Transformer-based Swedish Semantic Role Labeling through Transfer Learning

Lucy Yang Buhr[1], **Dana Dannélls**[2], **Richard Johansson**[1]
[1]Department of Computer Science and Engineering,
[2]Språkbanken Text, Department of Swedish, multilingualism, language technology
[1]Chalmers University of Technology and University of Gothenburg, [2]University of Gothenburg
lucyyangxl@gmail.com, dana.dannells@svenska.gu.se, richard.johansson@cse.gu.se

Abstract

Semantic Role Labeling (SRL) is a task in natural language understanding where the goal is to extract semantic roles for a given sentence. English SRL has achieved state-of-the-art performance using Transformer techniques and supervised learning. However, this technique is not a viable choice for smaller languages like Swedish due to the limited amount of training data. In this paper, we present the first effort in building a Transformer-based SRL system for Swedish by exploring multilingual and cross-lingual transfer learning methods and leveraging the Swedish FrameNet resource. We demonstrate that multilingual transfer learning outperforms two different cross-lingual transfer models. We also found some differences between frames in FrameNet that can either hinder or enhance the model's performance. The resulting end-to-end model is freely available and will be made accessible through Språkbanken Text's research infrastructure.

Keywords: FrameNet, Semantic Role Labeling, Transfer Learning

1. Introduction

Semantic Role Labeling (SRL) is a task in natural language understanding where the goal is to extract information that could answer relational questions such as *who*, *what*, *where*, *why* about events mentioned in a text. Traditional statistical methods for developing SRL models have required syntactic annotation of the sentences before semantic information could be added (Xue and Palmer, 2004; Gildea and Jurafsky, 2002). In recent years, however, thanks to the breakthroughs in deep neural networks, such as Long Short-Term Memory Recurrent Neural Networks (LSTM-RNNs) (Şahin and Steedman, 2018; Marcheggiani et al., 2017; He et al., 2017) and Transformers (Tan et al., 2018; Strubell et al., 2018), SRL can be formulated as an end-to-end deep learning task without the need for morphological or syntactic information. FrameNet (FN)(Johnson et al., 2001) has emerged as one of the most prominent data resources for training SRL models. Typically, leveraging this resource allows the task to be broken down into three distinct sub-tasks, as illustrated in Figure 1:

1. *Trigger Identification*: determining that *Played* is a trigger, or frame-evoking expression.
2. *Frame Classification*: classifying the triggered frame PERFORMERS_AND_ROLES.
3. *Argument Extraction*: identifying the locations of the semantic spans and labeling them with their respective semantic roles: PERFORMER, ROLE and PERFORMANCE.

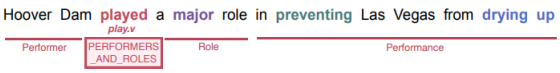

Figure 1: Frame semantic sentence annotations of the lexical unit *played* and its frame elements PERFORMER, ROLE and PERFORMANCE according to the frame PERFORMERS_AND_ROLES (from Swayamdipta et al. (2017)).

In case of English, SRL models trained on FN have achieved state-of-the-art (SOTA) performance. This is perhaps unsurprising given the large amount of annotated sentences available in the English resource, which is required to train these models from scratch. For languages with limited annotated semantic data, implementing Transformer modeling proves to be impractical. To address the challenge of requiring extensive training data, transfer learning (TL) methods have been suggested (Kalyanpur et al., 2020), but the method has never been tested for training Swedish SRL models. Furthermore, while text-to-text transfer techniques have previously been focused on new tasks or domains, to our knowledge, there are no studies on the suitability of these techniques for tasks involving multiple sub-tasks.

In this study we ask the following research questions: How suitable are pre-trained transfer learning models for developing SRL models? How should we develop a Swedish SRL model through transfer learning? Which semantic frames are more suitable for the training?

