EasyChair Preprint

№ 10380

Can the Rich Help the Poor? Transfer of Knowledge and Resources for Under Resourced Languages Semantic Role Labeling

Yesuf Mohamed and Wolfgang Menzel

June 11, 2023

Can the Rich Help the Poor? Transfer of Knowledge and Resources for under resourced languages Semantic Role Labeling

Yesuf Mohamed
yesufcomp1@gmail.com
Addis Ababa, Ethiopia

Wolfgang Menzel (Professor)
wolfgang.menzel@uni-hamburg.de
Hamburg University, Germany

Abstract

Semantic Role Labeling (SRL) is a crucial natural language processing task that involves assigning semantic roles to words or phrases in a sentence. While SRL techniques have been extensively studied for well-resourced languages, under-resourced languages face significant challenges due to the lack of annotated data and language-specific resources. In this paper, we explore the potential of leveraging knowledge and resources from richly-resourced languages/domains to improve SRL performance in under-resourced languages. We provide an overview of the SRL process, discuss the challenges faced in under-resourced languages, and present techniques for transferring knowledge and resources from rich languages/domains. By utilizing resources from rich languages, we can overcome some of the challenges in SRL for under-resourced languages. The advantages include access to pre-trained models, lexical resources, and annotated data, which can improve SRL accuracy. However, there are also challenges such as language divergence and domain mismatch that need to be addressed. We discuss these challenges and propose possible solutions, including domain adaptation and data augmentation techniques. Finally, we conclude by emphasizing the importance of leveraging the resources of rich languages to advance SRL in under-resourced languages.

Key Words: cross-lingual Transfer learning, domain adaptation, under resourced languages

1. Introduction

Semantic Role Labeling (SRL) is a natural language processing (NLP) task that aims to identify the underlying meaning or semantic structure of a sentence by identifying the relationships between its constituents. Specifically, SRL involves identifying the semantic roles played by each word or phrase in a sentence, such as the agent (the doer of an action), the patient (the entity that undergoes the action), the instrument (the tool or means used to perform an action), and the location (the place where an action occurs) [1]. SRL is a crucial step in many NLP applications, such as question answering, information retrieval, and machine translation, as it helps to disambiguate sentence meaning and improve accuracy [2]. Generally, SRL answers the question "Who did What to Whom, and How, When and Where?" in text [3].

The lack of annotated data and the linguistic differences between high-resourced and under-resourced languages pose significant challenges to this task. However, leveraging the resources and knowledge of riches[1], to improve SRL accuracy in under-resourced languages is an active area of research. Nowadays, one of the methods used to resolve this issue is cross-lingual transfer learning [4].

Cross-lingual transfer learning involves leveraging the knowledge learned from a high resourced language to improve the performance of a SRL model in an under resourced language, even if there is limited or no labeled data available in the target language [5]. This approach is particularly useful in scenarios where there is a scarcity of labeled data in the target language, which makes it challenging to train high-performing models. One of the most common approaches to cross-lingual transfer learning in SRL involves using models trained on large-scale datasets in a source language, such as English, and adapting them to a target language.

In this essay, we will discuss the cross-lingual transfer of knowledge and resources from riches to under-resourced languages for SRL. We will also mention the challenges and limitations to build automatic semantic role labeling system for an under-resourced language, and the question how can these challenges be addressed? Will be answered.

1. Riches refers to other languages or domains within the same language with high resources and knowledge

2. Overview of Semantic Role Labeling for Under-Resourced Languages

2.1. Explanation of the semantic role labeling process

The goal of SRL is to identify the relationships between the words or phrases in a sentence and the actions they describe [6]. The resulting semantic roles help to disambiguate the meaning of the sentence and facilitate various NLP applications, such as machine translation, information retrieval, and question answering. [7]

The semantic role labeling process typically involves several steps. The first step is to identify the main predicate or verb in the sentence, which is the word that describes the action being performed. Once the main predicate is identified, the next step is to identify the arguments of the predicate, which are the words or phrases that play specific roles in the action described by the predicate [8].

There are several types of semantic roles that can be assigned to words or phrases, including the agent, which is the entity that performs the action; the patient, which is the entity that is affected by the action; the instrument, which is the tool or means used to perform the action; and the location, which is the place where the action occurs [1].

