Human Language Technologies – The Baltic Perspective
A. Utka et al. (Eds.)

doi:10.3233/978-1-61499-442-8-167

Adding Constructicon to the Latvian FrameNet

Didzis GOSKO[a], Guntis BARZDINS[a]
[a]*Institute of Mathematics and Computer Science, University of Latvia*

Abstract. The paper describes an extension to the practical natural language processing and information extraction system implemented for a national news agency in Latvia. The constructicon extension introduces pattern based rules capturing the structure of frame annotations rather than identifying the frame constituents in an isolated manner. Reported are also results for English where this approach improves the accuracy of the FrameNet automatic parsing compared to the current state-of-the-art.

Keywords. FrameNet, Classification algorithms, Information extraction, Latvian

Introduction

In this paper we describe a further extension to Latvian FrameNet – the so called “Constructicon” [1], the concept of which was introduced by the FrameNet [2] inventor Charles J. Fillmore and further extended by the Construction Grammar community. Assumption of Constructicon is that language consists to quite a large extent of restricted, semi-productive constructions, which are highly problematic for regular rule-based language technology. The idea of building a Constructicon is to mine the same FrameNet annotated corpora for the idiomatic constructions (phrases) commonly used to express specific frames in Latvian.

The C6.0 rule-based binary classification algorithm[1] has been successfully used for creating a Latvian FrameNet parser achieving the state-of-the-art accuracy on par with the best English FrameNet parsers [3]. In this paper we report two extensions to the C6.0 algorithm: support for special kind of multi-class classification, and support for function features. Together these two extensions enable limited support for Constructicon for identification of the frame-bearing constructions rather than pure predicate-argument structures annotated in FrameNet so far.

We also briefly describe a practical application of the frame-semantic parsing system for structured information extraction (IE) from unstructured newswire texts, which is currently being implemented in a national news agency [4]. Since FrameNet itself does not define any Knowledge Representation (KR) paradigm, for the needs of this IE system FrameNet is combined with Named Entity Linking (NEL) to create the actual KR framework. The system is implemented and populated with data from more than 1 million newswire archive articles.

The paper is organized as follows: in Section 1 the Constructicon-enabling extension to the C6.0 rule-based classification is described. Section 2 describes the actual C6.0 FrameNet based information extraction and reports accuracy results for both Latvian and English. The paper concludes with the discussion on the frame

[1] Available at http://c60.ailab.lv

annotation quality, which has major impact on the overall FrameNet based information extraction system performance.

1. Constructicon Enabling Extensions to the C6.0 Classification Algorithm

The C6.0 rule-based binary classification algorithm [3] (C6.0 is a modification of the popular C4.5 algorithm [12]) has been successfully applied to creating a Latvian FrameNet parser achieving a state-of-the-art accuracy on par with the best English FrameNet parsers. Here we describe two extensions to the C6.0 algorithm: support for the special kind of multi-class classification, and support for function features.

Given *k* training examples of the form:

$(a_{11}, a_{12}, a_{13}, \ldots a_{1n}, \text{class}_1)$
$(a_{21}, a_{22}, a_{23}, \ldots a_{2n}, \text{class}_2)$
…
$(a_{k1}, a_{k2}, a_{k3}, \ldots a_{kn}, \text{class}_k)$

(where a_{ij} is an arbitrary character string (e.g. “chikmagalur”) and $class_i$ is only one of two strings “YES” or “NO”) a C6.0 classification algorithm uses exhaustive search to build a set of rules for identifying the YES-class examples in the training:

$\text{if}\,(x_1, x_2, x_3, \ldots x_n)\text{ then (class=”YES”)}$

where any of the positions x_i contains an arbitrary character string or an unspecified value denoted “_”. The rule-set built by C6.0 has the highest possible Laplace ratio $(tp+1)/(tp+fp+2)$ for each rule’s accuracy estimation, where *tp* is the number of true positives identified by the rule and *fp* is the number of false positives identified by the rule. High Laplace ratio is shown to ensure that such rules are likely to correctly identify YES-class samples also in the unseen data. The exhaustive search complexity of C6.0 is

$$k \times number\_of\_YES\_exemplars \times 2^n$$

which is a tractable number up to approximately n=20 (n is the number of features present in each example).

Although C6.0 has allowed creating state-of-the-art FrameNet parsers for both Latvian [1] and for English [6], in this paper we propose two extensions to the above-mentioned C6.0 algorithm.