We explore these questions by experimenting

LREC-COLING 2024, pages 16762–16769

	Frames	Lexical Units	Sentences
SweFN	1 195	39 212	9 K
BFN	1 221	13 572	202 K

Table 1: SweFN statistics compared to Berkeley FrameNet 1.7

with two state-of-the-art transfer learning models: cross-lingual and multilingual using the frame semantic resource, Swedish FrameNet (SweFN) (Dannélls et al., 2021).

2. Related Work

Since the introduction of the Transformer architecture (Vaswani et al., 2017) and its variants, several studies have demonstrated the strengths of this architecture in its ability to leverage multitask learning in SRL (Tan et al., 2018; Strubell et al., 2018). Studies have shown that Transformer-based models trained on a large semantically annotated corpus, e.g. FrameNet, can outperform earlier SOTA SRL approaches (Kalyanpur et al., 2020; Nair and Bindu, 2021; Bakker et al., 2022).

Recently, Kalyanpur et al. (2020) have explored the Transformer architecture and its variants for frame-semantic parsing techniques by training on the Berkeley FrameNet (BFN) dataset. They compared several pre-trained Transformer-based models, such as encoder-decoder T5 and language generative model GPT-2 (Radford et al., 2019), and showed that the T5 model has improved the SOTA benchmark by 12-17% in prediction accuracy.

Oliveira et al. (2021) investigated the improvement of Portuguese SRL with Transformers and transfer learning and demonstrated that through transfer learning of a BERT-based model by adding a few layers, the new SRL model can surpass the previous state-of-the-art in Portuguese SRL by over 15 points in prediction accuracy.

For Swedish, some prior work has been done on SRL using older versions of SweFN (Johansson et al., 2012), predating modern deep neural network and Transformer architectures.

3. Data

The Swedish FrameNet (SweFN) (Dannélls et al., 2021) is a rich semantic Swedish resource that builds on frame semantics theory (Fillmore, 1982). In SweFN, the lexical units (LUs) evoking frames are all linked to word senses in the Swedish dictionary, SALDO (Borin et al., 2013), which is the largest computational dictionary for modern Swedish with over 147,000 word senses. Following the theoretical apparatus of FrameNet, each sense in SALDO corresponds to no more than one

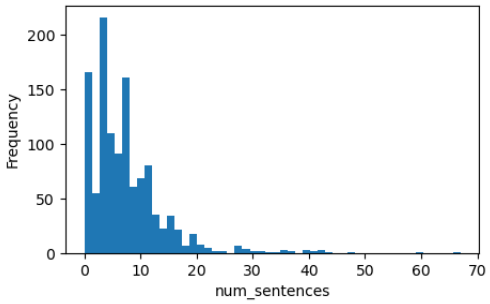

Figure 2: Distribution of number annotated sentences per frame in SweFN.

semantic frame in SweFN.[1]

Although SweFN is regarded as a Swedish version of Berkeley FrameNet (Fillmore et al., 2003), having identical frames and frame elements names, the information structure of these two databases are still different from each other in multiple ways. Moreover, there are frames in BFN that do not exist in SweFN and vise versa. Some frames have been modified and do not contain the same semantic roles.[2]

Furthermore, Berkeley FrameNet (BFN) contains both fully text annotation (approx. 6,000 sentences (Chanin, 2023)) and lexicographic annotation. With fully annotation, a sentence can be annotated to different frames with the associated LUs and frame elements (FEs) for each frame independently. Currently, SweFN has only annotations of single frame per sentence. If a sentence has several triggers that evoke the same frame, all the triggers will be annotated as LUs for this frame. Actually, 5% of the sentences in the SweFN dataset have more than one annotated trigger. This limitation of annotating a single frame per sentence becomes apparent when identifying potential triggers for multiple frames. In the current dataset and the chosen implementation, the Swedish SRL model developed in this study can only identify a single trigger in a sentence, even when there might be multiple frames present.