The semantic role labeling process can be performed manually, or automatically using machine learning techniques. Manual annotation involves human annotators identifying the semantic roles in a sentence, which can be time-consuming and expensive. On the other hand, automatic semantic role labeling involves training a machine learning model on annotated data, which can then be used to automatically assign semantic roles to new sentences. Recent advances in deep learning techniques have led to significant improvements in the accuracy and efficiency of automatic semantic role labeling, making it a critical component of many NLP applications.

2.2. Challenge in semantic role labeling of under-resourced languages

The main challenge in developing automatic SRL system for under-resourced language is the lack of annotated data. In general, the accuracy of automatic semantic role labeling system highly depends on the amount and quality of annotated data available for training. However, for many under-resourced languages, there is a scarcity of annotated data, making it difficult to develop an accurate semantic role labeling system.

2.3. Techniques for leveraging knowledge and resources from riches

There are various techniques to leverage knowledge and resources of riches and improve the performance of semantic role labeling (SRL) for under-resourced languages. These techniques include:

Cross-lingual transfer learning: This approach involves using pre-trained models in high-resourced languages and adapting them to under-resourced languages. The pre-trained model is fine-tuned using a small amount of labeled data in the under-resourced language, and the knowledge learned from the high-resourced language is transferred to the under-resourced language. This approach has shown promising results in improving SRL accuracy for under-resourced languages.

Cross-lingual transfer learning has become possible due to advances on machine learning and natural language processing (NLP) in recent years. One key development has been the availability of large amounts of data in multiple languages, which allows models to be trained on a diverse range of language samples. Advances in data availability, pre-trained language models, and evaluation methods have all contributed to the feasibility of cross-lingual transfer learning techniques.

Multi-task learning: In this approach, SRL is jointly learned with another related task such as part-of-speech labeling or named entity recognition. By learning multiple related tasks simultaneously, the model can learn shared representations that can improve the performance of SRL in under-resourced languages [11].

Domain Adaptation: Domain adaptation is a technique for adapting a model trained on one domain to another domain. In SRL, this means training a model on labeled data from one domain (e.g., news articles) and then adapting the model to perform well on data from another domain (e.g., scientific papers).

Domain adaptation can be useful for SRL because labeled data can be scarce or expensive in some domains, and it may be more efficient to use a pre-trained model and adapt it to the target domain than to train a new model from scratch. For example, suppose we have a pre-trained SRL model that has been trained on a large dataset of news articles. We want to adapt this model to perform well on a dataset of scientific papers. We could use a domain adaptation technique, such as fine-

tuning or transfer learning, to adapt the pre-trained model to the new domain. This could help us improve the performance of the SRL system on the new domain without requiring a large amount of labeled data.

Data selection is another way to perform domain adaptation. When a small corpus of domain-specific data is available, and a larger corpus of out-of-domain data is available, it is possible to select the most similar data items from the out-of-domain corpus to extend the in-domain collection. This approach can be useful when in-domain data is limited, but it is necessary to build a larger corpus for domain-specific model training.

2.4. Evaluation metrics of Different Techniques for Semantic Role Labeling

The following evaluation metrics are used in SRL papers [9] [10] individually or in combination to evaluate the efficiency and effectiveness of SRL systems.

Precision: Precision measures the percentage of correctly identified semantic roles out of all the roles predicted by the SRL system. A high precision score indicates that the SRL system is accurate in identifying the semantic roles.

Recall: Recall measures the percentage of correctly identified semantic roles out of all the roles present in the text. A high recall score indicates that the SRL system is effective in identifying all the semantic roles in the text.

F1-score: The F1-score is the harmonic mean of precision and recall. It provides a balanced measure of both precision and recall and is often used as a single evaluation metric for SRL systems.

Error rate: The error rate measures the percentage of incorrectly identified semantic roles out of all the roles predicted by the SRL system. A low error rate indicates that the SRL system is accurate in identifying the semantic roles.

2.5. Case studies and examples

To demonstrate the techniques for SRL in under-resourced languages, we will discuss specific case studies in which semantic role labeling was applied.

One such case study is the work of Ilseyar Alimova et al. [4] who applies cross-lingual transfer learning to improve SRL accuracy in the under-resourced language of Russian. They use a semantic role labeling model pre-trained on a high-resourced language, English, and a Russian FrameBank dataset for evaluation. The results shows that the transfer learning approach significantly improved the SRL accuracy in Russian, achieving a good performance.

The study of Fariz Ikhwantri et al. [12] use the multi-task active learning method, and they trained their model with only 6057 annotated sentences, the primary task is semantic role labeling and the second task is entity recognition. They achieved good results by using both active and multi-task learning methods.