The first extension nicknamed C6.0-C (for “Constructicon”) removes the limitation of the binary classification. Now we will assume that the *k* training examples are in the form:

$(a_{11}, a_{12}, a_{13}, \ldots a_{1n}, \text{label}_{11}, \text{label}_{12}, \ldots \text{label}_{1m})$
$(a_{21}, a_{22}, a_{23}, \ldots a_{2n}, \text{label}_{21}, \text{label}_{22}, \ldots \text{label}_{2m})$
…
$(a_{k1}, a_{k2}, a_{k3}, \ldots a_{kn}, \text{label}_{k1}, \text{label}_{k2}, \ldots \text{label}_{km})$

(where a_{ij} and $label_{ij}$ are arbitrary character strings) and the classification task of the extended C6.3 algorithm is to find the set of the highest Laplace scoring production rules in the following form

$$\text{if } (x_1, x_2, x_3, \dots x_n) \text{ then } (\text{label}_1, \text{label}_2, \dots \text{label}_m)$$

where any of the positions x_i and $label_i$ contains an arbitrary character string or an unspecified value denoted "_". The meaning of these rules is that they deduce right-hand-side label values from the left-hand-side parameters. It is important to note that this distinction between left and right hand side values is irrelevant during exhaustive search for the highest scoring rules, as both left and right hand side values are given in the training examples. The complexity of exhaustive search in the C6.0-C algorithm is

$$k^2 \times 2^{(n+m)}$$

Although C6.0-C classification algorithm has slightly higher complexity than C6.0, it enables constructicon based approach to FrameNet parsing.

The second extension nicknamed C6.0-F (for "Function-features") differs from C6.0 in that rules are not merely relaxed patterns of feature values in the positive training examples, but rather contain arbitrary functions with Boolean range:

$$\text{if } (f_1(x_1,y_1), f_2(x_2,y_2), f_3(x_3,y_3), \dots f_n(x_n,y_n)) \text{ then } (\text{class=''YES''})$$

where f_i is a predefined function, x_i is an arbitrary character string, and y_i is a placeholder for the feature value in the specific example to which this rule is applied; any position may contain instead an unspecified value denoted "_".

This allows employing more complex features as well as often reduces the total number of features required (and thus speeding up the exhaustive search).

As an example of a complex feature consider a function *substr(x,y),* which is *true* if the string *y* contains a substring *x*, and *false* otherwise.

If all the functions would be *equal(x,y),* then C6.0-F algorithm behaves exactly as C6.0 algorithm. Complexity of C6.0-F algorithm is the same as of C6.0, provided that the functions f_i and x_i value selection mechanisms are simple.

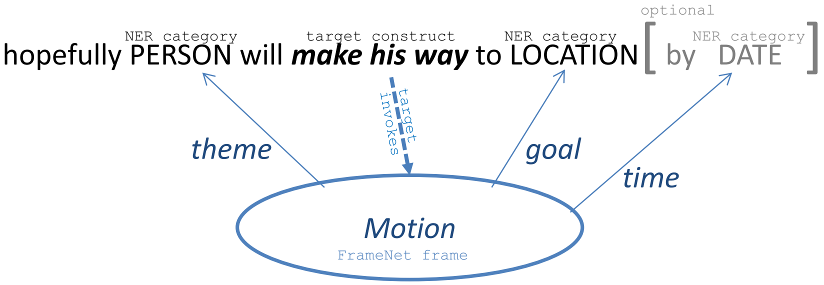


Figure 1. Example of an idiomatic construction in the Constructicon invoking a *Motion* frame.

Combining these two extensions into C6.0-CF algorithm (function-features only in the left-hand-side of the rule) enables efficient Constructicon like learning of frame-bearing construction patterns (as illustrated in Figure 1) capturing the structure of frame annotations rather than identifying the frame constituents in an isolated manner.

2. FrameNet Based Information Extraction

Latvian FrameNet originally was created for a practical information extraction system developed for a national news agency to automatically extract biographical data about publicly visible persons and organizations mentioned in the national archive of newswire articles [4]. Few design decisions helped to simplify the creation of Latvian FrameNet. The first design decision was to use for Latvian FrameNet only 26 English FrameNet frames (see Figure 2) which were of interest to the national news agency for the media monitoring purposes. The second design decision for Latvian FrameNet was to pre-process all input texts with a Latvian NLP pipeline to produce extended CoNLL-style annotations prior to any FrameNet annotation. Based on this approach, from various types of newswire sources a FrameNet annotated corpus for Latvian was created (see Table 1 for comparison with English FrameNet annotated corpus).

