

Semantic Role Labeling dane i modele

Seminarium ZIL IPI PAN

Michał Ulewicz

Warszawa 2022-06-13

Plan prezentacji

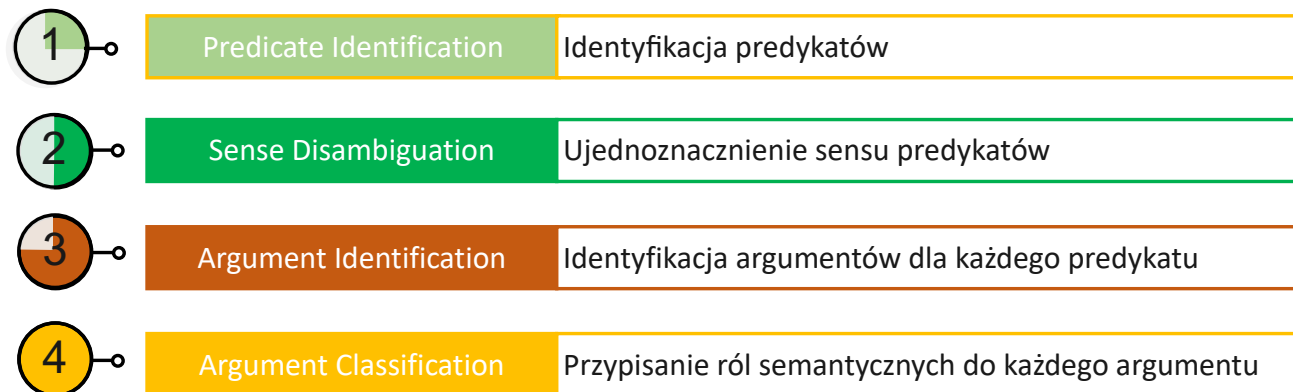
- Wstęp do SRL
- Dlaczego SRL?
- Universal Propositions 2.0
- SRL a język polski
- Modele SRL
- Podsumowanie

Semantic Role Labeling (SRL)

Who did what to whom, when, where and how?

(Gildea and Jurafsky, 2000; Màrquez et al., 2008)

Semantic Role Labeling (SRL)



Nazewnictwo argumentów: **A0** lub **ARG0**, **A1** lub **ARG1** ...

Jeden token w ramach danej ramki może mieć tylko jedną etykietę semantyczną. Predykat nie może mieć etykiety semantycznej.

Przykładowe ramki z Proposition Bank *:

break.01 (to cause not to be whole)

A0 – breaker/hitter

A1 – thing broken/hit

A2 – instrument

A3 – broken into ... (pieces)

break.18 (to violate the law, trust)

A0 – violator

A1 – thing violated

break.19 (to suspend an action)

A0 – person causing the break

A1 – activity interrupted

A2 – purpose of pause

* *The Proposition Bank: An Annotated Corpus of Semantic Roles* (Palmer et al., CL 2005)

Anotacije SRL (span/dependency)

This young man thought to himself , I have been here in America for ten years , and I did not send one letter to my family .

SPAN BASED

1	This	this	DET	DT	3	det	-	-	A0	-	-
2	young	young	ADJ	JJ	3	amod	-	-	A0	-	-
3	man	man	NOUN	NN	4	nsubj	-	-	A0	-	-
4	thought	think	VERB	VBD	0	root	Y	think.01	-	-	-
5	to	to	ADP	IN	6	case	-	-	A2	-	-
6	himself	himself	PRON	PRP	4	obl	-	-	A2	-	-
7	,	,	PUNCT	,	13	punct	-	-	-	-	-
8	I	I	PRON	PRP	13	nsubj	-	-	A1	-	-
9	have	have	AUX	VBP	13	aux	-	-	-	-	-
10	been	be	AUX	VBN	13	cop	Y	be.01	-	-	-
11	here	here	ADV	RB	8	ccomp	-	-	A2	-	-
12	in	in	ADP	IN	15	case	-	-	A2	-	-
13	America	America	PROPN	NNP	13	obl	-	-	A2	-	-
14	for	for	ADP	IN	18	case	-	-	-	-	-
15	ten	ten	NUM	CD	18	nummod	-	-	AM-TMP	-	-
16	years	year	NOUN	NNS	13	obl	-	-	AM-TMP	-	-
17	,	,	PUNCT	,	24	punct	-	-	-	-	-
18	and	and	CCONJ	CC	24	cc	-	-	-	-	-
19	I	I	PRON	PRP	24	nsubj	-	-	-	-	A0
20	did	do	AUX	VBD	24	aux	-	-	-	-	-
21	not	not	PART	RB	24	advmod	-	-	-	-	AM-NEG
22	send	send	VERB	VB	4	conj	Y	send.01	-	-	-
23	one	one	NUM	CD	26	nummod	-	-	-	-	A1
24	letter	letter	NOUN	NN	24	obj	-	-	-	-	A1
25	to	to	ADP	IN	29	case	-	-	-	-	A2
26	my	my	PRON	PRP\$	29	nmod:poss	-	-	-	-	A2
27	family	family	NOUN	NN	24	obl	-	-	-	-	A2
28	.	.	PUNCT	.	4	punct	-	-	-	-	-

DEPENDENCY BASED

1	This	this	DET	DT	3	det	-	-	-	-	-
2	young	young	ADJ	JJ	3	amod	-	-	-	-	-
3	man	man	NOUN	NN	4	nsubj	-	-	A0	-	-
4	thought	think	VERB	VBD	0	root	Y	think.01	-	-	-
5	to	to	ADP	IN	6	case	-	-	-	-	-
6	himself	himself	PRON	PRP	4	obl	-	-	A2	-	-
9	,	,	PUNCT	,	13	punct	-	-	-	-	-
10	I	I	PRON	PRP	13	nsubj	-	-	A1	-	-
11	have	have	AUX	VBP	13	aux	-	-	-	-	-
12	been	be	AUX	VBN	13	cop	Y	be.01	-	-	-
13	here	here	ADV	RB	8	ccomp	-	-	A2	-	-
14	in	in	ADP	IN	15	case	-	-	-	-	-
15	America	America	PROPN	NNP	13	obl	-	-	-	-	-
16	for	for	ADP	IN	18	case	-	-	-	-	-
17	ten	ten	NUM	CD	18	nummod	-	-	-	-	-
18	years	year	NOUN	NNS	13	obl	-	-	AM-TMP	-	-
19	,	,	PUNCT	,	24	punct	-	-	-	-	-
20	and	and	CCONJ	CC	24	cc	-	-	-	-	-
21	I	I	PRON	PRP	24	nsubj	-	-	-	-	A0
22	did	do	AUX	VBD	24	aux	-	-	-	-	-
23	not	not	PART	RB	24	advmod	-	-	-	-	AM-NEG
24	send	send	VERB	VB	4	conj	Y	send.01	-	-	-
25	one	one	NUM	CD	26	nummod	-	-	-	-	-
26	letter	letter	NOUN	NN	24	obj	-	-	-	-	A1
27	to	to	ADP	IN	29	case	-	-	-	-	-
28	my	my	PRON	PRP\$	29	nmod:poss	-	-	-	-	-
29	family	family	NOUN	NN	24	obl	-	-	-	-	A2
30	.	.	PUNCT	.	4	punct	-	-	-	-	-

Anotacje - role

A0, A1, A2, A3, A4, A5

Role, których znaczenie jest zdefiniowane w ramach ramek z PropBank

R-A0, R-A1

Referencje do innego argumentu w ramach tej samej ramki

A0 R-A0

The boy who played basketball.

C-A0, C-A1

Kontynuacja argumentu

A1 C-A1

Pojawiły się napięcia ... napięcia, które nie zawsze są związane z zasadami czy potencjałem gospodarczym.

Anotacje – modyfikatory (argumenty określające dalsze właściwości zdarzenia)

AM-TMP – when?

AM-LOC – where?

AM-DIR – where from ... to ...?

AM-MNR – how?

AM-PRP, AM-CAU – why?

AM-NEG – not ...

AM-MOD – modal verb

Modifier	Name	Description	Example	Span
COM	Comitatives	Comitative modifiers indicate who an action was done with. This can include people or organizations but excludes objects, which would be considered instrumental modifiers.	I sang a song with my sister. The man joined the club with his friend.	with my sister with his friend
LOC	Locatives	Locative modifiers indicate where some action takes place. The notion of a locative is not restricted to physical locations, but abstract locations are being marked as LOC as well.	The mass murder happening right now in Sudan In the Congo soldiers were accepting people into the party But the likely emission pattern of these man - made sounds is distinct from the pattern of underground tremors	Sudan Congo underground
DIR	Directional	Directional modifiers show motion along some path. 'walk along the road' is a directional, but 'walk around the countryside' is a location.	Or to turn this completely around the congressman snapped back that there had been enough studies of the issue	turn around back
GOL	Goal	This tag is for the goal of the action of the verb. This includes the final destination of motion verbs and benefactive arguments that receive	The child fed the cat for her mother. The couple translated for the Americans.	for her mother for the Americans
MNR	Manner	Manner adverbs specify how an action is performed. Manner tags should be used when an adverb could be an answer to a question	Among 33 men who worked closely with the substance, 28 have died. Workers mechanically mixed the dry fibers in a process used to make filters.	closely mechanically
TMP	Temporal	Temporal ARGMs show when an action took place such as 'in 1987', 'last Wednesday', 'soon' or 'immediately.' Also included in this	On the evening of the 23rd of the twelfth lunar month Le Beverly somehow struggled through the 1990s .	evening 1990s
EXT	Extent	Indicate the amount of change occurring from an action: Numerical adjuncts 'raised prices by 15 percent'; Quantifiers such as a lot, least,	New Hampshire shares closed yesterday at \$ 3.75 [*U*], off 25 cents, in New York Stock Exchange composite trading. An active 55-year-old in Boca Raton may care more about Senior Olympic games	\$ 3.75 [*U*], off 25 cents more
REC	Reciprocals	These include reflexives and reciprocals such as himself, itself, themselves, each other, or own, which refer back to one of the other arguments.	If the stadium was such a good idea someone would build it himself.	himself
PRD	Secondary Predication	These are used to show that an adjunct of a predicate is in itself capable of carrying some predicate structure.	The boys pinched them dead He will join the board as a nonexecutive director	dead as a nonexecutive director
PRP	Purpose Clause	Purpose clauses are used to show the motivation for some action. Clauses beginning with 'in order to' and 'so that' are canonical	he Commerce Commission said Commonwealth Edison could raise its electricity rates by \$ 49 million to pay for the plant The red-carpet treatment tempts them to return to a heartland city for future meetings.	to pay for the plant for future meetings
CAU	Cause Clauses	Similar to purpose clauses, these indicate the reason for an action. Clauses beginning with 'because' or 'due to' are canonical cause clauses.	They will remain on that so-called priority watch list because of an interim review.	because of an interim review
DIS	Discourse	These are markers which connect a sentence to a preceding sentence. Examples of discourse markers are: also, however, too, as well, but, and, as we've seen before, instead, on the other hand, for instance,	But for now , they're looking forward to their winter meeting Alan, will you go to the store? Gosh, I can't believe it	But Alan Gosh
MOD	Modals	Modals are: will, may, can, must, shall, might, should, could, and would.	Otherwise , Soviet forces would fall into a series of devastating traps on a future remote Where the new majority will face some difficult decisions .	would will
NEG	Negation	This tag is used for elements such as not, n't, never, no longer, and other markers of negative sentences.	I did not send one letter to my family Saddam Hussein did not invade Kuwait	not not
ADV	Adverbials	These are used for syntactic elements which clearly modify the event structure of the verb in question, but which do not fall under any of	though it is clear that But as cellphone users tend to buy new handsets only once every couple of years	clear tend
ADJ	Adjectival	This tag is used in a manner that is similar to ADV, but its use is restricted to noun annotation.	for what you find of cheap prices Unless builders are going to supply affordable housing at a loss	cheap affordable

Formaty plików: CoNLL-U/CoNLL-U Plus

Formaty plików:

- CoNLL-U (*.conllu) – Universal Dependencies – 10 atrybutów
- CoNLL-U Plus (*.conllup) – Uogólnienie formatu CoNLL-U

