ULSAna: Universal Language Semantic Analyzer

Ondřej Pražák
NTIS
Faculty of Applied Sciences
University of West Bohemia
Technická 8, 306 14 Plzeň
ondfa@ntis.zcu.cz

Miloslav Konopík
Dpt. of Computer Science and Engineering
Faculty of Applied Sciences
University of West Bohemia
Technická 8, 306 14 Plzeň
konopik@kiv.zcu.cz

Abstract

We present a live cross-lingual system capable of producing shallow semantic annotations of natural language sentences for 51 languages at this time. The domain of the input sentences is in principle unconstrained. The system uses single training data (in English) for all the languages. The resulting semantic annotations are therefore consistent across different languages. We use CoNLL Semantic Role Labeling training data and Universal dependencies as the basis for the system. The system is publicly available and supports processing data in batches; therefore, it can be easily used by the community for research tasks.

1 Introduction

In this work, we present a major outcome in our journey to build a system capable of producing semantic annotations for domain and language unconstrained natural language sentences. Currently, we rely on the Semantic Role Labeling (SRL) annotation scheme (Gildea and Jurafsky, 2002). The SRL goal is to determine semantic relationships (*semantic roles*) of given *predicates* (see examples in Figure 1). Verbs, such as "believe" or "cook", are natural predicates but certain nouns can be accepted as predicates as well (see the third line in the example). The semantic roles are specific for each predicate; however, the meaning of the roles is mostly shared across predicates. The core roles are denoted by *A0* (usually Agent), *A1* (usually Patient) and *A2*. Additional roles are modifier arguments (*AM-**), restriction arguments (*R-**) and others. We selected SRL because we believe that the annotations are simple enough to be generalized for different languages and target domains but at the same time expressive enough to bring a useful insight into the sentence semantics.

(1) [He]$_{A0}$ believes [in what he plays] $_{A1}$.

(2) Can [you] $_{A0}$ cook [the dinner] $_{A1}$?

(3) [The nation's] $_{AM\text{-}LOC}$ largest [pension]$_{A1}$ fund,

Figure 1: Three examples of shallow semantic annotations: 1) and 2) are examples of verb predicates and 3) of a noun predicate.

In order to be able to produce semantic annotations for more languages, we employ the *cross-lingual* SRL. Cross-lingual SRL takes the training data from a source language (usually English) and builds a language independent model that can be applied to target languages. The advantage of the cross-lingual SRL is that it ensures coherent annotation for all supported languages because it trains on single training data for all the languages. This does not apply to the monolingual SRL where the tagsets and annotation guidelines change with every training dataset.

In our approach, we heavily depend on Universal Dependencies (UD) (Nivre et al., 2016). They are the primary means to transfer the learned rules from one language to another. With universal dependencies, we can create language independent parse trees. It means that sentences with the same syntactic structure share (in theory) the same parse trees for all (supported) languages. We train our machine learning model on the UD trees to capture the syntactic patterns required for semantic role labeling. We do not use any lexical information or any other language dependent features. Our only information for SRL comes from UD trees. Thus, the resulting model can be applied to any of the supported languages.

The system we present in this paper is a web-based application written in Java – see the screen-

Proceedings of Recent Advances in Natural Language Processing, pages 967–972,
Varna, Bulgaria, Sep 2–4, 2019.
https://doi.org/10.26615/978-954-452-056-4_112

shots in Figures 3 and 2. The system allows a user to input a natural language sentence in any of the 51 languages. The system outputs SRL annotations (predicates and corresponding semantic roles) of the input sentences. The semantic roles are associated with syntactic tree nodes. The video demonstration of the system is available here: https://www.youtube.com/watch?v=8QPKCegHT_c. The system itself can be accessed at the following address: http://liks.fav.zcu.cz/ulsana. We intend to support the system for public use for several years.