There are several versions of BFN. The latest release is Berkeley FrameNet 1.7 (Ruppenhofer et al., 2016), containing more than 200K annotated examples (approx. 20 sentences per LU). As shown in Table 1, there is currently a significantly smaller number of annotated sentences in SweFN compared to BFN. In Swedish FrameNet, the annotated examples do not cover all LUs or frames. In addition, the available examples are not distributed

[1]The resource is available through Språkbanken Text's research infrastructure: https://spraakbanken.gu.se/.

[2]https://spraakbanken.gu.se/projekt/swedish-framenet-swefn/documentation-for-swefn

evenly among frames (shown in Figure 2), 860 out of 1195 frames contain less than 10 example sentences. The frame EMPTYING has 67 annotated sentences, the most in quantity.

With fewer annotated examples and an increased number of triggers in SweFN, training an SRL model becomes a more challenging task. Furthermore, the extremely unbalanced dataset distribution across semantic frames suggests that it is not reasonable to expect uniform performance across all frames for Swedish SRL if the model is developed based on the SweFN data only.

4. Experiments

4.1. Model 1: Cross-Lingual TL

Several studies have demonstrated that an English T5 model can perform cross-lingual transfer surprisingly well (Tran, 2020; Chi et al., 2020; Li et al., 2021; Chi et al., 2020). This may be due to the fact that the English T5 model has encountered 0.22% of non-English text during its pre-training phase, as well as the resemblance of the new target language to English (Blevins and Zettlemoyer, 2022).

T5-based *Frame Semantic Transformer* (Chanin, 2023), in particular, has several considerable advantages, including: (1) covering three sub-tasks in a pipeline fashion, (2) achieving state-of-the-art performance, (3) open-source and easy-of-use, and (4) allowing cross-lingual transfer capability of a T5 model. This English SRL model has therefore become an obvious choice to be used as the primary transfer model in this study.

The T5-based English-SRL model comes with two versions: "small" and "base". The latter performs better, making it our preferred choice. However, to speed-up the training setup and quickly gain experience of a Swedish SRL model, we initially begun with the small version, and subsequently transitioned to the base variant. Our Swedish SRL models developed from this approach are called MODEL-1-SMALL and MODEL-1-BASE respectively. Due to missing an existing Swedish SRL model to benchmark, we choose MODEL-1-SMALL as the baseline in this study.

However, we have to remember that a T5 model is not the best choice for cross-lingual transfer due to the lack of strong relationships established among multiple languages through significant exposure to these languages. In addition, Swedish special characters such as *ä, å, ö* are not in the vocabulary of T5 tokenizer by default, and therefore these unknown characters are converted to "<unk>". This conversion makes it harder for the model to learn the relationships among Swedish words that contain special characters.

4.2. Model 2: Multi-Task TL

Inspired by the English SRL model, particularly its ability to perform all three sub-tasks directly in one pass, we had initially intended to deploy a Swedish version of T5 as the second transfer model for transfer-learning on SRL tasks. However, lacking such a model, we opted for the multilingual T5 (mT5) as our alternative choice.

This alternative choice has its advantages too. After being pre-trained on more than 100 languages that include English and Swedish, a mT5 model has established weight-parameters that can represent the dependency between English and Swedish words. As mentioned earlier, in the SweFN corpus, both the frame names and the FEs names are in English, while the input sentence and the lexicon-units are in Swedish. Therefore, the Swedish SRL model actually needs to "understand" the relationship between these two languages. A mT5 model, with already obtained attention score between the targeted English names (Frame names or FEs names) and the relevant Swedish words in the input sentence, indeed has an advantage for the multi-task transfer learning on SRL sub-tasks.