Table 1. FrameNet data sets used for evaluation.

	Latvian FrameNet data	*English SemEval '07 data*
Exemplar sentences	4,079	139,439
Frame types	26	665
Frame Element types	80	720
Sentences in test data	844	120

This FrameNet annotated corpus was used to create an automatic Latvian frame-semantic parser reaching the state-of-the-art English frame-semantic parser [7, 8] accuracy (see Table 2) thanks to novel extensions to the C6.0 decision-tree classifier algorithm.

Table 2. Evaluation results for frame target and frame element identification.

	Target identification			***FE identification***		
	Precision	*Recall*	*F1*	*Precision*	*Recall*	*F1*
LTH (English data)	66.2%	50.6%	57.3%	51.6%	35.4%	42.0%
SEMAFOR (English data)	69.7%	54.9%	61.4%	58.1%	38.8%	46.5%
C6.0 RuleSet (English data)	**77.1%**	**53.7%**	**63.3%**	**47.3%**	**47.0%**	**47.1%**
C6.0 RuleSet (Latvian data)	**63.5%**	**62.7%**	**63.1%**	**65.9%**	**76.8%**	**70.9%**

FrameNet itself does not define any Knowledge Representation (KR) paradigm – it is merely a lexicographic annotation framework. FrameNet needs to be combined with Named Entity Linking (NEL) [10] to create a usable KR framework like an ontology shown in Figure 2.

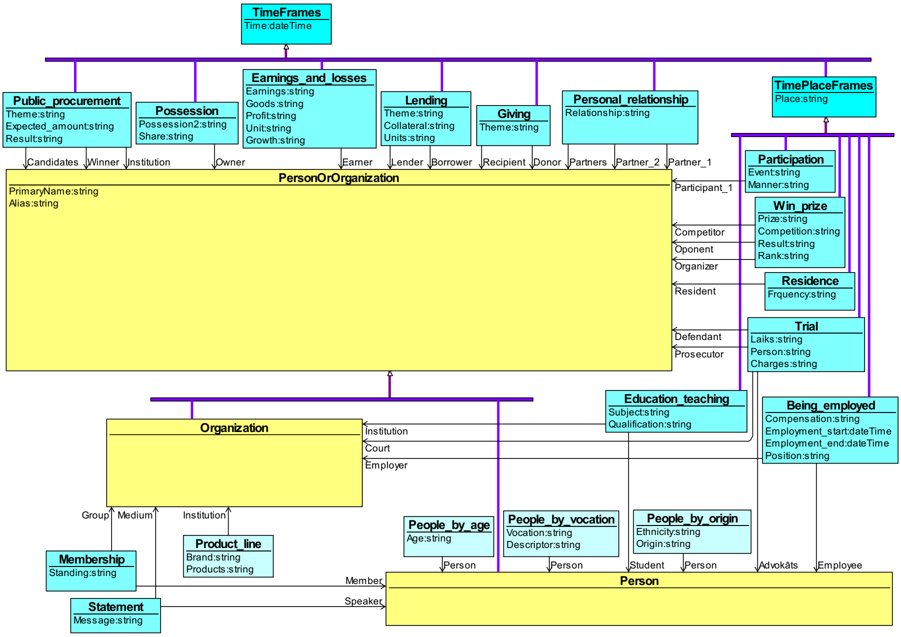


Figure 2. OWLGrEd visualization of 26 Latvian FrameNet Frames and core NEL categories.

It should be noted that the KR framework in Figure 2 does not define any constraints (such as cardinality constraints, e.g., "a person can have only one mother"). This means that an additional conversion and constraint-checking step is necessary to use the data from the KR framework in Figure 2 into similar, but more traditionally organized database such as DBpedia 3.9 ontology [9] (see a fragment in Figure 3).

