Multi-Task Deep Learning for Legal Document Translation, Summarization and Multi-Label Classification

Ahmed Elnaggar, Christoph Gebendorfer, Ingo Glaser and Florian Matthes
Software Engineering for Business Information Systems, Technische Universität München
Garching bei München, Bavaria
{ahmed.elnaggar,christoph.gebendorfer,ingo.glaser,matthes}@tum.de

ABSTRACT

The digitalization of the legal domain has been ongoing for a couple of years. In that process, the application of different machine learning (ML) techniques is crucial. Tasks such as the classification of legal documents or contract clauses as well as the translation of those are highly relevant. On the other side, digitized documents are barely accessible in this field, particularly in Germany. Today, deep learning (DL) is one of the hot topics with many publications and various applications. Sometimes it provides results outperforming the human level. Hence this technique may be feasible for the legal domain as well. However, DL requires thousands of samples to provide decent results. A potential solution to this problem is multi-task DL to enable transfer learning. This approach may be able to overcome the data scarcity problem in the legal domain, specifically for the German language. We applied the state of the art multi-task model on three tasks: translation, summarization, and multi-label classification. The experiments were conducted on legal document corpora utilizing several task combinations as well as various model parameters. The goal was to find the optimal configuration for the tasks at hand within the legal domain. The multi-task DL approach outperformed the state of the art results in all three tasks. This opens a new direction to integrate DL technology more efficiently in the legal domain.

CCS CONCEPTS

• **Computing methodologies** → **Machine learning**; **Multi-task learning**; **Natural language processing**;

KEYWORDS

Multi-task Deep Learning; Translation; Summarization; Multi-la bel; Classifcation

1 INTRODUCTION

In the past few years, deep learning yielded to great results in many fields, including computer vision, natural language processing (NLP), speech recognition and robotics. In many areas, it was able to outperform humans including, image classification [6], health [17] and reading comprehension [25]. The availability of large amount of annotated data and fast computing power are the two main reasons behind this big hype. In the legal domain, legal professionals are doing a lot of tasks related to natural language processing daily, which could be replaced by ML algorithms, but that didn't happen yet deeply due to a scarcity of annotated data. Despite the fact that there are exceptionally large text bases in the legal domain, they are not preprocessed and structured in a format to be used seamlessly in ML technology. The use of ML in the legal domain gained momentum and some work has already been done, like translating legal documents [9] or classifying verdicts of the French Supreme Court [23]. However, a lot of possible use cases are not exploited yet.
Generating annotated datasets is generally a cost-intensive process. It is even more difficult in the legal domain, because we can't easily crowd source it. For example "image net" the biggest image classification dataset and "SQUAD" the biggest reading comprehension dataset, were created through Amazon Mechanical Turk by sourcing people without very specific knowledge. In the legal domain, you need people with very specific knowledge and education to annotate these unstructured data. This circumstance makes the creation of new datasets difficult to crowd source and even more expensive. This leads to a particular problem in the legal domain:

- NLP is highly required for the legal domain, but annotated datasets barely exist at all.

One way to overcome this problem is by using multi-task deep learning [18]. In this approach, we train multiple tasks using only one model to provide better results of these problems through transfer learning, especially, tasks that suffer from data scarcity. Therefore, in our work, we needed to achieve two goals:

(1) Investigate the effect of transfer learning in the selected legal problems.
(2) Find a big legal text dataset that can be used for transfer learning in any other legal task.

Furthermore, we want to answer three questions regarding the usage of the multi-task deep learning in the legal domain:

(1) Is transfer learning through multi-tasking beneficial for tasks in the legal domain?
(2) What are the results of training multiple problems jointly versus separately?
(3) Can the multi-task approach outperform the state of the art in the legal domain?

2 RELATED WORK

Deep learning is not yet been used intensively in the legal domain. Specifically, according to our knowledge, multi-task deep learning was not deeply investigated by researchers and has not been applied in the legal domain. However, we will try to cover the most related research to our work.

AICCC '18, December 21–23, 2018, Tokyo, Japan

ACM ISBN 978-1-4503-6623-6/18/12.
https://doi.org/10.1145/3299819.3299844

Translation: A. Vaswani [24] proposed the Transformer which represents the current state of the art in general translation, with a BLEU [16] score of 41.8. P.Koehn [9] built 462 machine translation systems for all language pairs of the Acquis Communautaire corpus, which is the body of common rights and obligations which have been adopted by all European Union (EU) Member States.

Summarization: AM. Rush [19] initiated work on abstractive summarization with neural networks and induced researchers to continue with sequence-to-sequence models. Additional variants were proposed after that for both extractive and abstractive summarization [12]. C. Grover [3] built the HOLJ corpus for extractive summarization of British judgments. B. Hachey [4] used machine learning for extractive summarization using a corpus of judgments of the UK House of Lords.