CoNLL-U Plus - nagłówek w pliku z definicją kolumn:

global.columns = ID UP:PRED UP:ARGHEADS UP:ARGSPANS

global.columns = ID UP:FLAG UP:PRED UP:FRAME

global.columns = ID FORM LEMMA UPOS XPOS FEATS HEAD DEPREL DEPS MISC UP:PRED UP:ARGHEADS UP:ARGSPANS

global.columns = ID FORM LEMMA UPOS XPOS FEATS HEAD DEPREL DEPS MISC UP:FLAG UP:PRED UP:FRAME

Anotacje SRL w plikach CoNLL-U/CoNLL-U Plus

Anotacje SRL jako metadana

# srl = [{"args":["arg1",[7,8,10,11]],"roleset":"affirm.01","frameFile":"","target":[1]},{"args":["argm-tmp",[8]],["arg1",[4]],["arg2",[13]],"roleset":"point.01","frameFile":"","target":[11]]}									
# text = Afirmou que o conjunto de fatos, em princípio, aponta o envolvimento de Quércia.									
# sent_id = CF19-2									
# source = CETENFolha n=19 cad=Caderno Especial sec=nd sem=94a									
1	Afirmou	afirmar	VERB	<mv> V PS 3S	Mood=Ind Number=Sing	0	root	_	_
2	que	que	SCONJ	KS @SUB	_	11	mark	_	_
3	o	o	DET	<artd> ART M	Definite=Def Gender=Ma	4	det	_	_
4	conjunto	conjunto	NOUN	<np-def> N M	Gender=Masc Number=S	11	nsubj	_	_
5	de	de	ADP	PRP @N<	_	6	case	_	_
6	fatos	fato	NOUN	<np-idf> N M	Gender=Masc Number=P	4	nmod	_	SpaceAfter=No
7	,	,	PUNCT	PU @PU	_	9	punct	_	_
8	em	em	ADP	PRP @ADVL>	_	9	case	_	_
9	princípio	princípio	NOUN	N M S @P<	Gender=Masc Number=S	11	obl	_	SpaceAfter=No
10	,	,	PUNCT	PU @PU	_	9	punct	_	_
11	aponta	apontar	VERB	<mv> V PR 3S	Mood=Ind Number=Sing	1	ccomp	_	_
12	o	o	DET	<artd> ART M	Definite=Def Gender=Ma	13	det	_	_
13	envolvimento	envolvimento	NOUN	<np-def> N M	Gender=Masc Number=S	11	obj	_	_
14	de	de	ADP	PRP @N<	_	15	case	_	_
15	Quércia	Quércia	PROPN	PROP M S @f	Gender=Masc Number=S	13	nmod	_	SpaceAfter=No
16	.	.	PUNCT	PU @PU	_	1	punct	_	_

Anotacje SRL w plikach CoNLL-U/CoNLL-U Plus

Anotacje SRL – format “łatwy do edycji”

# global.columns = ID FORM LEMMA UPOS XPOS FEATS HEAD DEPREL UP:FLAG UP:PRED UP:FRAME										
# sentence-text = Chiny muszą polegać na sobie , aby stać się potężne .										
1	[	[	PUNCT	_	_	3	punct	_	_	_
2	Chiny	china	NOUN	_	_	3	nsubj	_	_	A0 A1
3	muszą	musić	VERB	_	_	0	ROOT	_	_	_
4	polegać	polegać	VERB	_	_	3	xcomp	Y	rely.01	_
5	na	na	ADP	_	_	6	case	_	_	_
6	sobie	sobie	PRON	_	_	4	obl:arg	_	_	A1
7	,	,	PUNCT	_	_	9	punct	_	_	_
8	aby	aby	SCONJ	_	_	9	mark	_	_	_
9	stać	stać	VERB	_	_	4	advcl	Y	become.01	_
10	się	się	PRON	_	_	9	expl:pv	_	_	_
11	potężne	potężny	ADJ	_	_	9	xcomp:pred	_	_	A2
12	.	.	PUNCT	_	_	3	punct	_	_	_

Anotacje SRL w plikach CoNLL-U/CoNLL-U Plus

Nowy format wprowadzony z Universal Propositions 2.0

# global.columns = ID FORM LEMMA UPOS XPOS FEATS HEAD DEPREL DEPS MISC UP:PRED UP:ARGHEADS UP:ARGSPANS												
# sent_id = dev-s20												
# text = Jasnobrązowy i czarny pies bawią się ze sobą na trawie.												
# orig_file_sentence = CDScorpus_6441_A#645												
1	Jasnobrązowy	jasnobrązowy	ADJ	adj:sg:nom:m2:	Animacy=Nhum Case=Nc	5	nsubj	5:nsubj	_	_	_	_
2	i	i	CCONJ	conj	_	4	cc	4:cc	_	_	_	_
3	czarny	czarny	ADJ	adj:sg:nom:m2:	Animacy=Nhum Case=Nc	4	amod	4:amod	_	_	_	_
4	pies	pies	NOUN	subst:sg:nom:m	Animacy=Nhum Case=Nc	1	conj	1:conj 5:n	_	_	_	_
5	bawią	bawić	VERB	fin:pl:ter:imper	Aspect=Imp Mood=Ind N	0	root	0:root	_	play.01	A1:4 A2:10	A1:1-4 A2:9-10
6	się	się	PRON	part	PronType=Prs Reflex=Yes	5	expl:pv	5:expl:pv	_	_	_	_
7	ze	z	ADP	prep:inst:wok	AdpType=Prep Variant=L	8	case	8:case	Case=Ins	_	_	_
8	sobą	siebie	PRON	siebie:inst	Case=Ins PronType=Prs F	5	obl:arg	5:obl:arg	_	_	_	_
9	na	na	ADP	prep:loc	AdpType=Prep	10	case	10:case	Case=Loc	_	_	_
10	trawie	trawa	NOUN	subst:sg:loc:f	Case=Loc Gender=Fem M	5	obl	5:obl	SpaceAfter=No	_	_	_
11	.	.	PUNCT	interp	PunctType=Peri	5	punct	5:punct	_	_	_	_

conllup – biblioteka python dla CoNLL-U Plus

```
1 #define conllup columns
Column_ID = Column("ID", ColumnType.ID)
Column_FORM = Column("FORM")
Column_LEMMA = Column("LEMMA")
Column_UPOS = Column("UPOS")
Column_XPOS = Column("XPOS")
Column_FEATS = Column("FEATS")
Column_HEAD = Column("HEAD")
Column_DEPREL = Column("DEPREL")
Column_DEPS = Column("DEPS")
Column_MISC = Column("MISC")

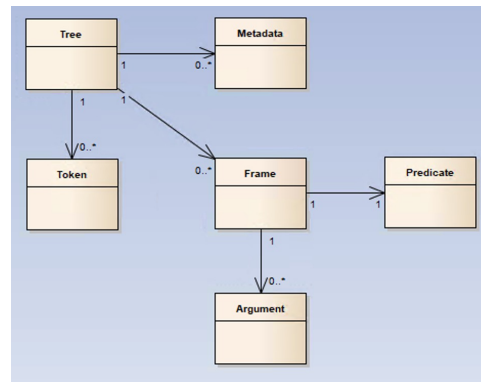
# define conllup column list
columns = [
    Column_ID,
    Column_FORM,
    Column_LEMMA,
    Column_UPOS,
    Column_XPOS,
    Column_FEATS,
    Column_HEAD,
    Column_DEPREL,
    Column_DEPS,
    Column_MISC,
    Column("UP:PRED", ColumnType.UP_PRED),
    Column("UP:ARGHEADS", ColumnType.UP_ARGHEADS),
    Column("UP:ARGSPANS", ColumnType.UP_ARGSPANS)
]
```

```
3 # create the first token
token = tree.add_token(1)
token.set_attribute(Column_FORM, "हालाँकि")
token.set_attribute(Column_LEMMA, "हालाँकि")
token.set_attribute(Column_UPOS, "SCONJ")
token.set_attribute(Column_XPOS, "CC")
token.set_attribute(Column_HEAD, 14)
token.set_attribute(Column_DEPREL, "mark")
token.set_attribute(
    Column_MISC, "AltTag=SCONJ-CONJ|ChunkId=CCP|ChunkType=head|Translit=hālāñki")

# create the second token
token = tree.add_token(2)
token.set_attribute("FORM", "स्थानीय")
token.set_attribute("LEMMA", "स्थानीय")
```

```
2 # create conllup tree
tree = Tree(columns)

# add metadata
tree.add_metadata(
    "source_sent_id", "conllu http://hdl.handle.net/11234/1-4611 UD_Hindi-HDTB/hi_hdtb-ud-dev.conllu dev-s32")
tree.add_metadata("sent_id", "dev-s32")
tree.add_metadata(
    "text", "हालाँकि स्थानीय वाशियों में यह रामाभार स्तूप के नाम से ही आज भी जाना जाता है।")
```



```
4 # create SRL frames for two predicates
frame = Frame()
tree.add_frame(frame)
frame.set_predicate(Predicate("know.02", token=14))
frame.add_argument(Argument("AM-DIS", head=1, start=1, end=3))
frame.add_argument(Argument("A1", head=5, start=5, end=9))
frame.add_argument(Argument("AM-TMP", head=12))

frame = Frame()
tree.add_frame(frame)
frame.set_predicate(Predicate("be.01", token=16))
frame.add_argument(Argument("AM-DIS", head=1, start=1, end=3))
frame.add_argument(Argument("A1", head=5, start=5, end=9))
frame.add_argument(Argument("AM-TMP", head=12))
frame.add_argument(Argument("A2", head=14, start=14, end=15))
```

```
5 # write conllup data to conllu file
with open('./tests/data/output/test_create_up.conllup', "w", encoding='utf8') as fo:
    fo.write(tree.to_conllup(definition=True))

# write conllup data to json file
with open('./tests/data/output/test_create_up.json', "w", encoding='utf8') as fo:
    fo.write(json.dumps(tree.to_json(), ensure_ascii=False, indent=3))

6 f = open('./tests/data/sample_up_1.conllu', encoding='utf8')

for up in parse(f):
    #up - Tree object
```

conllup – biblioteka python dla CoNLL-U Plus

CoNLLu Plus output file

# global.columns = ID FORM LEMMA UPOS XPOS FEATS HEAD DEPREL DEPS MISC UP:PRED UP:ARGHEADS UP:ARGSPANS													
# source_sent_id = conllu http://hdl.handle.net/11234/1-4611 UD_Hindi-HDTB/hi_hdtb-ud-dev.conllu dev-s32													
# sent_id = dev-s32													
# text = हालाँकि स्थानीय वाशिदों में यह रामाभार स्तूप के नाम से ही आज भी जाना जाता है ।													
1	हालाँकि	हालाँकि	SCONJ	CC	—	14	mark	—	AltTag=SCON	—	—	—	—
2	स्थानीय	स्थानीय	ADJ	JJ	Case=Acc	3	amod	—	AltTag=SCON	—	—	—	—
3	वाशिदों	वाशिदा	NOUN	NN	Case=Acc Ge	14	obl	—	Vib=0_में Tar	—	—	—	—
4	में	में	ADP	PSP	AdpType=Pos	3	case	—	ChunkId=NP1	—	—	—	—
5	यह	यह	PRON	PRP	Case=Nom f	14	nsubj	—	Vib=0 Tam=	—	—	—	—
6	रामाभार	रामाभार	PROPN	NNPC	Case=Nom C	7	compound	—	Vib=0 Tam=	—	—	—	—
7	स्तूप	स्तूप	PROPN	NNP	Case=Acc Ge	9	nmod	—	Vib=0_का Ta	—	—	—	—
8	के	का	ADP	PSP	AdpType=Pos	7	case	—	ChunkId=NP2	—	—	—	—
9	नाम	नाम	NOUN	NN	Case=Acc Ge	14	acl	—	Vib=0_से Tar	—	—	—	—
10	से	से	ADP	PSP	AdpType=Pos	9	case	—	ChunkId=NP4	—	—	—	—
11	ही	ही	PART	RP	—	9	dep	—	ChunkId=NP4	—	—	—	—
12	आज	आज	NOUN	NN	Case=Acc Ge	14	obl	—	Vib=0 Tam=	—	—	—	—
13	भी	भी	PART	RP	—	12	dep	—	ChunkId=NP5	—	—	—	—
14	जाना	जा	VERB	VM	Gender=Mas	0	root	—	Vib=ना_जा+र know.02	AM-DIS:1 A1:5 AM-TMP:12	AM-DIS:1-3 A1:5-9 AM-TMP:12-0	—	—
15	जाता	जा	AUX	VAUX	Aspect=Imp	14	aux	—	Vib=ता Tam=	—	—	—	—
16	है	है	AUX	VAUX	Mood=Ind N	14	aux:pass	—	Vib=है Tam=be.01	AM-DIS:1 A1:5 AM-TMP:12 A2:14	AM-DIS:1-3 A1:5-9 AM-TMP:12-0 A2:14-15	—	—
17	।	।	PUNCT	SYM	—	14	punct	—	ChunkId=BLK	—	—	—	—

JSON output file

```
{
  "columns": [ ...
  ],
  "metadata": {
    "source_sent_id": "conllu http://hdl.handle.net/11234/1-4611 UD_Hindi-HDTB/hi_hdtb-ud-dev.conllu dev-s32",
    "sent_id": "dev-s32",
    "text": "हालाँकि स्थानीय वाशिदों में यह रामाभार स्तूप के नाम से ही आज भी जाना जाता है ।"
  },
  "tokens": [ ...
  ],
  "frames": [
    {
      "predicate": {
        "sense": "know.02",
        "token": 14
      },
      "arguments": [
        {
          "name": "AM-DIS",
          "head": 1,
          "start": 1,
          "end": 3
        },
        {
          "name": "A1",
          "head": 5,
          "start": 5,
          "end": 9
        },
        {
          "name": "AM-TMP",
          "head": 12,
          "start": 12,
          "end": 15
        }
      ]
    }
  ]
}
```

Dlaczego SRL?