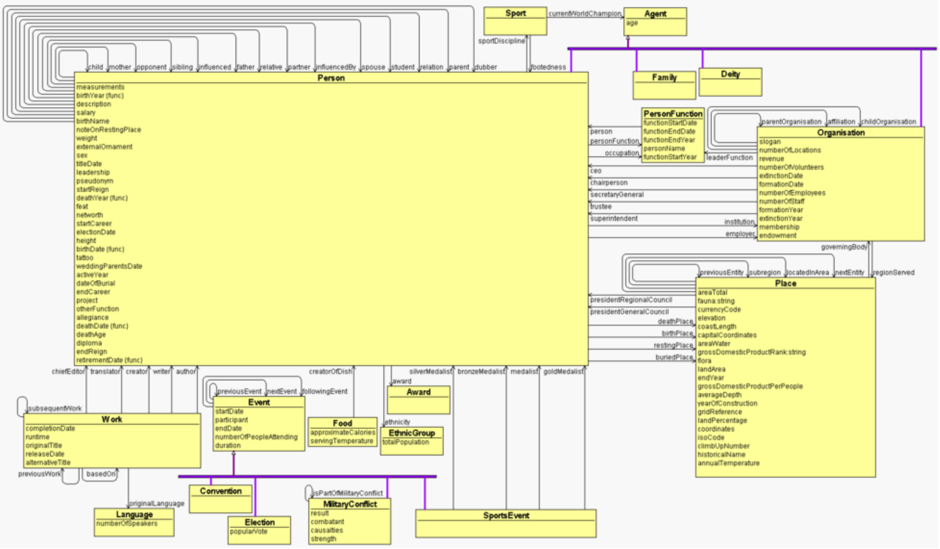
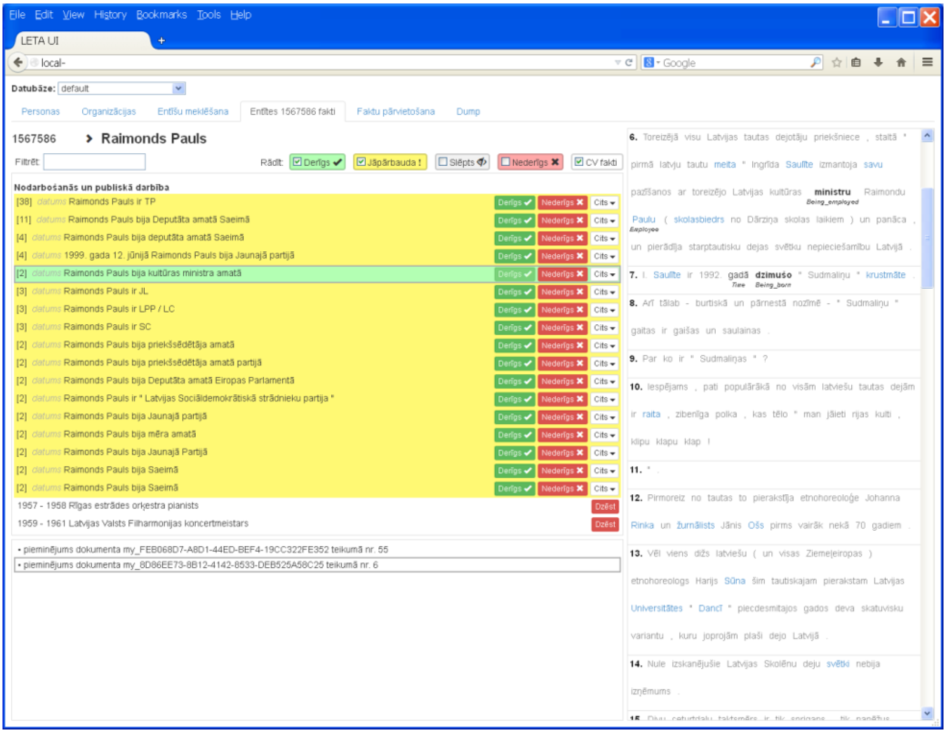


Figure 3. OWLGrEd visualisation of a fragment of DBpedia 3.9 ontology. It illustrates how Person and Organization classes are linked by binary relations according to traditional database structure.

The actual information extraction stands for populating the KR ontology with instance data retrieved from the source text. To this goal, frame-semantic parser (producing instances for the dark boxes in Figure 2) is combined with Named Entity Linking (NEL) techniques to automatically determine which mentions in the text refer to the same real-world entity (instances for the NEL categories in Figure 2).