The author of the mT5 model has shown that the largest version of mT5 (MT5-XXL) outperforms previous popular multilingual models for most NLP tasks (Xue et al., 2021). However, as the mT5 is pre-trained on more than 100 languages, even its smallest version (MT5-SMALL) is five times bigger than T5-SMALL. Due to the constraints of our computer capacity, it is hard for us to train a bigger variant of mT5. Therefore, in this study, we use only the MT5-SMALL, from which the further developed Swedish SRL model becomes the MODEL-2-SMALL. Experimenting with the small model variants can speed up the training process while still enabling a fair benchmarking between these two TL approaches. Moreover, since mT5 has not yet gone through any supervised training to learn any NLP downstream task, direct fine-tuning on a small dataset is not optimal, and will also need a longer learning process than a model like T5 that has been pre-trained by multiple downstream tasks.

The absence of supervised pre-training, combined with its unnecessarily large size for the Swedish SRL task, makes the mT5 model less computationally cost-effective than a T5 model. This emerges as the primary drawback of utilizing mT5 for the development of our Swedish SRL model.

5. Results and Discussion

5.1. Quantitative Results

After fine-tuning 61K steps (43 epochs) for MODEL-1-SMALL, 79K steps (14 epochs) for MODEL-1-BASE, and 152K steps (27 epochs) for MODEL-2-SMALL, we

F1-score (val/test)	Trigger ID	Frame CL	Argument EX
Model-1-small	0.50/0.50	0.60/0.60	0.52/0.53
Model-1-base	0.52/0.47	0.62/0.63	**0.57/0.58**
Model-2-small	**0.56/0.52**	**0.64/0.67**	0.54/0.55
English-SRL-small (Chanin, 2023)	0.74/0.70	0.83/0.81	0.68/0.70
English-SRL-base (Chanin, 2023)	0.78/0.71	0.89/0.87	0.74/0.72

Table 2: Model comparison through F1 scores on SweFN validation (val) set and test set respectively, reported per task type: trigger identification (Trigger ID), Frame classification (Frame CL), and Argument extraction (Argument EX). The last 2 rows are the reported F1 scores on BFN dataset of the English SRL models that is referred in this study.

observed the best performance of the respective model. These best results per model type and sub-task type are summarized in Table 2.[3]

Although, in terms of F1 scores, there are still significant gaps comparing to the SOTA English SRL models (Chanin, 2023), our Swedish SRL models have reached an acceptable performance when taking into account the more challenging SweFN dataset we have. Especially, on the most important sub-task Argument Extraction. The larger model developed from the English SRL (MODEL-1-BASE) outperforms previous Swedish SRL models and has a smaller performance gap compared to the targeted performance of the English SRL model on the English FrameNet data. The development process of fine-tuning the SOTA English SRL models is available in the Appendices A and B.

Notably, the results of the Swedish and the English FN models are not directly comparable because these results are based on our own test set, rather than on a translated version of the English test set. When we look more closely into the two test sets (the one used to evaluate the Swedish model compared to the set that was used to evaluate the English one) we find that the sentences included in the sets were taken from different frames. However, they are from the same distribution, genre and domain. What seems to differ between them is the amount of frame elements in each of the annotated sentences.

