Towards Semantic Role Labeling in Slovene and Croatian

Polona Gantar,* Kristina Štrkalj Despot,**
Simon Krek,† Nikola Ljubešić‡

* Department of translation, Faculty of Arts, University of Ljubljana
** Institute for the Croatian Language and Linguistics
† Artificial Intelligence Laboratory, Jožef Stefan Institute
‡ Department of Knowledge Technologies, Jožef Stefan Institute

Semantic Role Labeling (SRL)

- the process of detecting and assigning semantic roles to semantic arguments determined by the predicate or verb of a sentence.
- SRL is important for information extraction systems, question answering systems, improving syntactic parsing systems, machine translation etc.
- a very important step towards the understanding of the meaning of a sentence.

SRL in Slovene and Croatian

- both Slovene and Croatian may be considered as under-resourced languages in terms of language technologies, especially in the area of machine readable semantic resources and advanced tools for the processing of those resources (Krek et al. 2012)
- SRL will improve the existing levels of linguistic annotation of both Slovene and Croatian training corpora.

Semantic Role Labeling in Slovene and Croatian Project

- The **aim** of the project was to build a semantic role labeling system which will be added to the existing syntactic dependencies in both Slovene and Croatian training corpora used hitherto for machine learning algorithms.

The core project tasks

- 1) development of the common Slovene-Croatian semantic annotation scheme (entirely based on the previously developed SRL tagset from Krek et al. 2016.)
- 2) compiling the instructions for annotation;
- 3) manual annotation of the sample parts of both learning corpora using compatible tags.

This served as the basis for the automatic annotation experiments using supervised machine learning methods, performed later on both corpora.

Semantic Role Labeling framework for Slovene and Croatian

- In compiling the list of semantic roles and their respective formal descriptions, we follow the approach developed by:
- Prague Dependency Treebank, PDT (Mikulová et al., 2005)
- Valency Lexicon of Czech Verbs (Vallex)
- semantic role labeling within Croatian Dependency Treebank (SRL tagset compiled by Filko et al. 2012)
- Crovallex (Croatian version of Czech Vallex) (Mikelić Preradović et al. 2009).

SRL tagset

role	tag	role	tag
actant	ACT	cause	CAUSE
patient	PAT	aim	AIM
recipient	REC	condition	COND
origin	ORIG	regard	REG
result	RESLT	accompaniment	ACMP
location	LOC	restriction	RESTR
source	SOURCE	manner	MANN
goal	GOAL	means	MEANS
event	EVENT	quantification	QUANT
time	TIME	multi-word predicate	MWPRED
duration	DUR	modal	MODAL
frequency	FREQ	Phraseological unit	PHRAS

Corpora and Tools for Annotation SLO

- the **SSJ500k 2.0** (Krek et al. 2015) corpus
- 500,293 words (27,829 sentences) sampled from the FidaPLUS corpus (Arhar Holdt and Gorjanc 2007)
- manually annotated on morphosyntactic level (Grčar et al. 2012)
- partially on the syntactic level (Dobrovoljc et al. 2012)
- Named entities and multi-word expressions are also identified (Gantar et al. 2017)
- The total of 5,491 sentences were annotated with semantic roles
- the first 500 sentences used for test annotation by 4 annotators
- second phase: automatic annotation of the remaining 4,991 sentences
- manual check by 5 annotators
- Group discussions of problematic cases, consensus, no IAA

Corpora and Tools for Annotation CRO

- the SETimes.HR part of the hr500k corpus (Ljubešić et al. 2018)
- 3,757 sentences manually lemmatized and morphosyntactically tagged (Agić et al., 2013)
- annotated for syntactic dependencies using the Universal Dependencies formalism (Agić and Ljubešić, 2015)
- these sentences were being manually semantically annotated by 2 annotators.
- discussions of problematic cases, consensus, no IAA
- This then served as the resource for automatic labeling and quantitative analysis.