To overcome the dataset scarcity for multi-lingual conversational semantic role labeling, the study [13] applied a zero-shot cross-lingual semantic role labeling (CSRL) method to non-Chinese languages. To overcome the dataset scarcity problem and to implicitly learn the language's semantic representations, hierarchical encoders and pre-training objectives are used. The cross-lingual model outperforms different baselines by large margins and is robust to low-resource scenarios. The use of CSRL information helps downstream English conversational tasks, including question-in-context rewriting and multi-turn dialogue response generation, achieving substantial improvements.

To improve the performance of the automatic Semantic Role Labeling task for low-resource languages specifically Portuguese, Sofia Oliveira et al. [5], explored a model architecture that only uses a pre-trained Transformer-based model. They also leverage cross-lingual transfer learning using multilingual pre-trained models and transfer learning from dependency parsing in Portuguese to further improve the results. The authors were able to achieve a substantial improvement in the state-of-the-art performance for Portuguese by over 15 F1 percentage points

3. Advantages and challenges of using the resources of riches for semantic role labeling of under-resourced languages

In this section, we will discuss the advantages and challenges of using resources from riches for SRL in under-resourced languages.

3.1. Advantages:

Improved performance: The use of resources and knowledge from riches can significantly improve the performance of SRL in under-resourced languages. This is because riches have more annotated data, which can be used to train SRL models with higher accuracy.

Reduced annotation cost: The use of resources and knowledge from riches can also reduce the cost of annotation in under-resourced languages or tasks. Instead of annotating data from scratch, transfer learning techniques can be used to adapt pre-trained models from riches to under-resourced languages. This can significantly reduce the amount of labeled data needed for training.

Faster development: The use of resources from riches can also speed up the development of SRL systems for under-resourced languages. Instead of starting from scratch, pre-trained models and tools from riches can be used as a starting point for developing SRL systems for under-resourced languages.

3.2. Challenges:

The main challenges of using rich's knowledge and dataset to train a semantic role labeling system for an under-resourced language are: -

Linguistic differences: One major challenge to transfer resources and knowledge from high-resourced languages is the existence of linguistic differences between the languages involved. For example, languages with different word order, and morphological complexity can affect the performance of SRL models trained on high-resourced languages. High-resourced languages such as English have a grammatical structure that might not be present in under-resourced languages. For instance, Amharic, has a different grammatical structure than English for example, if we take subject verb agreement, an Amharic verb can take more information than an English verb. a single Amharic verb can take number, gender, and tense information in it. Therefore, directly applying a semantic role labeling system trained on English data to Amharic would not yield accurate results.

Domain differences: Another challenge is the existence of domain differences between riches dataset and under-resourced languages datasets [3]. The riches resources may be in a different domain, which may not be relevant to the under-resourced language. This can lead to poor performance of the SRL system in the target domain or language.

Data availability: Finally, the availability of annotated data in the under-resourced language can be a challenge. While transfer learning can reduce the amount of labeled data needed for training, for better accuracy, some labeled data is still required. If there is a lack of annotated data in the under-resourced language, it can be difficult to develop an accurate SRL system.

Resource constraints: cross lingual transfer learning needs large amount of high-resourced dataset. That means, if you want to train an SRL system for an under-resourced language like Amharic, using riches dataset, you may need significant computational resources to process it, which may not be available. This can lead to longer training times, increased costs, and reduced efficiency of the SRL system.

Overall, the use of resources from riches offers promising directions for improving SRL in under-resourced languages. However, as discussed above, there are several challenges that need to be addressed to ensure the effectiveness of these approaches.

3.3. Possible Solutions:

One possible solution to the challenges of using riches dataset for training a semantic role labeling system for an under-resourced language is, to adapt the dataset to the target language or task. This involves developing a mapping between the semantic roles in the riches and the target language or task. For instance, the system can identify the similarities and differences between the two languages' syntactic and sematic structures and create a mapping that takes into account the differences.

4. Future Directions of Semantic Role Labeling for Under-Resourced Languages

The development of accurate SRL systems for under-resourced languages is crucial for enabling natural language processing (NLP) applications for these languages. However, there is still much work to be done to improve the accuracy of SRL for under-resourced languages. In this section, we will discuss some future directions of SRL research for under-resourced languages.

