisation: merging tokens: ;->

Kidam nalijevo ;\ ->

New Span Annotation

Selected text: ;

Layer: Corrections

Features:

correction	;->
correction2	
correction3	
correction4	

Annotate

New Span Annotation

Selected text: ->

Layer: Corrections

Features:

correction	\$0
correction2	
correction3	
correction4	

Annotate

Kidam nalijevo ;\ ->

;-> \$0

Sentence segmentation

- Layer *Sentences*
- Value: only \$. (end of sentence)
- Not marked at end of tweet

Ne mogu vjerovati\ \$. To nisu ljudi :-(Bježim van\ !

Adding a sentence boundary

Ne mogu vjerovati\ . To nisu ljudi :-(Bježim van\ !

Edit Span Annotation

Selected text: :-(

Layer

Sentences

Features

sentence

\$.

sentence2

sentence3


sentence4

Ne mogu vjerovati\ . To nisu ljudi :-(Bježim van\ !

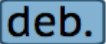

Annotate

Delete

Deleting a sentence boundary

moj deb\  . brat me opet gnjavi\ !

Edit Span Annotation	New Span Annotation	Edit Span Annotation
<p>Selected text: .</p> <p>Layer: Sentences</p> <p>Features:</p> <p>sentence: \$.</p> <p>sentence2: </p> <p>sentence3: </p> <p>sentence4: </p> <p>Annotate Delete</p>	<p>Selected text: .</p> <p>Layer: Corrections</p> <p>Features:</p> <p>correction: \$0 </p> <p>correction2: </p> <p>correction3: </p> <p>correction4: </p> <p>Annotate</p>	<p>Selected text: deb\</p> <p>Layer: Corrections</p> <p>Features:</p> <p>correction: deb.</p> <p>correction2: </p> <p>correction3: </p> <p>correction4: </p> <p>Annotate Delete</p>

moj   . brat me opet gnjavi\ !

Deleting a tweet

- If it makes no sense to annotate a tweet
- The first token should be annotated with \$0 on the layer „Sentences“

u žmaucu se ga pa kr sred belga dneva poha

Edit Span Annotation

Selected text: u

Layer

Sentences

Features

sentence

\$0

sentence2

sentence3

sentence4

Annotate

Delete

\$0

u žmaucu se ga pa kr sred belga dneva poha

Normalisation

- Layer *Normalisations*
- Original:
nis normalan\ !
- Corrected:

nisi



nis normalan\ !

Normalisation: splitting tokens

Nemrem verovat\ ...

New Span Annotation

Selected text: emre

Layer

Normalisations

Features

normalisation

Ne

normalisation2

•••••
mogu

normalisation3

normalisation4

Annotate

Ne | mogu vjerovati
Nemrem verovat\ ...

Normalisation: merging tokens

pre glupo buraz\ !

New Span Annotation

Selected text: **pr**

Layer

Features

normalisation

normalisation2

normalisation3

normalisation4

Annotate

Edit Span Annotation

Selected text: **glupo**

Layer

Features

normalisation

normalisation2

normalisation3

normalisation4


Annotate

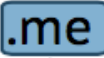

Delete

preglupo **\$0**
pre glupo buraz\ !

Combined corrections

- One error in automatic annotation causes others:

ovaj  \ me pun egotripera\ !

ovaj  \  me pun egotripera\ !

Annotating split tokens

ne odma.me kasnije pingniš ...

New Span Annotation

Selected text: **ma.me**

Layer: Corrections

Features

correction: odma

correction2: .

correction3: me

correction4:

Annotate

New Span Annotation

Selected text: **ma.m**

Layer: Normalisations

Features

normalisation: odmah

normalisation2: .

normalisation3: me|

normalisation4:

Annotate

New Span Annotation

Selected text: **ma.m**

Layer: Sentences

Features

sentence:

sentence2: \$.|

sentence3:

sentence4:

Annotate

. \$.
odmah . me
odma . me

ne odma.me kasnije pingniš ...

Overview

Annotation Layer	Use
Corrections	Tokenisation corrections
Sentences	Sentence segmentation
Normalisations	Normalised word

Char	Layer	Use
\	Tokens	No space after word
\$0	Corrections	Deletion of (merged) token
\$0	Sentences	Deletion of complete tweet (on first token)
\$0	Normalisations	Normalisation merged on previous token
\$.	Sentences	End of sentence (on token that ends the sentence)

Suggested method of annotation

1. Switch off layer *Normalisations*
2. In file correct *Corrections* and *Sentences*
3. Switch on layer *Normalisations*
4. In file correct *Normalisations*
5. Click *Done*
6. Go to next file