2 Related Work

Approaches to cross-lingual SRL can be divided into three main categories: **1)** *Annotation projection* methods attempt to transfer annotations from one language to another and then they train an SRL system on the transferred annotations. **2)** *Model transfer* approaches are designed to use language-independent features to train a universal model which can be applied to languages that support the designed features. **3)** Methods based on *unsupervised training* require no annotated data; however, the models have difficulties in assigning meaningful labels to predicate arguments.

Annotation projection Padó and Lapata (2009) transfer annotations to a target language via word alignments obtained from parallel corpora. Annesi and Basili (2010) use a similar approach and extend it with an HMM model to increase the transfer accuracy.

Model transfer Kozhevnikov and Titov (2013) use cross-lingual word mappings and cross-lingual semantic clusters obtained from parallel corpora, and cross-lingual features extracted from unlabelled syntactic dependencies to create a cross-lingual SRL system. In (Kozhevnikov and Titov, 2014), they try to find a mapping between language-specific models using parallel data automatically.

Unsupervised SRL Grenager and Manning (2006) deploy unsupervised learning using the EM algorithm based upon a structured probabilistic model of the domain. Lang and Lapata (2011) discover arguments of verb predicates with high accuracy using a small set of rules. A split-merge clustering is consequently applied to assign (nameless) roles to the discovered arguments. Titov and Klementiev (2012) propose a superior argument clustering by using the Chinese restaurant process.

Our approach belongs among the model transfer approaches. Most of the other state-of-the-art approaches to SRL rely on lexical features (e.g. word lemmas). In the cross-language scenario, such features require bilingual models (e.g. word mapping via machine translation or bilingual clusters). In our demonstration application, we show a multi-language model that is capable of producing annotations for many languages. Therefore, we omit all bilingual features including the lexical features from our model.

3 System Description

In this section, we describe the core of the cross-lingual system we use in our demo. The system is described in our original paper (Pražák and Konopik, 2017) in full details. Here, we explain only the basic principles. The system described in (Pražák and Konopik, 2017) is available for five languages only. In this demo, we extended the system for 51 languages.

3.1 Training Dataset and Annotation Conversion

We train our system on UD parse trees. However, there are no such training data that would contain SRL annotations on UD trees. Therefore, we proposed an algorithm to convert existing SRL annotations built on SD[1] trees.

The conversion process is by no means straightforward. The main source of complications stems from different approaches to choose head words for syntactic phrases in UD trees. To solve this issue, we proposed optimization algorithms that attempt to select the most appropriate heads for UD trees which would cover the same phrases as the heads in original SD trees. In many cases, there is no such word in UD trees which could be used as the new head. In such cases, we choose the head that minimizes the annotation error. The details of the proposed conversion algorithms are presented in the original paper – Section 4.

In our application, we use the CoNLL 2009 English dataset (Hajič et al., 2009). The corpus

[1]SD stands for Standard or language-Specific Dependencies, e.g. Stanford dependencies – urlhttps://nlp.stanford.edu/software/stanford-dependencies.shtml.