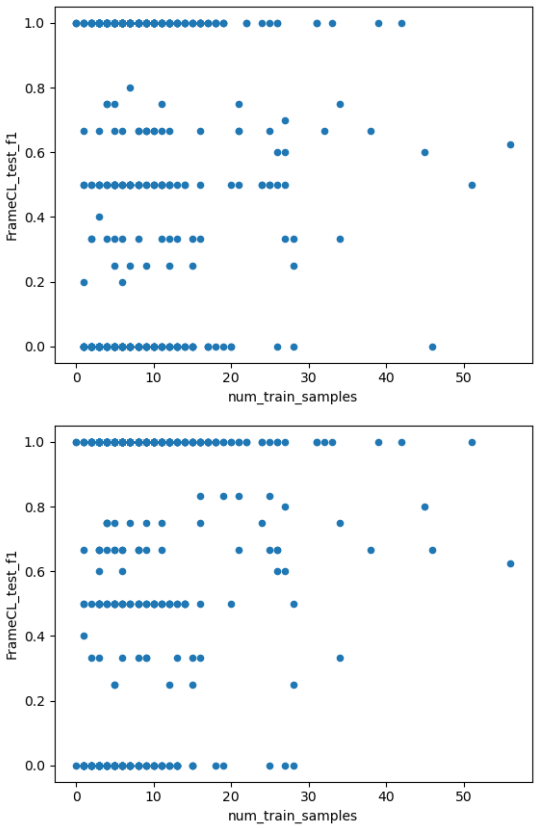

Figure 3: On Frame Classification task: the relationship between prediction performance (F1 test score on *Y-axis*) and number trained samples on *X-axis*. UPPER: predicted by MODEL-1-SMALL; LOWER: predicted by MODEL-2-SMALL

5.2. Qualitative Results

Performance per Frame Out of the three SRL sub-tasks, Frame Classification and Argument Extraction tasks should be the most relevant for analysis per frame.[4] Intuitively, the more training samples for a frame, the better prediction for this frame. We scattered two plots to examine this relationship, see Figure 3. To our surprise, we learn that more training samples do not have a straightforward positive impact. Indeed, frames trained with almost zero samples can also have 100% prediction accuracy. But for frames with more than 20 training sentences, the majority can have prediction accuracy above 50%. When comparing the prediction accuracy between MODEL-1-SMALL and MODEL-2-SMALL, we can see that our mT5 has improved performance on all frames regardless the size of the trained samples.

[3]All models are open source and are available here: https://github.com/lucyYB/SweFN-SRL

[4]As Argument Extraction is a more complex problem and harder to visualize in one graph, we choose to focus on Frame Classification in this analysis.

Easy frames are frames with fewer than 3 training samples that achieve 100% accuracy in Frame Classification on the test set. Thirteen common easy frames, listed in Appendix C, were identified for both MODEL-1-SMALL and MODEL-2-SMALL using this definition. Easy frames share some common features that make them easier to get a corrected prediction, such as: (1) they contain non-polysemous words, occurring with only one sense in SALDO, and therefore have a single, clear, and unambiguous meaning. For example, the word *begränsning* 'limit' occurs only with this particular sense in SALDO, therefore it is only listed under LIMITATION; (2) they include words containing annotated examples in this single frame only. For example, all the examples included in EVENT_INSTANCE are triggered by the lemma *gång* 'occasion', and this trigger has only been trained under this frame even though it is listed under multiple frames.

Surprisingly, there are also some frames which seem to be easy for MODEL-1 but not for MODEL-2, such as: COMMUTATIVE_PROCESS, LOSING_IT, ROBBERY, SERVING_IN_CAPACITY. Most often, this happens when the marked trigger has not been seen together with the target frame in the training set, but appears only in the list of this frame, hence in the hint list. Indeed, the English SRL-based model seems to rely more on the hints in the input prompt, but not the mT5 transferred model, which seems to be more capable of predicting with its established dependency between the English frame name and the Swedish trigger words. However, occasionally, the MODEL-2-SMALL can also make up a fake frame name through its pre-established connections among Swedish and English words.

Difficult frames are frames with more than 20 training examples but achieving less than 60% accuracy in Frame Classification on the test set. These frames, listed in Appendix D, are the most challenging to predict. These challenges typically arise due to the combination of the following common factors: (1) The trigger word is a compound word, but the frame lists a multi-word expression, for example, *vältra_bort* 'tip over' is one of the LUs listed in the frame CAUSE_MOTION but its compound format *bortvältrade* are used in the test sentences; (2) The trigger, appearing in its inflected form, and the LU, listed in the frame do not match each other, probably because of the limitation of Snowball stemmer, resulting in an empty hint in the prompt. For example, there is a mismatch between the trigger *åket* and the LU *åk..2* that is listed in the frame PART_ORDERED_SEGMENTS. (3) The trigger is less common and has never or very seldom been trained together with the target frame, thus no strong dependency can be established between the frame name and the trigger words. (4) Too many matched LUs for a trigger. Hence, many suggested frames in the hint, but this trigger has primarily been trained for other frames.