Transfer learning techniques: Transfer learning techniques have shown promising results in adapting pre-trained models from high-resourced languages to under-resourced languages. Future research can explore better transfer learning techniques and investigate their effectiveness in improving SRL accuracy for under-resourced languages.

Multi-task learning: The exploration of advancements on multi-task learning for semantic role labeling in under-resourced languages in future research has the potential to yield improved techniques for training models to execute multiple tasks concurrently with greater efficacy. This could involve the creation of new types of models that can effectively learn from multiple tasks and datasets, as well as the development of new methods for combining different types of tasks to improve overall performance.

Domain adaptation: Further investigation into domain adaptation in future research could open up the possibility of innovative and more efficient methods for adapting semantic role labeling models, originally trained on one domain, to operate effectively in another domain. This could involve the creation of new types of models that can effectively learn from limited domain-specific labeled data, as well as the development of new methods for leveraging unlabeled data in the target domain.

Advancements in technology and data collection methods like can also play a critical role in improving the accuracy of SRL in under-resourced languages.

5. Conclusion

In this paper, we have discussed semantic role labelling, the challenges and limitations of using rich's dataset for training a semantic role labeling system for an under-resourced language. We have also discussed the techniques used to overcome these challenges and improve SRL accuracy in under-resourced languages. We have highlighted the importance of case studies and examples in evaluating the effectiveness of different techniques for improving SRL accuracy. We have also discussed the advantages and challenges of using resources of riches to train SRL models for under-resourced languages.

Furthermore, we have discussed future directions of SRL research for under-resourced languages, including transfer learning techniques, multi-task learning, domain adaptation and advancements in the technology and methods for data collection.

In conclusion, developing accurate SRL systems for under-resourced languages is crucial for enabling NLP applications in these languages. While there are still many challenges to be addressed, recent advancements in transfer learning, multi-task learning, and data collection methods provide promising opportunities for improving SRL accuracy in under-resourced languages.

References

[1] Daniel Gildea et al, "Automatic Labeling of Semantic Roles," *Computational Linguistics,* vol. 28, p. 245–288, 2002.

[2] Jo˜ao Sequeira et al, "Semantic Role Labeling for Portuguese – A Preliminary Approach –," in *Computational Processing of the Portuguese Language*, Berlin, 2012.

[3] Quynh Ngoc Thi Do et al, "Facing the most difficult case of Semantic Role Labeling: A collaboration of word embeddings and co-training," in *International Conference on Computational Linguistics(COLING): Technical*, Osaka, Japan, 2016.

[4] Ilseyar Alimova et al, "Cross-lingual transfer learning for semantic role labeling in Russian," *Proceedings of the 4th International Conference on Computational Linguistics,* p. 72–80, 2020.

[5] Sofia Oliveira et al, "Improving Portuguese Semantic Role Labeling with Transformers and Transfer Learning," *8th International Conference on Data Science and Advanced Analytics (DSAA),* pp. 1-9, 2021.

[6] Matthew R. Gormley et al, "Low-Resource Semantic Role Labeling," in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, Baltimore, Maryland, USA,, 2014.

[7] B. M. Hailu, "SEMANTIC ROLE LABELING FOR AMHARIC TEXT USING DEEP LEARNING," 2021.

[8] Daniil Larionov et al, "Semantic Role Labeling with Pretrained Language Models for Known and Unknown Predicates," *Proceedings of Recent Advances in Natural Language Processing,* pp. 619–628,, 2019.

[9] Aashish Arora et al, "Multi-Task Learning for Joint Semantic Role and Proto-Role Labeling," 2022.

[10] Màrquez et al, "Semantic Role Labeling: An Introduction to the Special Issue," *Computational Linguistics,* vol. 34, p. 145–159, 2008.

[11] Mikhail Kozhevnikov, Ivan Titov, "Cross-lingual Transfer of Semantic Role Labeling Models," in *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, Sofia, Bulgaria, 2013.

[12] Fariz Ikhwantri et al, "Multi-Task Active Learning for Neural Semantic Role Labeling on Low Resource Conversational Corpus," in *Workshop on Deep Learning Approaches for Low-Resource NLP*, Melbourne, Australia, 2018.

[13] A. authors, "ZERO-SHOT CROSS-LINGUAL CONVERSATIONAL SEMANTIC ROLE LABELING," in *the Annual meeting of the Association for Computational Linguistics ACL*, 2022.

[14] Ilseyar Alimova et al, "Cross-lingual transfer learning for semantic role labeling in Russian," *Proceeding of CLIB,* 2020.