Rozważmy dwa zdania, które są:

- bardzo podobne do siebie jeżeli weźmiemy pod uwagę występujące w nich słowa
- mają zupełnie inne znaczenie (choć to jest dość względne i zależy od kontekstu lub wymagań)

*If the seller can not deliver the product, **the buyer** can cancel such a contract without a penalty, by giving **the seller** a written notice of cancellation.*

*If the seller can not deliver the product, **the seller** can not cancel such a contract without paying a penalty and giving **the buyer** a written notice of cancellation.*

Problem:

- Jak skonstruować dobrą funkcję określającą podobieństwo semantyczne między dwoma zdaniami?
- Jak wyodrębnić cechy istotne z tych zdań tak aby jak najprecyzyjniej oddać znaczenie tych zdań?
- Jakie są oczekiwania użytkownika jeżeli chodzi o podobieństwo między tymi zdaniami?
- Czy tak długie złożone zdania składające się z kilku niezależnych znaczeniowo części daje się porównywać precyzyjnie?
- Jak zaadresować problem niewielkiej liczby przykładów, które pozwolą nam zbudować precyzyjny mechanizm potrafiący rozróżniać i dobrze uogólniać

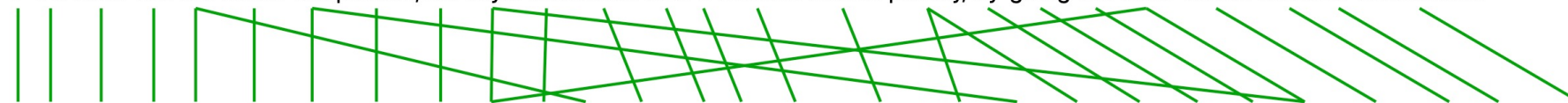
Opcje:

- Strojenie modelu bazowego do rozwiązania jakiegoś konkretnego problemu – dodatkowa warstwa liniowa/klasyfikator + fine tuning modelu
- Czy da się zbudować precyzyjną miarę podobieństwa w oparciu o reprezentacje wektorowe z bazowego modelu bez strojenia go pod konkretny przypadek użycia?

Dlaczego SRL?

Obliczenie reprezentacji wektorowych wykorzystując model transformer, a następnie wykonanie zrównoleglenia słów (word alignment)?

If the seller can not deliver the product, the buyer can cancel such a contract without penalty, by giving the seller written notice of cancellation.



If the seller can not deliver the product, the seller can not cancel such a contract without paying the penalty and giving the buyer written notice of cancellation.

Jeżeli będziemy porównywać odpowiadające sobie tokeny, to możemy wykorzystać np. miarę cosinusową do określenia na ile te zdania są do siebie podobne.

Niemniej pojawiają się dwa problemy:

- Nie zawsze dobrze trafiamy ze zrównolegleniem słów
- Szczególnie w przypadku dłuższych zdań wpływ pojedynczych niezgodności na całość będzie niewielki, a te pojedyncze niezgodności mogą być ważne z punktu widzenia podobieństwa – brakuje tu precyzji

Dlaczego SRL?

Modele sentence-transformers* – reprezentacje wektorowe na poziomie zdań

Wykorzystując jeden z gotowych modeli sentence-transformers otrzymujemy dla tych naszych dwóch zdań podobieństwo cosinusowe na poziomie: 0.9670 – czyli bardzo wysokie podobieństwo, mimo że znaczenie tych zdań jest zupełnie inne. Podobnie jak w poprzednim przypadku, pojedyncze niezgodności mogą być ważne z punktu widzenia podobieństwa – brakuje tu precyzji. Potrzebny byłby więc znowu fine-tuning pod kątem konkretnego problemu.

**Making Monolingual Sentence Embeddings Multilingual using Knowledge Distillation (Reimers, Nils and Gurevych, Iryna 2020)*

Dlaczego SRL?

Wykorzystanie SRL do precyzyjnego porównywania zdań

# text = If the seller can not deliver the product, the buyer can cancel such a contract without penalty, by giving the seller written notice of cancellation .						# text = If the seller can not deliver the product, the seller can not cancel such a contract without paying the penalty and giving the buyer written notice of cancellation .					
1	if	-	-	-	-	1	if	-	-	-	-
2	the	-	-	-	-	2	the	-	-	-	-
3	seller	-	A0	-	-	3	seller	-	A0	-	-
4	can	-	AM-MOD	-	-	4	can	-	AM-MOD	-	-
5	not	-	AM-NEG	-	-	5	not	-	AM-NEG	-	-
6	deliver	deliver.01	-	-	-	6	deliver	deliver.01	-	-	-
7	the	-	A1	-	-	7	the	-	A1	-	-
8	product,	-	A1	-	-	8	product,	-	A1	-	-
9	the	-	-	A0	A0	9	the	-	-	A0	A0
10	buyer	-	-	A0	A0	10	seller	-	-	A0	A0
11	can	-	-	AM-MOD	-	11	can	-	-	AM-MOD	-
12	cancel	cancel.01	-	-	-	12	not	-	-	AM-NEG	-
13	such	-	-	-	-	13	cancel	cancel.01	-	-	-
14	a	-	-	A1	-	14	such	-	-	-	-
15	contract	-	-	A1	-	15	a	-	-	A1	-
16	without	-	-	AM-NMR	-	16	contract	-	-	A1	-
17	a	-	-	AM-NMR	-	17	without	-	-	AM-NMR	-
18	penalty,	-	-	AM-NMR	-	18	a	-	-	AM-NMR	-
19	by	-	-	-	-	19	penalty	-	-	AM-NMR	-
20	giving	give.01	-	-	-	20	and	-	-	-	-
21	the	-	-	-	A2	21	giving	give.01	-	-	-
22	seller	-	-	-	A2	22	the	-	-	-	A2
23	a	-	-	-	A1	23	buyer	-	-	-	A2
24	written	write.01	-	-	A1	24	a	-	-	-	A1
25	notice	-	-	-	A1	25	written	write.01	-	-	A1
26	of	-	-	-	A1	26	notice	-	-	-	A1
27	cancellation	-	-	-	A1	27	of	-	-	-	A1
28	.	-	-	-	-	28	cancellation	-	-	-	A1
						29	.	-	-	-	-

Różnice w czerwonych ramkach.

Reprezentacje wektorowe z modelu bazowego dla tokenów.

Potencjalna możliwość włączenia wiedzy eksperckiej, na temat tego które ramki i argumenty są istotne z punktu widzenia porównania.

Określenie podobieństwa dla każdego argumentu (argument może składać się z kilku tokenów) wspierając się ewentualnie zrównolegleniami słów.

Zbudowanie ogólnej funkcji podobieństwa na podstawie tych podobieństw cząstkowych na poziomie argumentów.

Dlaczego SRL?

SRL w zastosowaniu do zadań rozumienia języka naturalnego (NLU):

- Podobieństwo semantyczne (Semantic Similarity) / Klasyfikacja tekstu (Text Classification)
- Czytanie maszynowe ze zrozumieniem (Machine Reading Comprehension)

<https://rajpurkar.github.io/SQuAD-explorer/>

- Wnioskowanie (Natural Language Inference/Recognizing Textual Entailment)

<https://nlp.stanford.edu/projects/snli/>

Budowa nowych modeli wykorzystując połączenie reprezentacji opartych na anotacjach SRL z reprezentacjami słów opartych na BERT:

- Semantics-aware BERT (SEMBERT)

Zhang, Z., Wu, Y., Zhao, H., Li, Z., Zhang, S., Zhou, X., & Zhou, X. (2020). Semantics-Aware BERT for Language Understanding. Proceedings of the AAAI Conference on Artificial Intelligence, 34(05), 9628-9635.

- Linguistic BERT (LingBERT)

Is Semantic-aware BERT more Linguistically Aware? A Case Study on Natural Language Inference, Ling Liu, Ishan Jindal, Yunyao Li, 2022

SRL – zbiór danych Ontonotes

Zbiór danych dla języka angielskiego z anotacjami SRL.

Liczba zdań:

TRAIN: 111 097

TEST: 11 623

DEV: 15 084

Każde zdanie może mieć 0 .. wiele ramek.

SPAN BASED ONTONOTES (LDC catalog LDC2013T19):

<https://catalog.ldc.upenn.edu/LDC2013T19>

DEPENDENCY BASED ONTONOTES:

Uzyskane poprzez przekształcenie z wykorzystaniem drzewa zależności.

Universal Propositions 2.0* - wstęp

<https://universalpropositions.github.io>

Do budowy skutecznych modeli SRL potrzebne są dane i o ile dla języka angielskiego istnieją wysokiej jakości zestawy danych (proppbank), to dla pozostałych języków takie zbiory praktycznie nie istnieją, głównie ze względu na wysoki nakład pracy i koszt potrzebny do ich stworzenia.

Celem projektu jest dodanie anotacji SRL do tekstów w różnych językach z wykorzystaniem ramek, ról i argumentów z angielskiego Proposition Bank.

Anotacje SRL dla tekstów w różnych językach generowane będą z wykorzystaniem techniki rzutowania anotacji.

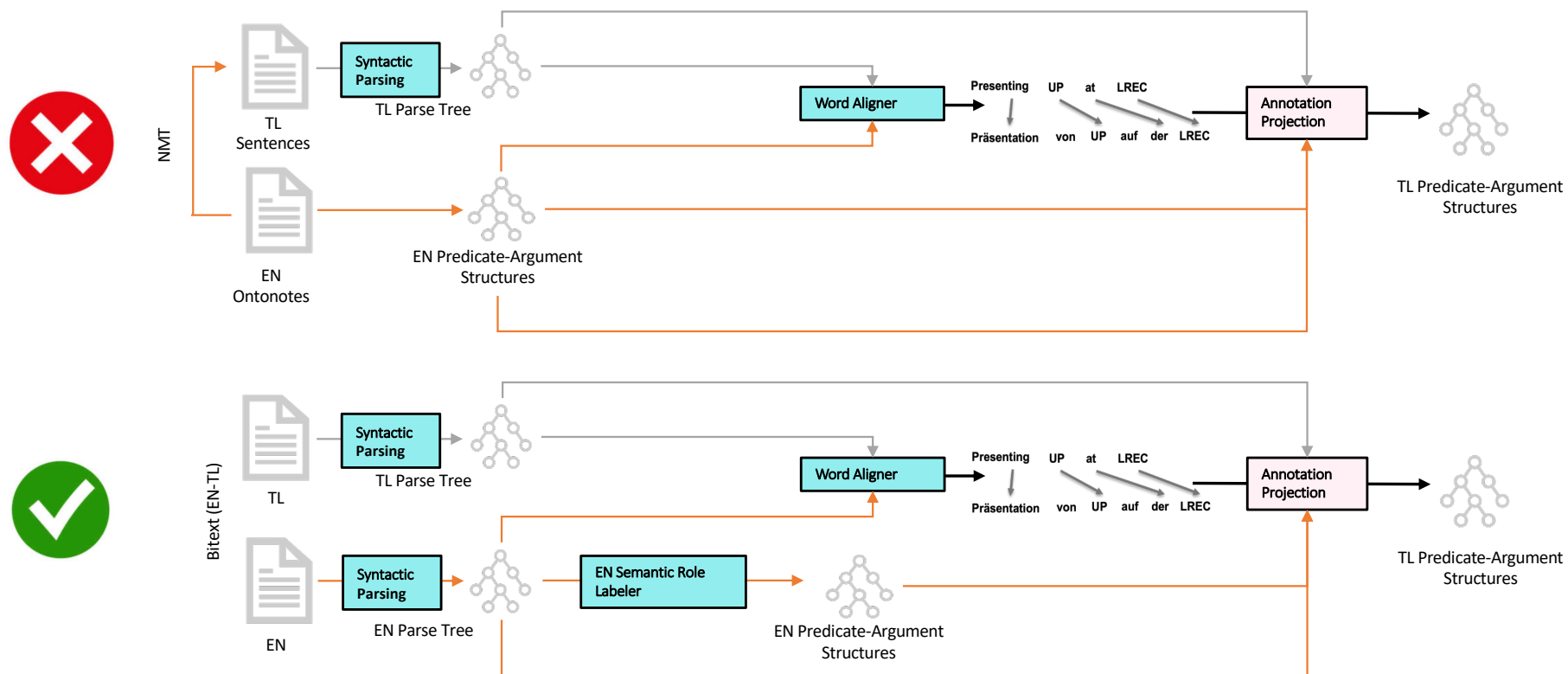
 LICENSE	up2.0
 README.md	update readme
 pl_pdb-up-dev.conllup	up2.0
 pl_pdb-up-test.conllup	up2.0
 pl_pdb-up-train.conllup	up2.0

 LICENSE	up2.0
 README.md	update readme
 pl_lfg-up-dev.conllup	up2.0
 pl_lfg-up-test.conllup	up2.0
 pl_lfg-up-train.conllup	up2.0

* Universal Proposition Bank 2.0. Ishan Jindal, Alexandre Rademaker, Michał Ulewicz, Linh Ha, Huyen Nguyen, Khoi-Nguyen Tran, Huaiyu Zhu and Yunyao Li. *LREC2022.