At the same time, as seen in Figure 3, MODEL-2-SMALL has significantly fewer difficult frames than MODEL-1-SMALL. This improved prediction usually happens when the hint list is empty in the input prompt. Indeed, the mT5-based model not only relies on the hints as much as the English SRL-based model does, but rather on the learned relationship between the trigger and the target frame, giving it the advantage of predicting the correct frame name.

5.3. Discussion

To further improve the model performance, especially on the difficult frames due to compound word or inflected form, we believe one should try with utilizing multiple stemmers and lemmatizers. While, for the rare triggers or too many possible LUs in the hint list, improving lemmatization will not help. For these kinds of difficult frames, introducing diversity and enriching the data set can be an alternative. Furthermore, if our experiments had not been constrained by the limited computer capacity, we would have tried with larger mT5 models. A larger model has much higher capacity of storing and recognizing relationships between words, which leads to the model being less dependent on the small SweFN training examples.

6. Conclusion

In this paper, we have presented the first effort in building a Transformer-based SRL system for Swedish. In our study, we have explored the suitability of applying multilingual and cross-lingual transfer learning methods for developing an SRL model for Swedish. The results of our experiments show that fine-tuning mT5 on English and Swedish data improves performance for the first two tasks of SRL, namely trigger identification and frame classification. However, the English-only T5 model is still competitive in its base version even though it is 5 times smaller and, surprisingly, it outperforms the multilingual version on argument extraction when using the English-only T5 base model. We also found some differences between frames that can either hinder or enhance the model's performance. We conclude that by leveraging the rich annotations from Berkeley FrameNet and the cross-lingual dependencies of a multilingual Transformer system, we can not only increase prediction accuracy but also identify multiple triggers in a sentence.

For model comparison, a possible future direction includes altering the training sequence of the three sub-tasks and translating the test data from English to provide comparable model results.

7. Ethics Statement

The FrameNet data is open-source. In the original resources all sentences were semantically annotated manually. They may, however, potentially contain some offensive language that our models were trained on.

8. Acknowledgements

The research presented here has been enabled by the Swedish national research infrastructure Nationella språkbanken, funded jointly by the Swedish Research Council (2018–2024, contract 2017-00626) and the 10 participating partner institutions. We gratefully acknowledge Huminfra, a Swedish national infrastructure for the Humanities, funded by the Swedish Research Council and the consortium nodes, grant number 2021-00176. This research was funded by the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation, and the Wallenberg AI, Autonomous Systems and Software Program – Humanities and Society (WASP-HS) funded by the Marianne and Marcus Wallenberg Foundation and the Marcus and Amalia Wallenberg Foundation.

9. Bibliographical References

Roos Bakker, Romy A.N. van Drie, Maaike de Boer, Robert van Doesburg, and Tom van Engers. 2022. Semantic role labelling for Dutch law texts. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 448–457, Marseille, France. European Language Resources Association.

Terra Blevins and Luke Zettlemoyer. 2022. Language contamination helps explains the cross-lingual capabilities of English pretrained models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3563–3574, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Lars Borin, Markus Forsberg, and Lennart Lönngren. 2013. SALDO: a touch of yin to WordNet's yang. *Language Resources and Evaluation*, 47(4):1191–1211.

David Chanin. 2023. Open-source frame semantic parsing. *arXiv preprint arXiv:2303.12788*.

Zewen Chi, Li Dong, Furu Wei, Xianling Mao, and Heyan Huang. 2020. Can monolingual pretrained models help cross-lingual classification? In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 12–17, Suzhou, China. Association for Computational Linguistics.