Automatic Semantic Role Labeling

- Both annotated corpora were split in training and test data in a 80:20 fashion. This data split is available for each of the languages at https://github.com/clarinsi/bilateral-srl/tree/master/data.
- the well-known baseline mate-tools semantic role labeler (Björkelund et al. 2009) was benchmarked on the data
- weighted F1 score for all classes for Croatian was 0.72, while for Slovene it was 0.75.
- The data on both languages are quite similar, with F1 metrics correlating to the frequency of each phenomenon (coefficients of 0.517 and 0.611)

Verbs with frequency f>=50 in SSJ500k and SETimes.HR.

biti	7203	**biti**	4969
imeti	333	**htjeti**	670
morati	178	**kazati**	276
iti	114	izjaviti	210
vedeti	95	**moći**	195
dobiti	83	**imati**	163
moči	83	**reći**	160
začeti	80	trebati	146
videti	75	**morati**	117
reči	74	**željeti**	65
priti	72	očekivati	62
povedati	72	**dobiti**	57
hoteti	69	**postati**	57
želeti	59	postojati	56
postati	54	priopćiti	54
govoriti	51	predstavljati	53
misliti	50	navoditi	50

Syntactic-semantic patterns - SLO

‘to have’ imeti (333)

● WHO (ACT) has WHAT (PAT 316) [for WHOM (REC), from whom (ORIG), where (LOC), when (TIME) ...]: Na zadnji hrbtni bodici ima veliko črno piko.

‘must’ morati (178)

● WHO (ACT) must INF (MODAL): Država bi morala plačati stroške presoje vplivov na okolje.

‘to go’ iti (114)

● WHO (ACT) goes WHERE (GOAL) [how(MANN), when (TIME), under what conditions (COND) …]: Šel sem prvič k vedeževalki.

● to go (PHRAS 11): Zgodba mi ni in ni šla iz glave

● to go SUPINE (MWPRED): Verjetno bom šla smučat na Krvavec.

‘to get’ dobiti (83)

● WHO (ACT) gets WHAT (PAT) [from whom (ORIG), in regard to what (REG), with what means (MEANS), when (TIME), under what conditions (COND) …]: Mala je dobila ime po Prometeju

Syntactic-semantic patterns - CRO

‘to want’ htjeti (670), željeti (65)

WHO (ACT) wants WHAT (PAT) [for WHOM (REC), from WHOM (ORIG)...]: Oni žele autonomiju sjevera, a za druge enklave žele takozvani Ahtisaari plus.

WHO (ACT) wants INF (MODAL) [(WHAT) (PAT)]: Mnoge žrtve ne žele podnijeti tužbu.

‘to tell, say’ kazati (276), izjaviti (210), reći (160)

WHO (ACT) says WHAT (RESLT) to WHOM (REC) about WHAT (PAT) [WHERE (LOC), WHEN (TIME)]: “U suprotnom ćemo biti neozbiljni političari”, rekao je Lagumdžija novinarima u Beogradu nakon sastanka s Jeremićem 14. ožujka

‘can’ moći (195)

WHO (ACT) can INF (MODAL) WHAT (PAT): Privatizacija je mogla donijeti bolje usluge

‘to have’ imati (163)

WHO (ACT) has WHAT (PAT) [WHEN (TIME) for WHOM (REC), from WHOM/WHAT (ORIG)...]: Moldavija sada ima novog predsjednika.

Summary and Conclusions

- the data obtained from the experimental automatic semantic role labeling based on supervised machine learning methods
- the preliminary quantitative analyses of Slovene and Croatian training corpora (in terms of verbs range and frequencies, semantic roles, and typical syntactic-semantic patterns for the most frequent verbs)
- The data for both languages are quite similar from all the above perspectives, despite the differences in corpora design.

Summary and Conclusions

- the SRL framework that was being developed within this bilateral project is suitable for semantic role labeling tasks in both languages.
- the framework has been successfully implemented to serve as the solid base for the automatic SRL (using supervised machine learning methods).
- a common framework
  - saving time and resources
  - mutual evaluation and corrections

Future developments

- Building a corpus with SRL annotations is an ongoing work and both corpora will be upgraded in the future.
- increase in size, calculation of inter-annotator agreement and segmentation of patterns according verb senses (when compatible semantic resources for both languages are available).

A big ‘Thank you’ to our annotators

- Lucija Jezeršek
- Taja Kuzman
- Dafne Marko
- Ivan Pandžić
- Iza Škrjanec
- Anja Zajc

Hvala!

- apolonija.gantar@guest.arnes.si
- kdespot@ihjj.hr
- krek@ijs.si
- nikola.ljubesic@ijs.si

- From both corpora, we have extracted stable syntactic-semantic patterns characteristic for each individual verb. Those patterns are similar in both languages despite the differences in the corpus design. To make the formalizations of these patterns more readable, we use "**Who did What to Whom, and How, When and Where**?" form (ACT = Who, PAT = What, RESLT=Who/What, LOC = Where etc.).