Classification: [21] Multi-label classification of legal document of the JRC-Acquis using the EuroVoc thesaurus [1, 15] is one of the difficult tasks, because it has more than 6000 labels and low number of samples per label. R. Steinberger [21] achieved a respectable accuracy of 47.3% on German and 48% on English documents of the JRC-Acquis involving the EuroVoc thesaurus.

Multi-Task: R. Collobert [2] built a unified multi-task architecture for various NLP tasks such as SRL, NER, POS, chunking and language modeling. They demonstrated that learning tasks simultaneously can improve performance, and they achieved state-of-the-art performance in SRL by training the SRL task jointly with a language model. X. Liu [14] successfully develop a multi-task DNN to combine tasks as disparate as classification and web page ranking. The experimental results demonstrate that the model consistently outperforms strong baselines. P. Liu [13] proposed three RNN based architectures to model text sequence with multi-task learning. They focused their work on four different text classification tasks about movie reviews. H. Zhang [26] proposed a multi-task learning architecture for text classification with four types of recurrent neural layers. Their model outperforms the single task models for various datasets consisting of product and movie reviews. L. Kaiser [7] took the next step of multi-task learning by combining tasks from different domains including image classification, image caption generation, text translation, text parsing and speech recognition. They showed that adding these tasks together never hurts performance and in most cases improves it on all tasks. They also showed that tasks with less data benefit largely from joint training with other tasks, while performance on large tasks degrades only slightly if at all.

3 LEGAL CORPORA

The three datasets that were used include the proceedings of the European Parliament (Europarl) [8], digital corpus of the European parliament (DCEP) [5] and Joint Research Centre - Acquis Communautaire (JRC-Acquis) [22].

The Europarl corpus provides the proceedings of the European Parliament between the years 1996 and 2011 for 20 languages. Usually, the documents cover the discussions of political topics. Therefore, sentences often contain first-person narrative text expressing political opinions and positions. The DCEP covers different areas including press releases, session protocols, reports of the parliamentary committees and written questions for 23 languages. The JRC-Acquis is a collection of legislative documents, retrieved from the European Union (EU) law, stating EU laws and policies for 22 languages, which have to be implemented by each member state.

Only seven major languages were selected for training as proof of concept including English, German, French, Italian, Spanish, Czech and Swedish. Furthermore, the three datasets were preprocessed from their original format to Moses format [10], which eases the integration of any machine learning platform or library.

Translation Dataset: The three datasets which were used for the translation are Europarl[1], DCEP[2] and JRC-Acquis[3]. Including only the previous mentioned 7 languages and 21 language pairs as shown on Table 1. The final combined translation dataset contains 4 to 8 million samples sentences per language pair. It is considered a good source for transfer learning, because summarization and multi-labeling datasets are only 0.5% and 0.3% of its size.

Summarization Dataset: The JRC-Acquis[4] dataset was used for summarization where each document contains a title element holding a short description of the document body. This summary usually varies between one to three sentences representing the semantic core of each document. The dataset contains between 18k to 22k samples per each language. Table 2 shows the number of samples for training and test datasets for each language.

Multi-Labeling Dataset: The JRC-Acquis[5] dataset was used for multi-labeling, where each document is assigned with multiple EuroVoc labels. These labels originate from the EuroVoc thesaurus, a hierarchical structure of legal topics divided in more than 6000 classes, e.g. covering agriculture, food, health, information technology, law or politics. Between one and seven classes are usually assigned to each document. The dataset contains between 11k to 14k samples per language. Table 2 shows the number of samples for training and test datasets for each language.

4 MULTI-TASK LEGAL SYSTEM

The algorithm we used for multi-task learning is the MultiModel algorithm. The algorithm was proposed by the Google Brain Team [7] to create a single generalized deep learning model which is capable of solving tasks across multiple areas (natural language processing, computer vision and speech recognition). This single model was originally trained concurrently on general tasks including image classification, image captioning generation, language translation, English parsing task and speech recognition. However, in our work we used the algorithm for language translation, summarization and document classification specifically in the legal domain.