ID	Language	Corpus	Notes
cs	Czech	CAC ↗ , CLTT ↗ , FicTree ↗ , PDT ↗	
de	German	GSD ↗ , HDT ↗	
el	Greek	GDT ↗	
es	Spanish	GSD ↗ , AnCora ↗	
fi	Finnish	TDT ↗ , FTB ↗	
fr	French	GSD ↗ , Rhapsodie ↗ , Sequoia ↗	
hi	Hindi	HDTB ↗	
hu	Hungarian	Szeged ↗	
id	Indonesian	GSD ↗	
it	Italian	ISDT ↗ , ParTUT ↗ , PoSTWITA ↗ , TWITTURO ↗ , VIT ↗	
ja	Japanese	GSD ↗	
ko	Korean	GSD ↗ , Kaist ↗	
mr	Marathi	UFAL ↗	
nl	Dutch	Alpino ↗ , LassySmall ↗	
pl	Polish	LFG ↗ , PDB ↗	
pt	Portuguese	Bosque ↗ , GSD ↗	
ro	Romanian	Nonstandard ↗ , RRT ↗ , SiMoNERo ↗	
ru	Russian	GSD ↗ , Taiga ↗ , SynTagRus ↗	
ta	Tamil	TTB ↗	
te	Telugu	MTG ↗	
uk	Ukrainian	IU ↗	
vi	Vietnamese	VTB ↗	
zh	Chinese	GSD ↗	

Universal Propositions 2.0 – wybór podejścia



Universal Propositions 2.0 – parser

Porównanie parserów na danych UD 2.5.

Wybór: **STANZA**

Qi, P., Zhang, Y., Zhang, Y., Bolton, J., and Manning, C. D. (2020). Stanza: A python natural language processing toolkit for many human languages. *arXiv preprint arXiv:2003.07082*

Straka, M. (2018). UDPipe 2.0 prototype at CoNLL 2018 UD shared task. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 197–207, Brussels, Belgium, October. Association for Computational Linguistics.

Lang.	Parser	POS	UAS	LAS
DE	spaCy/de_core_news_lg	93.79	47.78	3.85
	spaCy/de_core_news_sm	93.58	47.65	3.77
	spaCy/de_dep_news_trf	92.55	48.08	3.89
	Stanza	94.95	88.38	82.74
	UDPipe	92.47	80.91	74.06
EN	spaCy/en_core_web_lg	91.40	59.04	42.89
	spaCy/en_core_web_sm	90.82	58.54	42.47
	spaCy/en_core_web_trf	93.12	61.50	44.71
	Stanza	96.30	89.84	87.03
	UDPipe	93.63	84.85	80.62
ES	spaCy/es_core_news_lg	91.39	82.73	70.61
	spaCy/es_core_news_sm	90.95	81.72	69.34
	spaCy/es_dep_news_trf	91.82	85.80	73.43
	Stanza	91.68	84.90	72.55
	UDPipe	91.22	80.97	68.53
FR	spaCy/fr_core_news_lg	88.33	78.07	69.11
	spaCy/fr_core_news_sm	85.87	73.11	63.90
	spaCy/fr_dep_news_trf	96.62	91.50	85.16
	Stanza	97.99	93.15	88.68
	UDPipe	94.05	87.42	81.03
IT	spaCy/it_core_news_lg	95.75	90.97	87.44
	spaCy/it_core_news_sm	94.02	88.61	83.64
	Stanza	97.93	93.49	90.02
	UDPipe	97.33	90.39	86.94
PL	spaCy/pl_core_news_lg	97.24	90.01	82.24
	spaCy/pl_core_news_sm	96.37	87.49	78.71
	Stanza	98.40	93.81	87.88
	UDPipe	97.12	86.93	78.64
PT	spaCy/pt_core_news_lg	97.56	91.02	87.32
	spaCy/pt_core_news_sm	97.51	91.08	87.62
	Stanza	98.10	95.33	93.29
	UDPipe	91.30	80.74	67.74
ZH	spaCy/zh_core_web_lg	69.55	47.24	14.26
	spaCy/zh_core_web_sm	61.09	37.45	9.43
	spaCy / trf	74.02	54.21	19.21
	Stanza	83.31	69.03	59.73
	UDPipe	88.58	72.23	59.98

Universal Propositions 2.0 – zrównoleglenia słów (word alignments)

Porównanie rozwiązań dla zrównoleglenia słów:

Wybór: **SimAlign / itermix**

SimAlign/AwesomeAlign - reprezentacje wektorowe
z 8 ukrytej warstwy modelu.

Lang.	Method	P	R	F1
EN-DE	Awesome-align	89.19	71.88	79.60
	BerkeleyAligner	67.38	55.07	60.60
	SimAlign/argmax	92.15	64.44	75.84
	SimAlign/itermax	84.92	72.58	78.27
	SimAlign/match	78.34	70.06	73.97
EN-FR	Awesome-align	63.05	93.96	75.46
	BerkeleyAligner	14.37	23.06	17.70
	SimAlign/argmax	68.28	91.36	78.15
	SimAlign/itermax	58.67	95.10	72.57
	SimAlign/match	55.54	92.92	69.52
EN-ZH	Awesome-align	81.82	81.32	81.57
	BerkeleyAligner	62.11	62.12	62.11
	SimAlign/argmax	88.39	68.28	77.04
	SimAlign/itermax	78.00	77.83	77.92
	SimAlign/match	72.65	75.08	73.85

Sabet, M. J., Dufter, P., Yvon, F., and Schütze, H. (2020). Simalign: High quality word alignments without parallel training data using static and contextualized embeddings. *arXiv preprint arXiv:2004.08728*.

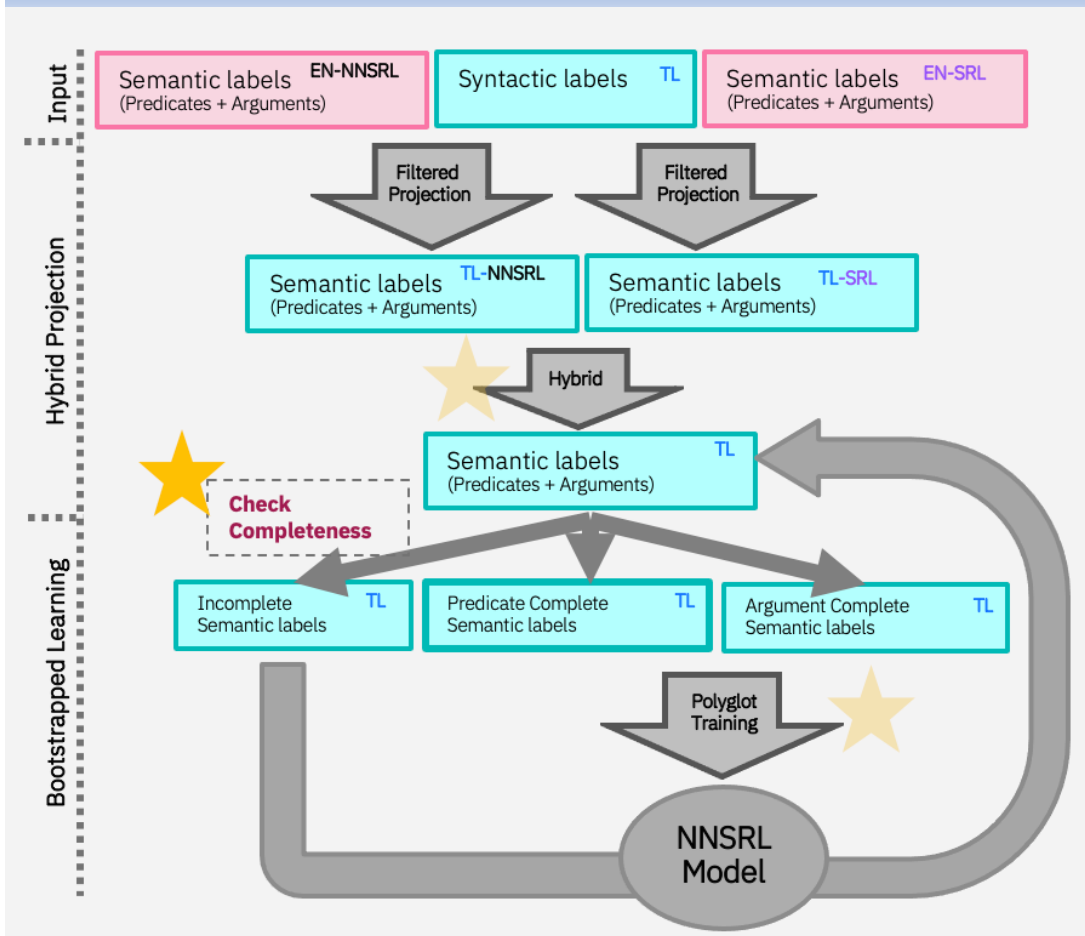
Dou, Z.-Y. and Neubig, G. (2021). Word alignment by fine-tuning embeddings on parallel corpora.

Universal Propositions 2.0 – zrównoleglenia słów

	W	rzecz	##yw	##istości	pro	##ste	w	##ys	##zuki	##wanie	w	Alta	##V	##ista
In	0,997151	0,423296	0,34415	0,445843	0,443248	0,468799	0,456038	0,364217	0,409283	0,494416	0,506391	0,398756	0,311068	0,400922
fact	0,404764	0,629365	0,490273	0,598919	0,485861	0,421796	0,428313	0,374891	0,363135	0,445087	0,385687	0,381551	0,351003	0,404279
,	0,448227	0,491219	0,520937	0,614135	0,508852	0,519261	0,516599	0,454458	0,432924	0,531479	0,498123	0,388248	0,357061	0,40614
a	0,443042	0,488407	0,417848	0,503012	0,577987	0,600183	0,615075	0,537218	0,46578	0,610088	0,53169	0,45178	0,355456	0,413962
simple	0,389304	0,481473	0,43781	0,493269	0,588606	0,603659	0,605551	0,574081	0,54541	0,592489	0,518548	0,475918	0,378793	0,458464
search	0,398947	0,411868	0,357442	0,439451	0,50354	0,478047	0,596796	0,585741	0,655389	0,631176	0,492155	0,431765	0,384263	0,431623
on	0,417674	0,437862	0,376641	0,465994	0,508853	0,515983	0,583656	0,56257	0,538832	0,622521	0,706927	0,493043	0,437719	0,489905
Alta	0,363763	0,403805	0,341307	0,37091	0,442394	0,45131	0,490049	0,475718	0,455567	0,442839	0,499345	0,937723	0,514889	0,563564
##V	0,280048	0,359816	0,458075	0,365545	0,407673	0,378255	0,464567	0,428334	0,347365	0,395476	0,430784	0,461736	0,940195	0,555179
##ista	0,368437	0,388931	0,388664	0,492938	0,413752	0,492276	0,46233	0,470773	0,433468	0,484561	0,486375	0,554944	0,557591	0,909832

	W	rzeczywisto	proste	wyszukiwal	w	AltaVista
In	0,997151	0,445843	0,468799	0,494416	0,506391	0,400922
fact	0,404764	0,629365	0,485861	0,445087	0,385687	0,404279
,	0,448227	0,614135	0,519261	0,531479	0,498123	0,40614
a	0,443042	0,503012	0,600183	0,615075	0,53169	0,45178
simple	0,389304	0,493269	0,603659	0,605551	0,518548	0,475918
search	0,398947	0,439451	0,50354	0,655389	0,492155	0,431765
on	0,417674	0,465994	0,515983	0,622521	0,706927	0,493043
AltaVista	0,368437	0,492938	0,492276	0,490049	0,499345	0,940195

Universal Propositions 2.0 – rzutowanie anotacji



Dwa modele EN trenowane na Ontonotes:

- NNSRL – wyższa dokładność (precision)
- K-SRL – niższa dokładność (precision), wyższa kompletność (recall) szczególnie dla identyfikacji predykatów

Predykaty i argumenty SRL zwracane przez te modele są niezależnie rzutowane do języka docelowego TL (Target Language).