Dana Dannélls, Lars Borin, Markus Forsberg, Karin Friberg Heppin, and Maria Toporowska Gronostaj. 2021. Swedish FrameNet. In *The Swedish FrameNet++. Harmonization, integration, method development and practical language technology applications*, pages 37 – 66. John Benjamins Publishing Company, Amsterdam / Philadelphia.

Charles J. Fillmore. 1982. Frame semantics. In Linguistic Society of Korea, editor, *Linguistics in the Morning Calm*, pages 111–137. Hanshin Publishing Co., Seoul.

Charles J. Fillmore, Christopher R. Johnson, and Miriam R.L. Petruck. 2003. Background to FrameNet. *International Journal of Lexicography*, 16(3):235–250.

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

Luheng He, Kenton Lee, Mike Lewis, and Luke Zettlemoyer. 2017. Deep semantic role labeling: What works and what's next. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 473–483, Vancouver, Canada. Association for Computational Linguistics.

Richard Johansson, Karin Friberg Heppin, and Dimitrios Kokkinakis. 2012. Semantic role labeling with the Swedish FrameNet. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 3697–3700, Istanbul, Turkey. European Language Resources Association (ELRA).

Christopher Johnson, Charles J Fillmore, Esther Wood, Josef Ruppenhofer, Margaret Urban, Miriam Petruck, and COLLIN Baker. 2001. The framenet project: Tools for lexicon building. *Manuscript. Berkeley, CA, International Computer Science Institute*.

Aditya Kalyanpur, Or Biran, Tom Breloff, Jennifer Chu-Carroll, Ariel Diertani, Owen Rambow, and Mark Sammons. 2020. Open-domain frame semantic parsing using transformers. *arXiv preprint arXiv:2010.10998*.

Zuchao Li, Kevin Parnow, Hai Zhao, Zhuosheng Zhang, Rui Wang, Masao Utiyama, and Eiichiro Sumita. 2021. Cross-lingual transferring of pretrained contextualized language models. *arXiv preprint arXiv:2107.12627*.

Diego Marcheggiani, Anton Frolov, and Ivan Titov. 2017. A simple and accurate syntax-agnostic neural model for dependency-based semantic role labeling. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 411–420, Vancouver, Canada. Association for Computational Linguistics.

Archana M Nair and KR Bindu. 2021. Semantic role labelling using transfer learning model. *Journal of Physics: Conference Series*, 1767(1).

Sofia Oliveira, Daniel Loureiro, and Alípio Jorge. 2021. Improving Portuguese Semantic Role Labeling with Transformers and Transfer Learning. *arXiv preprint arXiv:2101.01213*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Josef Ruppenhofer, Michael Ellsworth, Miriam R.L. Petruck, Christopher R. Johnson, Collin F. Baker, and Jan Scheffczyk. 2016. FrameNet II: Extended theory and practice. Technical report. Berkeley: ICSI.

Gözde Gül Şahin and Mark Steedman. 2018. Character-level models versus morphology in semantic role labeling. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 386–396, Melbourne, Australia. Association for Computational Linguistics.

Emma Strubell, Patrick Verga, Daniel Andor, David Weiss, and Andrew McCallum. 2018. Linguistically-informed self-attention for semantic role labeling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5027–5038, Brussels, Belgium. Association for Computational Linguistics.

Swabha Swayamdipta, Sam Thomson, Chris Dyer, and Noah A. Smith. 2017. Frame-semantic parsing with softmax-margin segmental rnns and a syntactic scaffold. *CoRR*, abs/1706.09528.

Zhixing Tan, Mingxuan Wang, Jun Xie, Yidong Chen, and Xiaodong Shi. 2018. Deep semantic role labeling with self-attention. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA*, pages 4929–4936. AAAI Press.

Ke Tran. 2020. From English to foreign languages: Transferring pre-trained language models. *arXiv preprint arXiv:2002.07306*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6000–6010, Red Hook, NY, USA. Curran Associates Inc.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Nianwen Xue and Martha Palmer. 2004. Calibrating features for semantic role labeling. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 88–94, Barcelona, Spain. Association for Computational Linguistics.