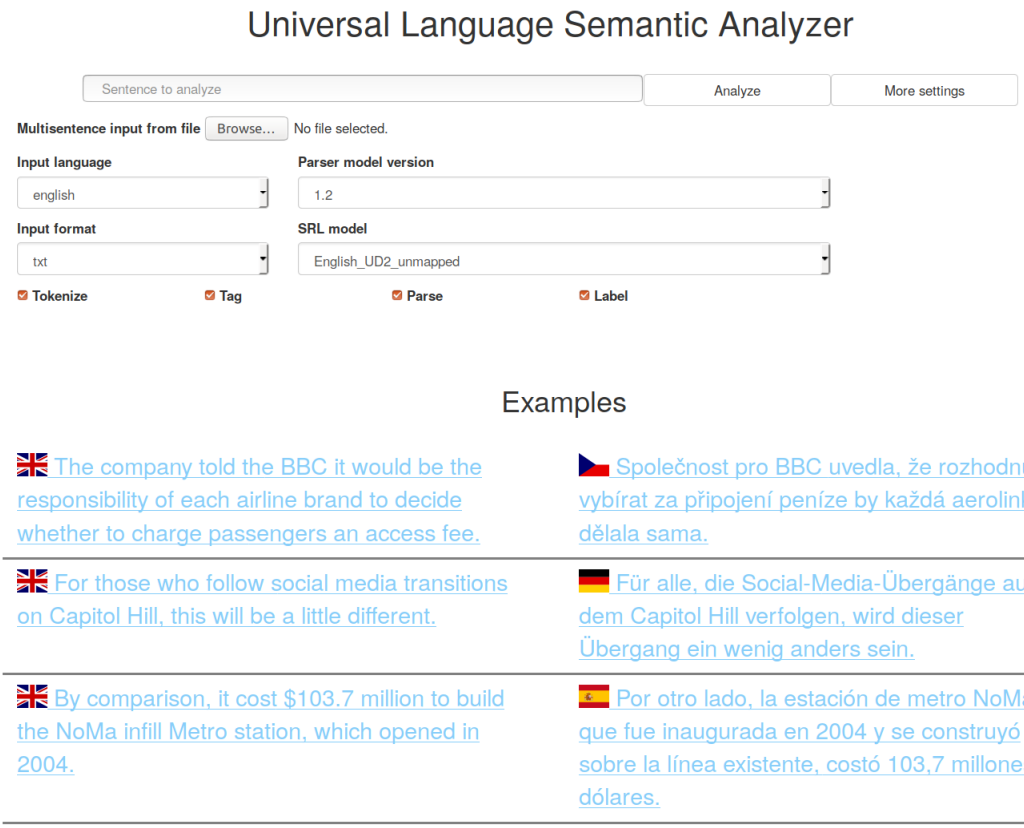

Figure 2: Application screenshot

includes syntactic dependencies (from the Penn Treebank [TB]) and semantic dependencies (from PropBank [PB] and NomBank [NB]).

3.2 Universal Dependencies Parser

Our system requires syntactic trees in the UD annotation scheme. We rely on the freely available tool UDPipe (Straka et al., 2016). It contains pretrained models for all the languages we support in our application. We use models provided with parsers based on UD. The algorithms were developed on UD v1.2 models based on UD 2.0 were also added, and they achieve better results (about +1% of labeling accuracy).

3.3 Classifier & Features

We train a supervised system based upon the Maximum Entropy classifier using the Brainy tool (Konkol, 2014). We use separate models for verb and non-verb predicates.

All features employed in our system are syntactic:

- *Predicate-argument distance* – the distance between the locations of a predicate and an argument in a sentence.
- *POS* – part-of-speech of the predicate, the argument and their parent nodes.
- *Dependency relation* – dependency tree relation of the predicate, the argument and their parent nodes.
- *Directed path* – dependency tree path from the predicate to the argument including the indication of the dependency directions.
- *Undirected path* – the list of relations from the predicate to the argument.
- *Verb voice* – indication of active/passive voice.
- *Other syntactic features* – `feats` column in CoNLL 2009 format with additional information about the words.
- *Bigram features* – predicate-argument bigrams of the part-of-speech and the dependency relations.

The dependency path features are encoded as a probability of a word being a predicate argument (or having a specific role) given the path. These features are more general, and the resulting vectors have a smaller dimension. Also, the cost function is smoother, and thus the model is easier to train.

3.4 Web Application Description

We created a Java web UI for the SRL annotation and its visualization. We use *TikZ* for visualization of the trees. The *TikZ* output is converted to *SVG* which is then shown in the browser. The application takes its input either in plain text or in various CoNLL formats. Input can be a single sentence or a file with sentences separated by new lines. On the input, a user has to select an input language (one of 51 supported languages) because syntactic parsers are language-dependent and the application cannot determine the language automatically at this time. The application can parse the sentences syntactically and semantically. After these steps (if requested) the annotations are visualized in the SVG format and showed in the browser. The user can download analyzed sentences in *svg*, *pdf* or raw *CoNLLu* format.