Augmentacja zidentyfikowanych predykatów w wyniku NNSRL wynikami K-SRL.

Polyglot training - modele trenowane jednocześnie na danych dla języka angielskiego i docelowego języka pozwalają osiągnąć lepsze wyniki.

Universal Propositions 2.0 – rzutowanie anotacji

Celem rzutowania jest transfer semantycznych etykiet dla zdań angielskich do zdań w docelowym języku w oparciu o zrównoleglenia słów.

Filtrowane rzutowanie:

- **Verb Filter** - rzutowanie etykiety predykatu jedynie gdy źródłowy i docelowy token są czasownikami
- **Translation Filter** – rzutowanie jedynie gdy docelowy token jest tłumaczeniem źródłowego i dodatkowo jest to najbardziej popularne tłumaczenie
- **Reattachment Heuristic** – zapewnia, że etykiety są zawsze podłączone do nagłówków składniowych zgodnie z drzewem zależności

Universal Propositions 2.0 – kompletność rzutowania, trening iteracyjny

Kompletność predykatów (Predicate Completeness)

Zdanie w docelowym języku jest **kompletne** jeżeli posiada tą samą liczbę predykatów co odpowiadające mu zdanie w źródłowym języku – angielskim. Predykaty w języku angielskim są identyfikowane przez wytrenowany model dla języka angielskiego, a predykaty dla docelowego języka są otrzymywane poprzez rzutowanie anotacji.

Kompletność argumentów (Argument Completeness)

Dla każdego predykatu w zdaniu wybieramy podrzędne tokeny na podstawie drzewa zależności, które mają określony typ relacji [nsubj, obj, obl] i sprawdzamy, czy dla nich istnieje rzutowana etykieta argumentu. Para predykat-argumenty jest **k-kompletna** jeżeli zawiera k lub mniej różnic w argumentach pomiędzy wynikiem predykcji z wytrenowanego modelu dla języka angielskiego a wynikiem rzutowania anotacji. Dla $k = 0$ mamy pełną kompletność i taką wartość $k = 0$ przyjęliśmy w iteracyjnym treningu UP 2.0.

Iteracyjny trening

Wykorzystujemy ideę ponownego etykietowania (Van der Plas i in., 2011), które polega na przetrenowaniu modelu SRL na kompletnych w sensie predykatu i argumentów danych wyznaczonych w trakcie rzutowania anotacji, a następnie użyciu tego modelu SRL do predykcji na pozostałych niekompletnych danych. Wyniki predykcji zostają wykorzystane do ponownego etykietowania wyników rzutowania anotacji, nadpisując rzutowane etykiety potencjalnie mniej zaszumionymi etykietami z predykcji. Dzięki takiemu podejściu iteracyjnie poprawiamy kompletność (recall) starając się nie zmniejszać dokładności (precision).

Universal Propositions 2.0 - pliki GOLD

W ramach UP 2.0 do ewaluacji modeli wykorzystujemy 4 pliki GOLD:

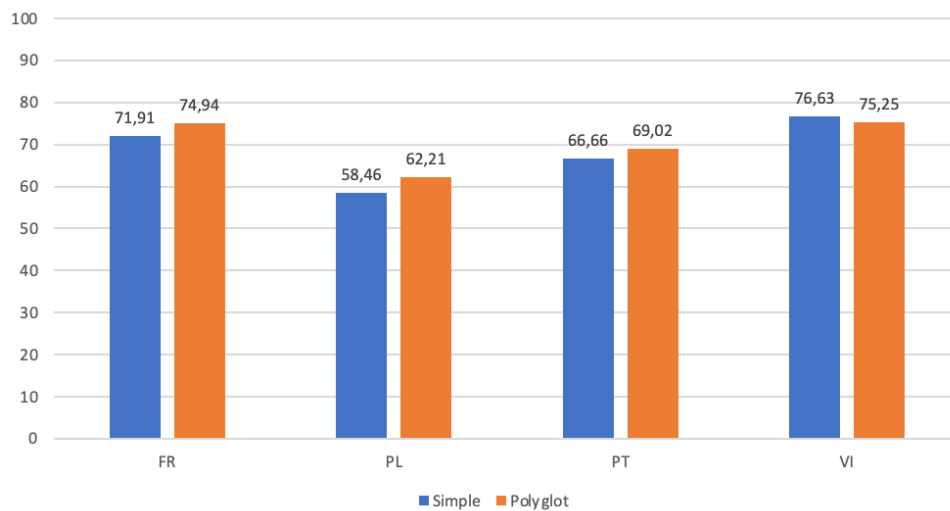
- Dla języka francuskiego (FR)*
- Dla PL, PT, VI przygotowaliśmy nowe pliki GOLD z anotacjami SRL

Language	Sentences	Predicates	Arguments
PL	100	223	495
PT	3779	6173	15097
VI	378	770	1625

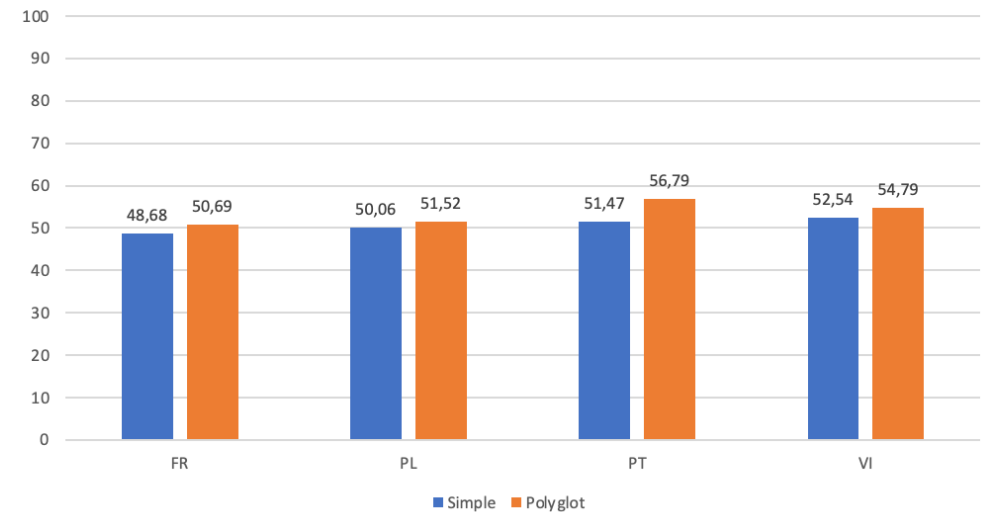
**Cross-lingual validity of PropBank in the manual annotation of French L Van der Plas, T Samardzic, P Merlo (2010)*

Universal Propositions 2.0 – wyniki w oparciu o GOLD: simple vs polyglot

Predicate identification and sense disambiguation (F1)



Arguments identification and classification (F1)



Universal Propositions 2.0 – co udostępniamy - Universal Propositions a Universal Dependencies

pl_lfg-ud-train.conllu

# sent_id = train-4											
# text = 19-letnia mieszkanka Koszalina przekonała kolegów, aby pomogli jej oszukać bank.											
# converted_from_file = NKJP1M_1305000000581_morph_1-p_morph_1.12-s-dis@1.xml											
# genre = news											
1	19-letnia	19-letni	ADJ	adj:sg:nom:f	Case=Nom G	2	amod	2:amod			
2	mieszkanka	mieszkanka	NOUN	subst:sg:nom	Case=Nom G	4	nsubj	4:nsubj			
3	Koszalina	Koszalin	PROPN	subst:sg:gen	Case=Gen G	2	nmod:poss	2:nmod:poss			
4	przekonała	przekonać	VERB	praet:sg:f:pe	Aspect=Perf	0	root	0:root			
5	kolegów	kolega	NOUN	subst:pl:acc:m	Case=Acc G	4	obj	4:obj	SpaceAfter=No		
6	,	,	PUNCT	interp	PunctType=C	8	punct	8:punct			
7	aby	aby	SCONJ	comp		8	mark	8:mark			
8	pomogli	pomóc	VERB	praet:pl:m1:j	Aspect=Perf	4	ccomp	4:ccomp			
9	jej	on	PRON	ppron3:sg:da	Case=Dat G	8	iobj	8:iobj 10:nsubj			
10	oszukać	oszukać	VERB	inf:perf	Aspect=Perf	8	xcomp	8:xcomp			
11	bank	bank	NOUN	subst:sg:acc:m	Case=Acc G	10	obj	10:obj	SpaceAfter=No		
12	.	.	PUNCT	interp	PunctType=P	4	punct	4:punct			



pl_lfg-up-train.conllup

# source_sent_id = conllu http://hdl.handle.net/11234/1-4611 UD_Po											
# sent_id = train-4											
# text = 19-let.....											
1											
2											
3											
4	persuade.01	A0:2 A1:5 A2:8	A0:1-3 A1:5-5 A2:7-11								
5											
6											
7											
8	help.01	A0:5 A1:10	A0:5-5 A1:10-11								
9											
10	deceive.01	A0:9 A1:11	A0:9-9 A1:11-11								
11											
12											

pl_lfg-ud-up-train.conllup



# global.columns = ID FORM LEMMA UPOS XPOS FEATS HEAD DEPREL DEPS MISC UP:PRED UP:ARGHEADS UP:ARGSPANS											
# sent_id = train-4											
# text = 19-letnia mieszkanka Koszalina przekonała kolegów, aby pomogli jej oszukać bank.											
# converted_from_file = NKJP1M_1305000000581_morph_1-p_morph_1.12-s-dis@1.xml											
# genre = news											
1	19-letnia	19-letni	ADJ	adj:sg:nom:f	Case=Nom G	2	amod	2:amod			
2	mieszkanka	mieszkanka	NOUN	subst:sg:nom	Case=Nom G	4	nsubj	4:nsubj			
3	Koszalina	Koszalin	PROPN	subst:sg:gen	Case=Gen G	2	nmod:poss	2:nmod:poss			
4	przekonała	przekonać	VERB	praet:sg:f:pe	Aspect=Perf	0	root	0:root	persuade.01	A0:2 A1:5 A2:8	A0:1-3 A1:5-5 A2:7-11
5	kolegów	kolega	NOUN	subst:pl:acc:m	Case=Acc G	4	obj	4:obj	SpaceAfter=		
6	,	,	PUNCT	interp	PunctType=C	8	punct	8:punct			
7	aby	aby	SCONJ	comp		8	mark	8:mark			
8	pomogli	pomóc	VERB	praet:pl:m1:j	Aspect=Perf	4	ccomp	4:ccomp	help.01	A0:5 A1:10	A0:5-5 A1:10-11
9	jej	on	PRON	ppron3:sg:da	Case=Dat G	8	iobj	8:iobj 10:nsubj			
10	oszukać	oszukać	VERB	inf:perf	Aspect=Perf	8	xcomp	8:xcomp	deceive.01	A0:9 A1:11	A0:9-9 A1:11-11
11	bank	bank	NOUN	subst:sg:acc:m	Case=Acc G	10	obj	10:obj	SpaceAfter=		
12	.	.	PUNCT	interp	PunctType=P	4	punct	4:punct			

▾ UD_Polish-LFG
 ≡ pl_lfg-ud-up-dev.conllup
 ≡ pl_lfg-ud-up-test.conllup
 ≡ pl_lfg-ud-up-train.conllup
 ▾ UD_Polish-PDB
 ≡ pl_pdb-ud-up-dev.conllup
 ≡ pl_pdb-ud-up-test.conllup
 ≡ pl_pdb-ud-up-train.conllup

Universal Propositions 2.0 – ramki

PropBank

- [abase.html](#)
- [abash.html](#)
- [abate.html](#)
- [abbreviate.html](#)
- [abdicate.html](#)
- [abdominal.html](#)
- [abduct.html](#)
- [aberrate.html](#)
- [abet.html](#)
- [abhor.html](#)
- [abide.html](#)
- [ablate.html](#)
- [able.html](#)
- [abnegate.html](#)
- [abnormal.html](#)
- [abolish.html](#)
- [abominate.html](#)
- [abort.html](#)
- [abound.html](#)
- [about.html](#)
- [abrade.html](#)
- [abridge.html](#)
- [abrogate.html](#)
- [absent.html](#)
- [absolute.html](#)
- [absolve.html](#)
- [absorb.html](#)
- [abstain.html](#)
- [abstract.html](#)



UP 2.0

- [brzmieć.html](#)
- [budować.html](#)
- [budzić.html](#)
- [burzyć.html](#)
- [być.html](#)
- [bywać.html](#)
- [całować.html](#)
- [cechować.html](#)
- [celebrować.html](#)
- [cenić.html](#)
- [charakteryzować.html](#)
- [chcieć.html](#)
- [chłubić.html](#)
- [chodźć.html](#)
- [chować.html](#)
- [chronić.html](#)
- [chwalić.html](#)
- [chwycić.html](#)
- [ciąć.html](#)
- [ciągnąć.html](#)
- [ciążyć.html](#)
- [cierpieć.html](#)
- [cieszyć.html](#)
- [cofać.html](#)
- [cofnąć.html](#)
- [cytować.html](#)
- [czekać.html](#)
- [czerpać.html](#)
- [czuć.html](#)

Predicate: *chodźć*

Roleset id: **chodźć.04**, refer, call, label, refer to **Source:** [refer.01](#) (*refer.01*)

Roles:

A0: *labeller*
A1: *labelled*
A2: *label*

Example:

Chodzi mi przede wszystkim o zwalczanie terroryzmu i przestępczości zorganizowanej .