1567586 › Raimonds Pauls

Nodarbošanās un publiskā darbība

[38]	datums	Raimonds Pauls ir TP	Derīgs ✓	Nederīgs ✗	Cits ▾
[11]	datums	Raimonds Pauls bija Deputāta amatā Saeimā	Derīgs ✓	Nederīgs ✗	Cits ▾
[4]	datums	Raimonds Pauls bija deputāta amatā Saeimā	Derīgs ✓	Nederīgs ✗	Cits ▾
[4]	datums	1999. gada 12. jūnijā Raimonds Pauls bija Jaunajā partijā	Derīgs ✓	Nederīgs ✗	Cits ▾
[2]	datums	Raimonds Pauls bija kultūras ministra amatā	Derīgs ✓	Nederīgs ✗	Cits ▾
[3]	datums	Raimonds Pauls ir JL	Derīgs ✓	Nederīgs ✗	Cits ▾
[3]	datums	Raimonds Pauls ir LPP / LC	Derīgs ✓	Nederīgs ✗	Cits ▾
[3]	datums	Raimonds Pauls ir SC	Derīgs ✓	Nederīgs ✗	Cits ▾
[2]	datums	Raimonds Pauls bija priekšsēdētāja amatā	Derīgs ✓	Nederīgs ✗	Cits ▾
[2]	datums	Raimonds Pauls bija priekšsēdētāja amatā partijā	Derīgs ✓	Nederīgs ✗	Cits ▾
[2]	datums	Raimonds Pauls bija Deputāta amatā Eiropas Parlamentā	Derīgs ✓	Nederīgs ✗	Cits ▾
[2]	datums	Raimonds Pauls ir " Latvijas Sociāldemokrātiskā strādnieku partija "	Derīgs ✓	Nederīgs ✗	Cits ▾
[2]	datums	Raimonds Pauls bija Jaunajā partijā	Derīgs ✓	Nederīgs ✗	Cits ▾
[2]	datums	Raimonds Pauls bija mēra amatā	Derīgs ✓	Nederīgs ✗	Cits ▾
[2]	datums	Raimonds Pauls bija Jaunajā Partijā	Derīgs ✓	Nederīgs ✗	Cits ▾
[2]	datums	Raimonds Pauls bija Saeimā	Derīgs ✓	Nederīgs ✗	Cits ▾
[2]	datums	Raimonds Pauls bija Saeimā	Derīgs ✓	Nederīgs ✗	Cits ▾
		1957 - 1958 Rīgas estrādes orķestra pianists			Dzēst
		1959 - 1961 Latvijas Valsts Filharmonijas koncertmeistars			Dzēst

- pieminējums dokumenta my_FEB068D7-A8D1-44ED-BEF4-19CC322FE352 teikumā nr. 55
- pieminējums dokumenta my_8D86EE73-8B12-4142-8533-DEB525A58C25 teikumā nr. 6

6. Toreizējā visu Latvijas tautas dejotāju priekšniece , staltā " pirmā latvju tautu meita " Ingrīda Saulīte izmantoja savu pazīšanos ar toreizējo Latvijas kultūras ministru Raimondu Paulu (skolasbiedrs no Dārziņa skolas laikiem) un panāca , un pierādīja starptautisku dejas svētku nepieciešamību Latvijā .

7. I. Saulīte ir 1992. gadā dzimušo " Sudmaliņu " krustmāte .

8. Arī tālab - burtiskā un pārnestā nozīmē - " Sudmaliņu " gaitas ir gaišas un saulainas .

9. Par ko ir " Sudmaliņas " ?

10. Iespējams , pati populārākā no visām latviešu tautas dejām ir raita , zibenīga polka , kas tēlo " man jāieti rījas kulti , klipu klapu klap !

11. " .

12. Pirmoreiz no tautas to pierakstīja etnohoreoloģe Johanna Rinka un žurnālists Jānis Ošs pirms vairāk nekā 70 gadiem .

13. Vēl viens dižs latviešu (un visas Ziemeļeiropas) etnohoreologs Harijs Sūna šim tautiskajam pierakstam Latvijas Universitātes " DancI " piecdesmitajos gados deva skatuvisku variantu , kuru joprojām plaši dejo Latvijā .

14. Nule izskanējušie Latvijas Skolēnu deju svētki nebija izņēmums .

Figure 4. User interface for viewing and correcting the automatically extracted information.

We have implemented this integrated information extraction system and populated it with data from more than 1 million newswire articles. Figure 4 shows the automatic information extraction system user interface, where instance data from the KR database in Figure 2 is verbalized using a light version of [5] producing simple sentences shown in the left part of Figure 4 along with the found duplicate counts indicate the confidence level. Although the accuracy of the implemented system is insufficient for autonomous use, it provides an important assistance to the human data curators extracting this kind of information from the public news sources.