3.5 Application Use Cases

Research Experiments The primary purpose of our application is to help the researchers to get familiar with the capabilities of cross-lingual semantic processing. We also want to demonstrate the power and limitations of Universal Dependencies. Users can work with examples entered into the input field, but they can also use the batch processing feature. In this way, users can obtain SRL annotations of larger corpora that can be used in the consequent research.

Language Learning The application can also help users who are learning a new language. Our application shows the structure of a sentence and the basic roles of the main phrases in the sentence. In this way, users can more easily understand the semantic structure of the sentence.

Translation Cross-lingual SRL can be used either in the machine or human translation. When translating a sentence, we aim to preserve the semantic structure of the sentence. We can achieve that by studying both structures of the source and target input sentences.

3.6 Known issues

- Parser errors – Since our system relies solely on syntactic features, it is very sensitive to parser errors. When the sentences match the domain of training data of UD annotations (mostly news domain) the parse trees are generally quite correct. We produce mostly correct SRL annotations with correct parser trees. However, our system is usually unable to classify correctly when the parse trees contain significant errors.

- Visualizing complex relations – Our system sometimes struggles with visualizing complex relations. It those circumstances the resulting visualization can be confusing.

4 Future Work

In future work, we plan to adopt the end-2-end approach to Semantic Role Labeling. We intend to attach the SRL annotation after UD parsing and use a global cost function to optimize the UD parsing and the SRL annotation simultaneously. In order to apply the end-2-end approach, we might have to switch to the SyntaxNet[2] UD parser. We expect to be able to produce more robust SRL annotations with one global optimization function.

Next, we plan to focus on lexical features. We want to stay with the idea of one model for many languages. Therefore, we need to use cross-lingual embeddings as lexical features.

5 Conclusion

We have created a semantic role labeling system with a massively multilingual model. A single model can be used for SRL in 51 different languages. The system supports large inputs, and therefore it can be used to annotate entire datasets for various NLP tasks.

Acknowledgments

This publication was supported by the project LO1506 of the Czech Ministry of Education, Youth and Sports and by Grant No. SGS-2019-018 Processing of heterogeneous data and its specialized applications. Computational resources were provided by the CESNET LM2015042 and