Rel: Chodzi

A0: mi

AM-TMP: przede wszystkim

A1: o zwalczanie terroryzmu i przestępczości zorganizowanej

Predicate: *chodźć*

Roleset id: **chodźć.05**, walk, walk, be a pedestrian, forward motion, one foot in front of the other **Source:** [walk.01](#) (*walk.01*)

Roles:

A0: *walker, thing in motion*
A1: *entity being walked, where there is outside Agent*
A2: *path/ underspecified location walked*

Example:

Tom i Mary lubią **chodźć** razem po lasach .

A0: Tom i Mary

AM-DIS: lubią .

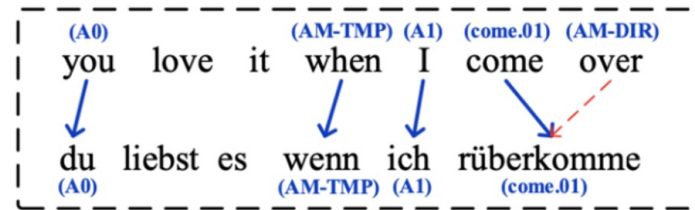
Rel: chodźć

AM-MNR: razem

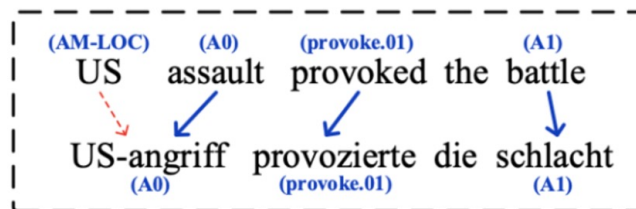
AM-LOC: po lasach

Universal Propositions 2.0 – zidentyfikowane problemy związane z rzutowaniem EN-> PL

Handling collisions*:



(b) Predicate-argument collision. Only keep predicate.



(c) Argument-argument collision. Only keep the one with higher confidence.

* Cross-Lingual Semantic Role Labeling with High-Quality Translated Training Corpus Hao Fei and Meishan Zhang and Donghong Ji

Universal Propositions 2.0 – zidentyfikowane problemy związane z rzutowaniem EN -> PL

Predicate one-to-many (pomijamy? zostawiamy predicate for HEAD? Token w języku polskim nie jest czasownikiem ...)

English				Word alignments	Polish			
role/context	verbClass	upos	token		token	upos	verbClass	role/context
	reevaluate.01	VERB	re-evaluating		ponownej	ADJ	reevaluate.01	
					oceny	NOUN	reevaluate.01	

Predicate many-to-one (rzutujemy tylko ostatni?)

English				Word alignments	Polish			
role/context	verbClass	upos	token		token	upos	verbClass	role/context
	let.01	VERB	let		zapomnijmy	VERB	let.01/forget.01	
		PRON	us					
	forget.01	VERB	forget					

Role or Modifier for Predicate (pomijamy modifier)

English				Word alignments	Polish			
role/context	verbClass	upos	token		token	upos	verbClass	role/context
		PRON	There		Nie	PART		
AM-MOD		AUX	will		będzie	VERB	be.02	AM-MOD
	be.02	VERB	be					
		DET	no					

Universal Propositions 2.0 – zidentyfikowane problemy związane z rzutowaniem EN -> PL

Role for Predicate (nie pozwalamy? Tracimy ważną informację o roli!)

English				Word	Polish			
role/context	verbClass	upos	token	alignments	token	upos	verbClass	role/context
A0		PRON	I		myślałem	VERB	think.01	A0
	think.01	VERB	thought					

Modifier tries to override Role (nie pozwalamy?)

English				Word	Polish			
role/context	verbClass	upos	token	alignments	token	upos	verbClass	role/context
AM-DIS		CCONJ	and		i	CCONJ		AM-DIS
A0		PRON	I		będę	AUX		AM-MOD/A0
AM-MOD		AUX	'll		korzystać	VERB	use.01	
	use.01	VERB	use					

Role tries to override Modifier (pozwalamy?)

English				Word	Polish			
role/context	verbClass	upos	token	alignments	token	upos	verbClass	role/context
AM-NMR			How		Jak			AM-NMR
AM-MOD			could		mógł			AM-MOD/A0
A0			he		mieć		have.03	
	have.03		have					

Universal Propositions 2.0 – zidentyfikowane problemy związane z rzutowaniem EN -> PL

Ramka angielska oparta jest na rzeczowniku.

Występuje tutaj także konflikt argumentów.

English				Word	Polish			
role/context	verbClass	upos	token	alignments	token	upos	verbClass	role/context
A0			Hamas		Hamas		A0	
AM-MOD			would		zawarłby			AM-MOD/AM-LVB
AM-LVB			make		pokój		peace.01	
	peace.01		peace					

Modele SRL – architektura

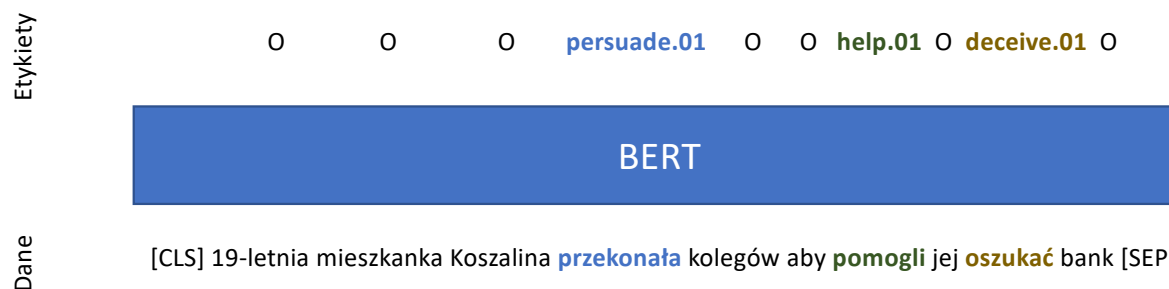
Dwa oddzielne modele:

- model identyfikacji predykatów i ich ujednoznaczniania
- model identyfikacji argumentów i ich klasyfikacji

Obydwa modele oparte na bert-base-multilingual-cased - zadanie klasyfikacji tokenów (podobnie jak dla NER)

Modele SRL – architektura

Model klasyfikacji predykatów i ujednoznacznienia sensu



Modele SRL – architektura

Model identyfikacji i klasyfikacji argumentów



[CLS] zdanie [SEP] predykat/sens [SEP]

Modele SRL- predykcja

[CLS] 19-letnia mieszkanka Koszalina przekonała kolegów aby pomogli jej oszukać bank [SEP]



Model identyfikacji predykatów i
ujednoznaczniania sensu



przekonała – persuade.01

pomogli – help.01

oszukać – deceive.01

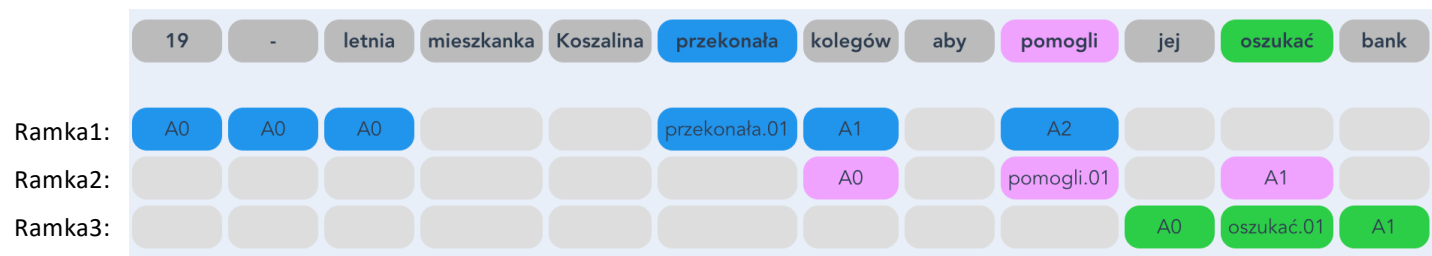
Ramka1: [CLS] 19-letnia mieszkanka Koszalina przekonała kolegów aby pomogli jej oszukać bank [SEP] persuade.01 [SEP]

Ramka2: [CLS] 19-letnia mieszkanka Koszalina przekonała kolegów aby pomogli jej oszukać bank [SEP] help.01 [SEP]

Ramka3: [CLS] 19-letnia mieszkanka Koszalina przekonała kolegów aby pomogli jej oszukać bank [SEP] deceive.01 [SEP]

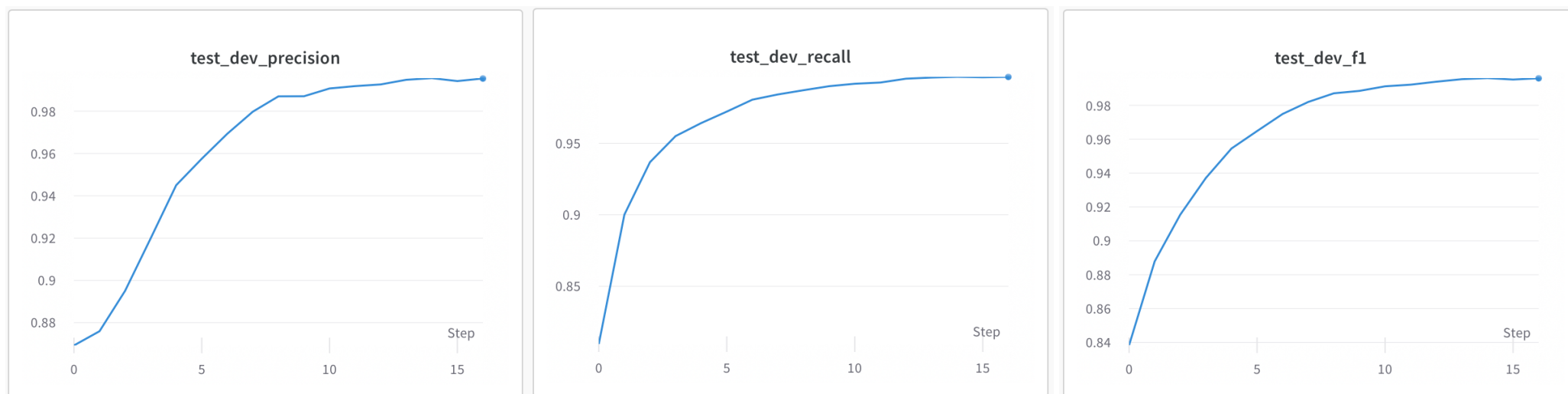


Model identyfikacji i klasyfikacji
argumentów



Modele SRL – trening PRED EN Ontonotes

Ewaluacja modelu w oparciu o zbiór **treningowy** po każdej epoce



Po 13 epoce:

```
"micro_avg": {  
  "precision": 0.9948653343259081,  
  "recall": 0.9960060387165817,  
  "f1-score": 0.9954353597280315  
}
```