10. Appendix

Appendix A. Model 1: Fine-tuning the English SRL

Figure 4 and Figure 5 show the development when fine-tuning the SOTA English SRL models. As we can see in the upper graph of each figure, the validation loss starts to increase after a while, even though the F1 scores are continually improving on validation set. Note, it is not overfitting as long as the accuracy is still increasing on unseen data.[5]

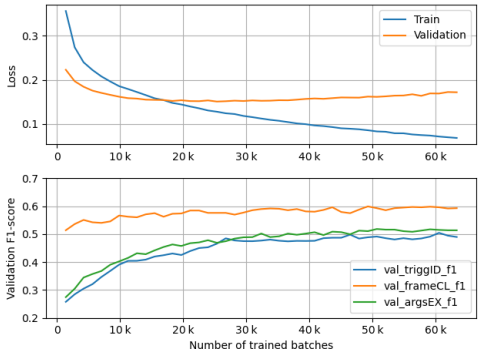

Figure 4: MODEL-1-SMALL: loss development (UPPER) and validation F1-score for each sub-task (LOWER) during 45 epochs with batch-size of 16.

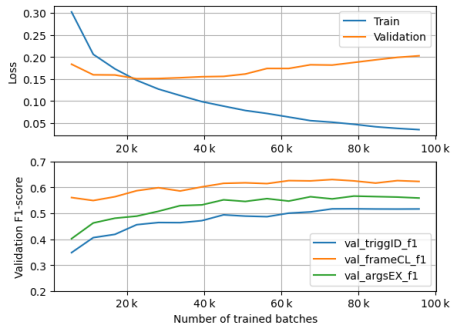

Figure 5: MODEL-1-BASE: loss development (UPPER) and validation F1-score for each sub-task (LOWER) during 17 epochs with batch-size of 4.

Appendix B. Model 2: Fine-tuning the mT5

As we can see from Figure 6, because the mT5-base model has not been fine-tuned on any NLP downstream task before, we expect that the F1-scores start from a much lower level than the previous two models. It is the case on all sub-tasks except for "Frame classification", the performance on which has already reached a similar level as the Model-1-small after just one epoch.

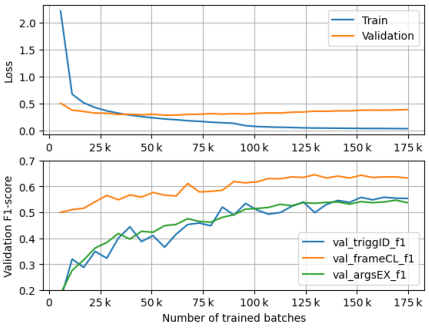

Figure 6: MODEL-2-SMALL: loss development (UPPER) and validation F1-score for each sub-task (LOWER) during 31 epochs with batch-size of 4.

Appendix C. Easy Frames

BEING_QUESTIONABLE, CAUSE_EMOTION, CONFERRING_BENEFIT, DYNAMISM, EVENT_INSTANCE, JUDICIAL_BODY, GIZMO, LIMITATION, PROCESS_RESUME, SUBJECTIVE_TEMPERATURE, SURRENDERING, SURRENDERING_POSSESSION, and TYPICALITY.

Appendix D. Difficult Frames

SIMULTANEITY, REMOVING, FILLING, CAUSE_MOTION, PART_ORDERED_SEGMENTS, and PROLIFERATING_IN_NUMBER.

[5] T5 (as well as mT5) uses the standard Cross-Entropy loss function to compare each token of the model predicated sequence against the target sequence. Cross-Entropy loss is commonly used by any language model due to its mathematical property, and is explained *here*. While the confusion metrics used for evaluation of the model performance, are entirely different calculations than Cross-Entropy loss, and therefore do not always follow the same developing trend.