3. Discussion on the Frame Quality

Our comparative study of large English and smaller Latvian FrameNets highlights the fact that not all 665 frames in English FrameNet version 1.3 or 877 frames in English FrameNet version 1.5 are of equal quality: some of the frames have clearly cut meaning and sufficient training examples, while others are vague of with too few examples (see Table 3). Therefore, the average accuracy evaluation in the previous section is somewhat misleading about the actual accuracy of the described methods for well-defined frames.

Table 3. Target identification F1 scores for some FrameNet frames.

Being born	100	Residence	67	Participation	40
Earnings and losses	89	Statement	67	Employment end	33
Death	80	Hiring	62	Product line	33
Education teaching	71	Membership	50	Lending	29
Being employed	70	Possession	48	Personal relationship	25
Change of leadership	67	People by vocation	46	Trial	18
Intentionally create	67	Win prize	45	People by origin	16

For information extraction tasks it makes sense to hand-pick only a high-quality subset of all FrameNet frames, like the ones in the left column in Table 3, or to create additional training resources for the low-scoring frames.

An alternative approach to boosting semantic parsing accuracy might be to aim for the complete semantic parsing of the entire sentence according to Abstract Meaning Representation (AMR) [11] framework, as it uses predefined Named Entity types and explicit identifiers for coreference and missing argument identification to address core reasons for often low FrameNet annotation accuracies.

4. Conclusion

The described approach illustrates the possibility of bootstrapping a competitive frame-semantic parser and Constructicon for a new language by merely hand-annotating at least 1000 sentences with the FrameNet frames of interest. Having an accurate frame-semantic parser enables creation of a practically useful information extraction system.

5. Acknowledgment

This research was partially supported by the Project Nr.2DP/2.1.1.1.0/13/APIA/VIAA/014 (ERAF) “Identification of relations in newswire texts and graph visualization of the extracted relation database” under contract Nr. 1/5-2013, LU MII Nr. 3-27.3-5-2013.

References

[1] Fillmore, Charles J., Russell Lee-Goldman & Russell Rhomieux. The FrameNet Constructicon. In Hans C. Boas & Ivan A. Sag (eds.), Sign-based Construction Grammar, pp. 309-372. Stanford: CSLI (2012)

[2] Fillmore, C.J., Johnson, C.R., Petruck, M.R.L.: Background to FrameNet. International Journal of Lexicography, 16, pp. 235-250 (2003)

[3] Barzdins, G., Gosko, D., Rituma, L., Paikens, P.: Using C5.0 and Exhaustive Search for Boosting Frame-Semantic Parsing Accuracy. In: Proceedings of the 9th Language Resources and Evaluation Conference (LREC), pp. 4476-4482. Reykjavik (2014)

[4] Paikens, P. Latvian newswire information extraction system and entity knowledge base. In: Baltic HLT-2014, Frontiers in Artificial Intelligence and Applications, IOS Press, (2014, this volume)

[5] Dannells, D., Gruzitis, N. Extracting a bilingual semantic grammar from FrameNet - annotated corpora. In: Proceedings of the 9th Language Resources and Evaluation Conference (LREC), pp. 2466-2473. Reykjavik (2014)

[6] Barzdins G.: FrameNet CNL: a Knowledge Representation and Information Extraction Language. In: CNL 2014 Workshop, LNCS/LNAI 8625, pp. 90-101. Springer, Heidelberg (2014)

[7] Johansson, R., Nugues, P.: LTH: semantic structure extraction using nonprojective dependency trees. In: Proceedings of SemEval-2007: 4th International Workshop on Semantic Evaluations. Prague, pp. 227-230 (2007)

[8] Das, D., Chen, D., Martins, A.F.T, Schneider, N., Smith, N.A.: Frame-Semantic Parsing, Computational Linguistics, 40(1), pp. 9-56. (2014)

[9] Daiber, J., Jakob, M., Hokamp, C., Mendes, P.N.: Improving efficiency and accuracy in multilingual entity extraction, In: Proceedings of the 9th International Conference on Semantic Systems, pp. 121-124. ACM (2013)

[10] Wick, M., Singh, S., Pandya, H., McCallum, A.: A Joint Model for Discovering and Linking Entities. In: Proceedings of the 2013 workshop on Automated knowledge base construction, pp. 67-72. ACM (2013)

[11] Banarescu, L., Bonial, C., Cai, S., Georgescu, M., Griffitt, K., Hermjakob, U., Knight, K., Koehn, P., Palmer, M., and Schneider, N.: Abstract Meaning Representation for Sembanking. In: Proc. Linguistic Annotation Workshop (2013)

[12] Quinlan, J.R.: C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers (1993)