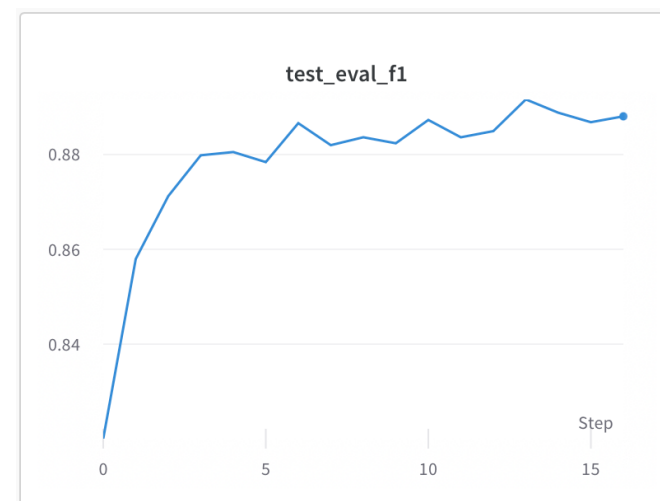
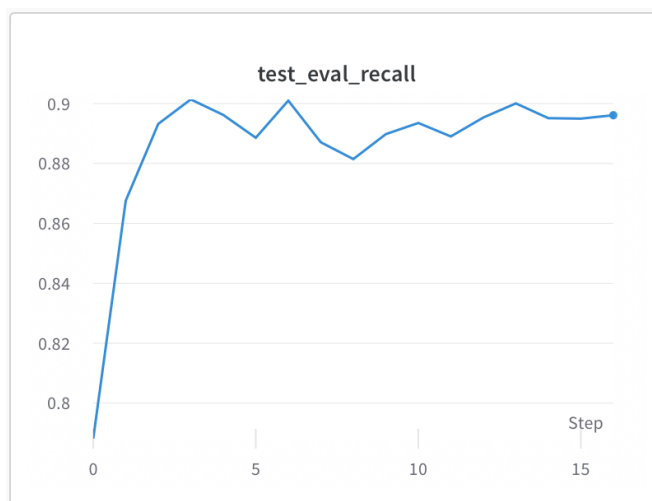
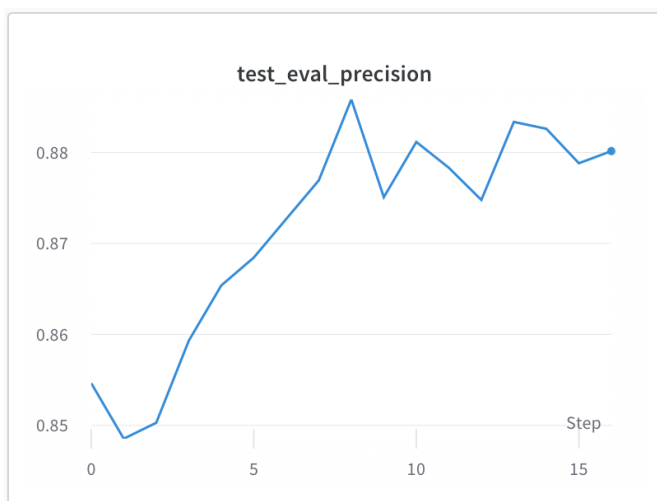

Modele SRL – trening PRED EN Ontonotes

Postęp treningu (F1) – pierwsze 33 z 5455 pozycji

		Epoch																
Predicate/Sense	Support	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
be.02	6271	0.93	0.95	0.95	0.97	0.98	0.98	0.98	0.98	0.99	0.99	0.99	1	1	1	0.99	1	1
smell.01	46	0.08	0.82	0.73	0.96	0.9	0.96	0.95	1	0.95	1	1	1	1	1	1	0.93	1
pass.02	56	0	0	0.12	0.41	0.64	0.77	0.88	0.83	0.89	0.93	0.97	0.98	1	1	0.97	1	1
cross.02	88	0.6	0.81	0.8	0.83	0.88	0.91	0.92	0.99	0.98	0.98	0.93	0.99	1	0.96	0.98	1	1
range.01	134	0.94	0.98	0.99	0.99	0.99	0.99	0.99	1	0.98	0.99	1	0.99	0.99	1	0.99	0.99	1
be.01	73665	0.96	0.97	0.98	0.98	0.99	0.99	0.99	1	1	1	1	1	1	1	1	1	1
prospect.02	26	0	0	0.72	0.53	0.59	0.58	0.63	0.79	0.9	0.93	1	0.84	0.9	1	1	1	1
call.01	1551	0.85	0.94	0.97	0.98	0.96	0.98	0.98	0.99	0.99	1	0.99	0.99	1	1	1	1	1
sign.02	227	0	0.76	0.82	0.89	0.91	0.91	0.93	0.9	0.97	0.99	0.98	1	1	1	1	1	1
have.01	26790	0.98	0.99	0.99	0.99	0.99	1	1	1	1	1	1	1	1	1	1	1	1
prove.01	502	0.9	0.96	0.98	0.98	0.99	0.96	0.99	1	1	0.99	0.99	1	1	1	1	1	1
adapt.01	76	0.1	0.9	0.95	0.97	0.99	1	1	1	1	1	1	1	1	1	1	0.99	1
converge.01	10	0	0	0.89	1	0.83	0.91	1	1	1	1	1	1	1	1	1	1	1
come.01	4481	0.89	0.9	0.93	0.95	0.96	0.97	0.98	0.98	0.99	0.99	0.99	0.99	0.99	1	1	0.99	1
get.03	1381	0.83	0.9	0.9	0.94	0.96	0.96	0.98	0.98	0.98	0.95	0.98	0.99	0.96	1	0.99	0.99	0.99
involve.01	709	0.9	0.91	0.92	0.93	0.93	0.93	0.97	0.99	1	1	0.99	1	1	1	0.99	1	1
contribute.01	308	0.85	0.85	0.91	0.9	0.92	0.99	0.97	0.98	0.98	0.99	1	1	1	1	1	1	1
inflamm.01	20	0	0.89	0.95	0.87	0.95	0.95	0.95	0.95	0.95	0.95	0.95	1	1	1	1	1	1
purchase.01	457	0.86	0.87	0.88	0.9	0.97	0.97	0.98	0.98	0.99	1	1	1	1	1	1	1	1
make.LV	1757	0.73	0.83	0.86	0.92	0.94	0.96	0.97	0.98	0.99	0.99	0.98	0.99	0.99	0.99	1	0.99	0.99
say.01	23239	0.98	0.98	0.99	0.99	0.99	1	1	1	1	1	1	1	1	1	1	1	1
pick_up.04	215	0.62	0.88	0.84	0.88	0.89	0.95	0.93	0.97	0.98	0.98	0.99	0.99	0.99	0.98	0.98	0.99	0.99
be.03	54917	0.94	0.95	0.97	0.97	0.98	0.98	0.99	0.99	0.99	0.99	1	1	1	1	1	1	1
authorize.01	78	0	0.97	0.99	1	0.99	0.99	1	1	1	0.99	0.99	1	1	1	1	1	1
order.01	331	0.76	0.82	0.82	0.89	0.95	0.99	0.97	0.99	1	1	1	0.99	0.93	1	1	0.99	1
probation.01	20	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
mean.01	2108	0.97	0.98	0.98	0.98	0.99	0.99	1	0.99	1	1	1	1	1	1	1	1	1
do.LV	263	0	0.49	0.64	0.79	0.87	0.93	0.95	0.98	0.95	0.96	0.99	0.98	0.97	0.99	0.94	0.97	1
write.01	1367	0.91	0.93	0.97	0.96	0.99	0.99	0.99	1	1	1	1	1	1	1	1	1	1
bear.02	360	0.89	0.91	0.87	0.92	0.95	0.97	0.96	0.98	0.99	1	0.99	0.99	1	1	1	1	1
have.03	10528	0.9	0.93	0.95	0.96	0.96	0.98	0.98	0.99	0.99	0.99	1	1	1	1	1	0.99	1
festoon.01	6	0	0	1	1	1	0.8	1	1	1	1	1	1	1	1	1	1	1
go.02	3764	0.83	0.86	0.87	0.93	0.94	0.97	0.97	0.98	0.99	0.98	0.99	0.99	1	0.99	0.99	0.99	0.99
take.01	2795	0.78	0.82	0.87	0.91	0.92	0.95	0.96	0.97	0.99	0.98	0.99	0.99	0.99	0.99	0.99	0.99	0.99
know.01	5065	0.82	0.81	0.89	0.9	0.91	0.93	0.94	0.93	0.96	0.94	0.96	0.95	0.97	0.97	0.99	0.98	0.98
call.02	518	0.48	0.74	0.84	0.87	0.91	0.91	0.96	0.97	0.98	0.97	0.99	0.99	0.98	0.96	0.99	0.99	1
ask.01	1064	0.88	0.91	0.97	0.98	0.99	0.96	0.99	1	0.99	1	1	1	0.99	1	1	1	1

Modele SRL – trening PRED EN Ontonotes

Ewaluacja modelu w oparciu o zbiór **testowy** po każdej epoce

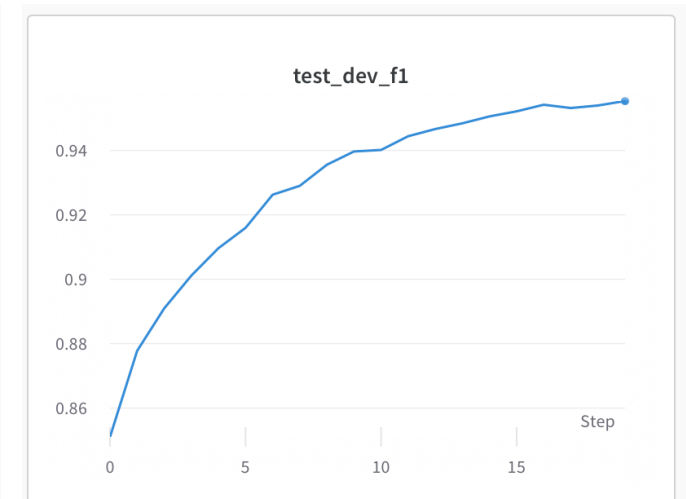
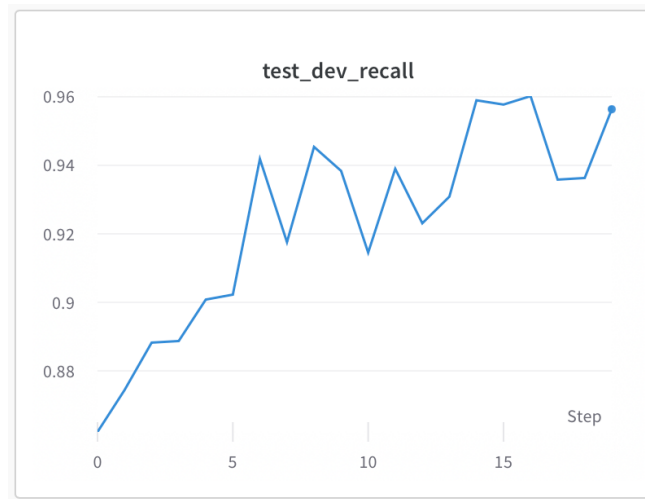
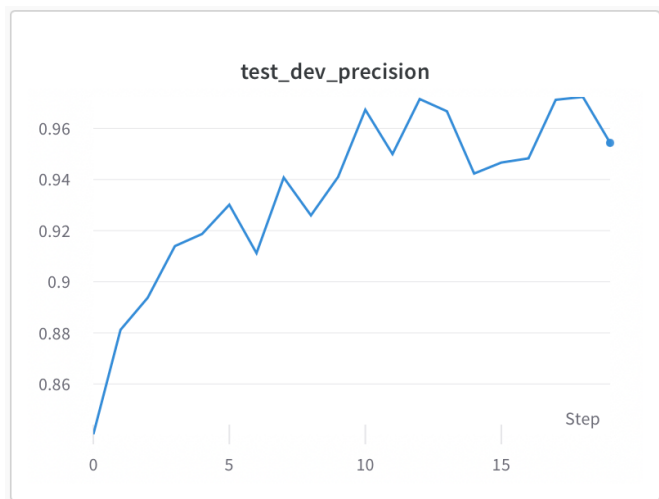


Po 13 epoce:

```
"micro_avg": {  
  "precision": 0.8833605713939834,  
  "recall": 0.900058589534059,  
  "f1-score": 0.8916314093262666  
}
```

Modele SRL – trening ARG EN Ontonotes

Ewaluacja modelu w oparciu o zbiór **treningowy** po każdej epoce



Po 16 epoce:

```
"micro_avg": {  
  "precision": 0.9482844122263591,  
  "recall": 0.9602106863430102,  
  "f1-score": 0.9542102853654814  
}
```