[2] https://opensource.google.com/projects/syntaxnet

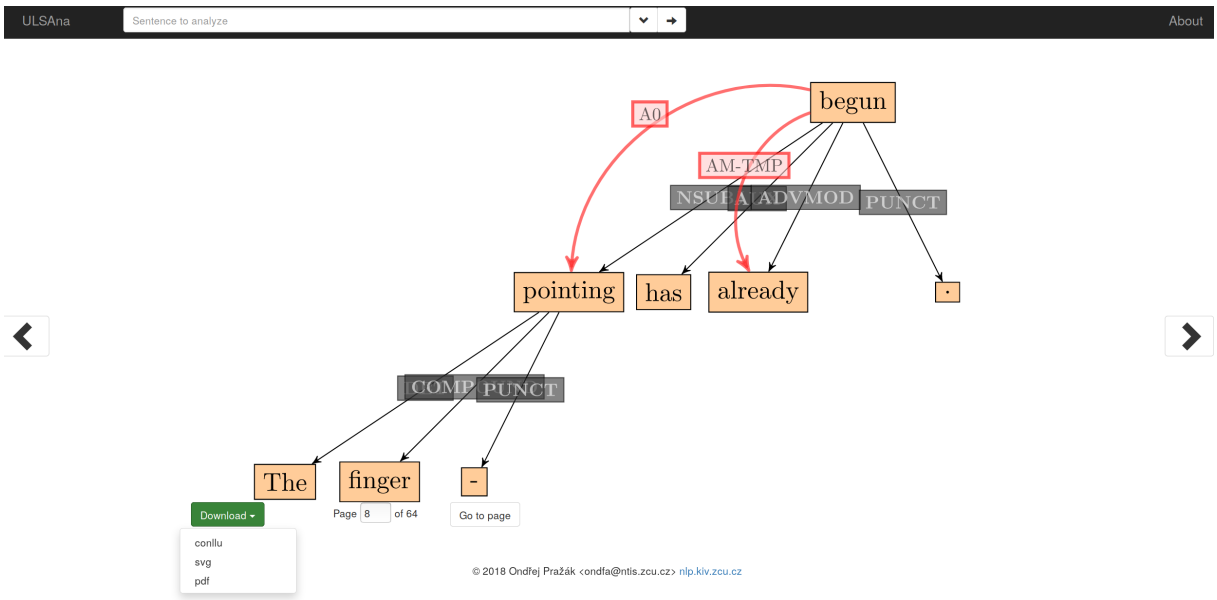

Figure 3: Application screenshot – sentence visualization example

the CERIT Scientific Cloud LM2015085, provided under the programme "Projects of Large Research, Development, and Innovations Infrastructures".

References

Paolo Annesi and Roberto Basili. 2010. Cross-lingual alignment of FrameNet annotations through hidden markov models. In *Proceedings of the 11th International Conference on Computational Linguistics and Intelligent Text Processing*, CICLing'10, pages 12–25, Berlin, Heidelberg. Springer-Verlag.

Daniel Gildea and Daniel Jurafsky. 2002. Automatic labeling of semantic roles. *Computational linguistics*, 28(3):245–288.

Trond Grenager and Christopher D. Manning. 2006. Unsupervised discovery of a statistical verb lexicon. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, EMNLP '06, pages 1–8, Stroudsburg, PA, USA. Association for Computational Linguistics.

Jan Hajič, Massimiliano Ciaramita, Richard Johansson, Daisuke Kawahara, Maria Antònia Martí, Lluís Màrquez, Adam Meyers, Joakim Nivre, Sebastian Padó, Jan Štěpánek, et al. 2009. The conll-2009 shared task: Syntactic and semantic dependencies in multiple languages. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning: Shared Task*, pages 1–18. Association for Computational Linguistics.

Michal Konkol. 2014. Brainy: A machine learning library. In *Artificial Intelligence and Soft Computing*, volume 8468 of *Lecture Notes in Computer Science*, pages 490–499. Springer International Publishing.

Mikhail Kozhevnikov and Ivan Titov. 2013. Cross-lingual transfer of semantic role labeling models. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, ACL 2013, 4-9 August 2013, Sofia, Bulgaria, Volume 1: Long Papers*, pages 1190–1200.

Mikhail Kozhevnikov and Ivan Titov. 2014. Cross-lingual model transfer using feature representation projection. In *ACL (2)*, pages 579–585.

Joel Lang and Mirella Lapata. 2011. Unsupervised semantic role induction via split-merge clustering. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1117–1126, Portland, Oregon, USA. Association for Computational Linguistics.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, Christopher D Manning, Ryan T McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, et al. 2016. Universal dependencies v1: A multilingual treebank collection. In *LREC*.

Sebastian Padó and Mirella Lapata. 2009. Cross-lingual annotation projection for semantic roles. *Journal of Artificial Intelligence Research*, 36:307–340.

Ondřej Pražák and Miloslav Konopik. 2017. Cross-lingual srl based upon universal dependencies. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 592–600, Varna, Bulgaria. INCOMA Ltd.

Milan Straka, Jan Hajič, and Jana Straková. 2016. UDPipe: trainable pipeline for processing CoNLL-U files performing tokenization, morphological analysis, pos tagging and parsing. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, Paris, France. European Language Resources Association (ELRA).

Ivan Titov and Alexandre Klementiev. 2012. A bayesian approach to unsupervised semantic role induction. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '12, pages 12–22, Stroudsburg, PA, USA. Association for Computational Linguistics.