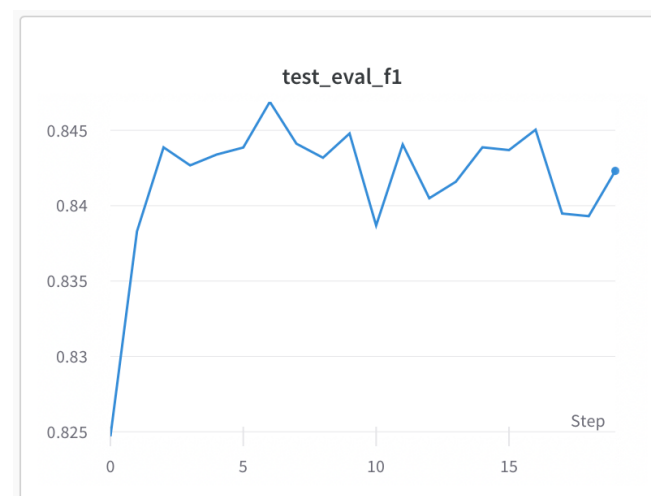
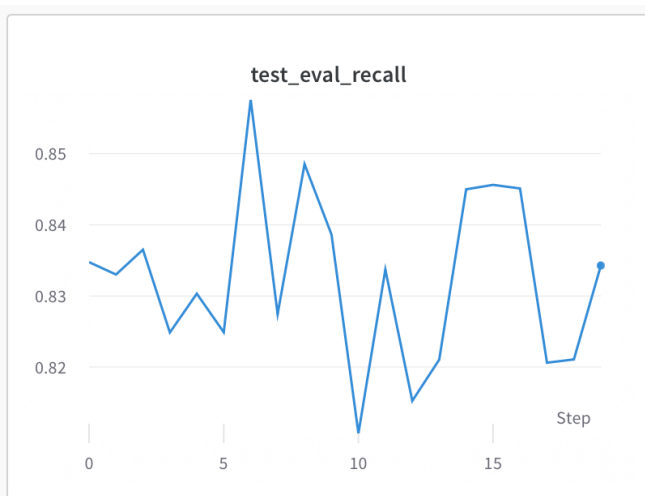
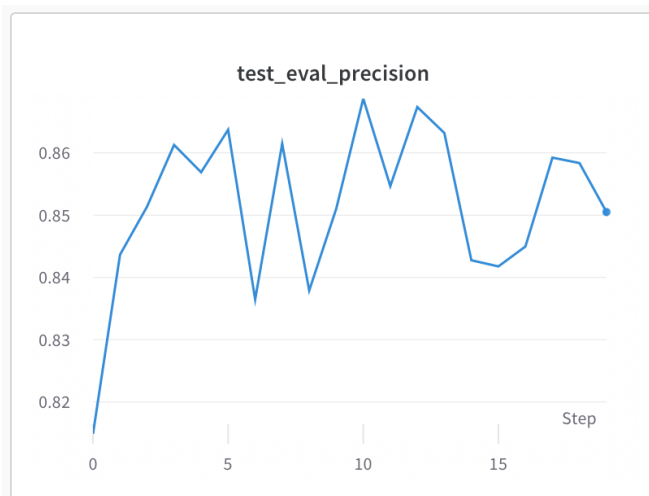
Modele SRL – trening ARG EN Ontonotes

Postęp treningu (F1)

		0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
A1	446955	0,9	0,92	0,93	0,93	0,94	0,94	0,95	0,95	0,95	0,95	0,95	0,95	0,95	0,95	0,96	0,96	0,96	0,95	0,96	0,96
A2	152905	0,82	0,85	0,88	0,88	0,89	0,9	0,91	0,91	0,91	0,91	0,91	0,91	0,92	0,92	0,92	0,93	0,93	0,93	0,92	0,92
A0	265866	0,92	0,93	0,94	0,95	0,96	0,96	0,96	0,96	0,97	0,97	0,97	0,97	0,97	0,97	0,97	0,97	0,97	0,97	0,97	0,97
AM-LOC	30886	0,71	0,79	0,82	0,85	0,87	0,88	0,91	0,92	0,93	0,94	0,95	0,95	0,95	0,95	0,95	0,96	0,96	0,96	0,96	0,96
AM-MOD	48894	0,94	0,94	0,95	0,95	0,95	0,95	0,95	0,95	0,95	0,95	0,95	0,96	0,95	0,95	0,95	0,96	0,96	0,96	0,96	0,96
AM-ADV	46315	0,64	0,68	0,69	0,7	0,72	0,75	0,78	0,79	0,83	0,85	0,86	0,86	0,88	0,9	0,9	0,91	0,92	0,92	0,92	0,93
AM-DIS	45343	0,81	0,83	0,83	0,84	0,85	0,86	0,88	0,88	0,9	0,9	0,91	0,92	0,93	0,92	0,93	0,94	0,94	0,94	0,95	0,95
C-A1	7302	0,56	0,66	0,7	0,73	0,79	0,81	0,85	0,86	0,9	0,9	0,9	0,92	0,92	0,93	0,93	0,94	0,94	0,94	0,94	0,95
AM-TMP	84197	0,85	0,87	0,89	0,9	0,91	0,91	0,93	0,92	0,94	0,94	0,95	0,95	0,95	0,95	0,95	0,96	0,96	0,96	0,96	0,96
AM-DIR	7886	0,63	0,71	0,74	0,79	0,82	0,83	0,87	0,88	0,87	0,91	0,92	0,93	0,93	0,94	0,95	0,96	0,96	0,96	0,96	0,96
AM-MNR	32586	0,71	0,76	0,8	0,82	0,84	0,86	0,88	0,9	0,91	0,92	0,93	0,94	0,94	0,95	0,95	0,95	0,95	0,96	0,96	0,96
AM-PRR	3560	0,97	0,97	0,98	0,98	0,98	0,99	0,99	0,99	0,99	0,99	0,99	0,99	0,99	0,99	0,99	0,99	0,99	0,99	0,99	0,99
AM-NEG	23579	0,89	0,91	0,92	0,92	0,92	0,92	0,92	0,92	0,93	0,93	0,93	0,93	0,93	0,93	0,93	0,93	0,94	0,94	0,93	0,93
AM-PRD	7573	0,25	0,36	0,38	0,42	0,49	0,54	0,56	0,68	0,71	0,77	0,78	0,79	0,84	0,87	0,88	0,89	0,91	0,92	0,91	0,92
AM-ADJ	11241	0,68	0,73	0,77	0,82	0,85	0,85	0,9	0,9	0,91	0,92	0,93	0,94	0,94	0,94	0,94	0,94	0,95	0,95	0,95	0,95
R-A0	15473	0,92	0,94	0,94	0,96	0,96	0,96	0,97	0,97	0,97	0,97	0,97	0,97	0,97	0,97	0,97	0,97	0,97	0,97	0,97	0,97
A4	8502	0,81	0,83	0,86	0,89	0,9	0,91	0,93	0,93	0,93	0,94	0,95	0,95	0,95	0,96	0,96	0,96	0,97	0,96	0,97	0,97
AM-CAU	8097	0,72	0,75	0,77	0,78	0,81	0,84	0,86	0,85	0,87	0,87	0,89	0,89	0,89	0,9	0,91	0,91	0,91	0,91	0,92	0,92
AM-LVB	3646	0,81	0,85	0,88	0,92	0,92	0,94	0,95	0,95	0,94	0,96	0,95	0,95	0,96	0,97	0,97	0,96	0,97	0,96	0,97	0,97
R-A2	969	0,57	0,78	0,83	0,85	0,88	0,89	0,93	0,93	0,93	0,94	0,93	0,94	0,95	0,95	0,94	0,94	0,95	0,95	0,95	0,95
AM-GOL	2566	0,31	0,42	0,51	0,56	0,63	0,63	0,73	0,73	0,78	0,8	0,85	0,83	0,86	0,88	0,91	0,9	0,91	0,9	0,92	0,92
AM-PRP	10515	0,67	0,74	0,77	0,79	0,83	0,85	0,88	0,89	0,9	0,92	0,92	0,93	0,94	0,94	0,94	0,95	0,95	0,96	0,96	0,95
R-A1	12638	0,84	0,88	0,89	0,89	0,89	0,91	0,92	0,91	0,92	0,91	0,91	0,91	0,92	0,91	0,92	0,91	0,92	0,91	0,91	0,92
A3	11034	0,63	0,75	0,78	0,83	0,86	0,87	0,9	0,91	0,92	0,94	0,94	0,94	0,95	0,96	0,96	0,96	0,96	0,96	0,96	0,97
C-A0	605	0	0,2	0,35	0,52	0,6	0,67	0,74	0,79	0,82	0,82	0,86	0,86	0,88	0,91	0,87	0,87	0,91	0,9	0,93	0,92
AM-EXT	4296	0,54	0,63	0,65	0,73	0,75	0,73	0,79	0,84	0,85	0,87	0,91	0,91	0,91	0,93	0,93	0,93	0,94	0,94	0,95	0,95
R-AM-LOC	1710	0,79	0,84	0,85	0,85	0,89	0,9	0,92	0,94	0,89	0,94	0,93	0,94	0,92	0,94	0,95	0,96	0,95	0,95	0,94	0,95
C-A2	1168	0,46	0,45	0,59	0,65	0,68	0,72	0,76	0,79	0,81	0,83	0,84	0,85	0,84	0,86	0,87	0,88	0,88	0,88	0,87	0,88
R-AM-TMP	1129	0,68	0,76	0,79	0,81	0,84	0,87	0,89	0,89	0,87	0,92	0,91	0,92	0,88	0,94	0,94	0,94	0,94	0,94	0,92	0,95
AM-COM	999	0,43	0,57	0,55	0,67	0,7	0,72	0,81	0,77	0,82	0,87	0,92	0,91	0,9	0,93	0,93	0,93	0,93	0,91	0,93	0,94
R-AM-MNR	131	0,51	0,71	0,75	0,75	0,78	0,85	0,87	0,87	0,84	0,9	0,87	0,9	0,92	0,76	0,93	0,93	0,89	0,9	0,89	0,92
AM-REC	324	0,43	0,65	0,68	0,72	0,72	0,76	0,82	0,81	0,84	0,88	0,92	0,89	0,87	0,89	0,89	0,93	0,93	0,93	0,9	0,93
C-AM-TMP	51	0	0	0	0	0	0	0	0	0,15	0,28	0,08	0,28	0,48	0,53	0,57	0,47	0,61	0,68	0,7	0,74
C-AM-LOC	77	0,09	0,67	0,62	0,7	0,66	0,68	0,75	0,75	0,84	0,81	0,92	0,88	0,89	0,91	0,91	0,91	0,92	0,87	0,94	0,88
A1-DSP	1872	0,01	0,36	0,69	0,24	0,68	0,77	0,84	0,8	0,83	0,83	0,87	0,92	0,93	0,94	0,93	0,94	0,93	0,94	0,96	0,89
C-A1-DSP	1896	0,02	0,23	0,68	0,59	0,7	0,76	0,81	0,8	0,79	0,89	0,92	0,92	0,94	0,94	0,95	0,94	0,94	0,94	0,95	0,91
C-A4	49	0	0	0,33	0,2	0,59	0,51	0,57	0,73	0,8	0,85	0,82	0,86	0,87	0,91	0,91	0,92	0,92	0,95	0,92	0,9
C-A3	105	0,2	0,51	0,34	0,69	0,71	0,76	0,64	0,78	0,84	0,88	0,91	0,92	0,89	0,95	0,89	0,89	0,91	0,88	0,97	0,97
C-AM-MNR	94	0	0	0,18	0,29	0,49	0,54	0,63	0,65	0,69	0,77	0,82	0,83	0,81	0,94	0,91	0,9	0,91	0,94	0,96	0,89
C-AM-DIR	32	0	0,12	0,57	0,71	0,82	0,69	0,8	0,91	1	0,94	0,94	0,93	0,97	1	0,91	1	0,94	1	1	1

Modele SRL – trening ARG EN Ontonotes

Ewaluacja modelu w oparciu o zbiór **testowy** po każdej epoce



Po 16 epoce:

```
"micro_avg": {  
  "precision": 0.8449655049363626,  
  "recall": 0.845116292903456,  
  "f1-score": 0.8450408921933086  
}
```

Modele SRL - wyniki

Język	Dane	Model	P	R	F1
EN	ontonotes	predicate disambiguation	0,88	0,90	0,89
		argument detection/labeling	0,84	0,85	0,85

Podsumowanie

Potencjalne kierunki prac:

- Budowa/treningi własnych modeli SRL dla języka polskiego
- Eksperymenty z innymi niż bert-base-multilingual-* modelami bazowymi
- Ulepszanie mechanizmu zrównoleglania słów
- Ulepszanie mechanizmów rzutowania anotacji, zaadresowanie specyficznych dla języka polskiego problemów
- Budowa modeli opartych na połączeniu reprezentacji BERT i reprezentacji pochodzących z SRL
- Modele i eksperymenty z zero-shot transfer learning z EN do PL w oparciu o modele multilingual
- Manualna budowa polskiego propbank'u dla SRL
- Aplikacja WEB do oznaczania zdań anotacjami SRL pod kątem manualnej budowy propbank'ów w różnych językach