4.1 MultiModel Architecture

The MultiModel consists of four parts facilitating multi-task learning across multiple different areas of application. The core is based on a

[1] https://mediatum.ub.tum.de/1446650
[2] https://mediatum.ub.tum.de/1446648
[3] https://mediatum.ub.tum.de/1446655
[4] https://mediatum.ub.tum.de/1446654
[5] https://mediatum.ub.tum.de/1446653

	legal-europarl		legal-dcep		legal-jrc-acquis		Combined	
Task	Train	Test	Train	Test	Train	Test	Train	Test
cs-de	554785	12877	3322863	26365	956206	7926	4833854	47168
cs-en	632331	13475	3429669	26023	954139	7731	5016139	47229
cs-es	605198	13293	3331565	24771	965210	8003	4901973	46067
cs-fr	614135	12797	3353646	27185	961109	7979	4928890	47961
cs-it	593176	12533	3427214	25216	956954	7939	4977344	45688
cs-sv	617190	13088	3353962	24759	912310	7447	4883462	45294
de-en	1920519	39310	5389867	47179	1227338	9800	8537724	96289
de-es	1848928	38030	5244580	46939	1234976	10129	8328484	95098
de-fr	1904020	37733	5353860	50266	1239724	10100	8497604	98099
de-it	1794869	37183	5399213	47908	1230221	10098	8424303	95189
de-sv	1803663	37445	5223641	45001	1150149	9416	8177453	91862
en-es	1968689	39069	5782727	50650	1231178	9916	8982594	99635
en-fr	2011292	38370	5730964	52431	1237570	9916	8979826	100717
en-it	1907616	37237	5617352	45408	1222257	9874	8747225	92519
en-sv	1854436	37007	5684684	48163	1144536	9300	8683656	94470
es-fr	1944351	37359	5507250	50581	1244162	10234	8695763	98174
es-it	1843115	36557	5458678	46341	1242123	10229	8543916	93127
es-sv	1789877	35696	5539768	46588	1155662	9382	8485307	91666
fr-it	1906101	36276	5558798	49214	1234846	10196	8699745	95686
fr-sv	1842905	36220	5321630	45969	1157695	9423	8322230	91612
it-sv	1730666	34133	5402000	43877	1155158	9354	8287824	87364
Total	**31687862**	**635688**	**102433931**	**870834**	**23813523**	**194392**	**157935316**	**1700914**

Table 1: Number of translation samples in training and test sets of the legal translation corpora (legal-dcep, legal-europarl, legal-jrc-acquis)

	legal-jrc-acquis-summarize		legal-jrc-acquis-label	
Task	Train	Test	Train	Test
cs	17956	264	12571	253
de	22707	327	14153	295
en	22448	328	14391	302
es	22751	327	14065	296
fr	22586	326	14147	297
it	22371	322	14086	293
sv	19255	265	11561	236
Total	**150074**	**2159**	**94974**	**1972**

Table 2: Number of samples in training and test sets of the legal summarization and labeling corpus

fully convolutional sequence-to-sequence approach which includes three actors (encoder, decoder, mixer). For the purpose of multi-task learning, the MultiModel uses so-called modality nets to cope with input from different application domains. These four building blocks are shown on figure 1 and are briefly presented [7] bellow.

Modality Nets: There are four different modality nets available for the MultiModel (language, image, audio, categorical). These allow it to accept and produce different input and output types. The language, image and audio modality nets are responsible for converting input data into a variable-size joint representation which is fed into the encoder. On the other side, the language and categorical modality nets are used to transfer the variable-size joint output of the decoder into the expected output format. Different tasks with the same input or output format share a modality net in order to promote generalization and allow the quick addition of further tasks. In our case, only the language modality was used.

Encoder: The encoder takes the unified embeddings of the input modality nets and processes it with six custom-built convolutional blocks with one mixture-of-expert layer in between.

Decoder: The decoder takes the encoded input data (encoder) and encoded output data (mixer) from previous steps to generate variable-size decoded outputs for an output modality. It consists of four convolutional attention blocks with one mixture of expert layer in the middle. Furthermore, at the beginning of each decoding run a command token for the current task is passed to it. This way the decoder learns to produce decoded outputs of different tasks for the same modality net.

I/O Mixer: The mixer takes the output of the encoder and the previous output from the decoder. Through this autoregressive scheme it is highly capable to learn long term dependencies. The mixer consists of two convolutional blocks and one attention block.

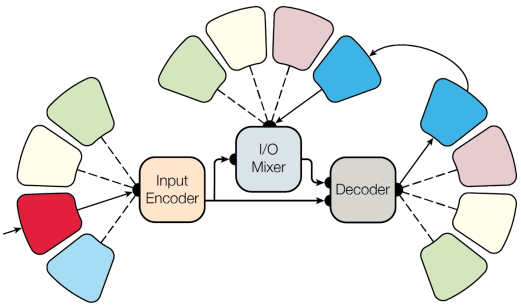

Figure 1: Google Multi-Model Building Blocks

5 EXPERIMENTAL SETTINGS

5.1 Training Details

Generally, every model was trained until it converged, and sometimes we used early stop to prevent overfitting. For the MultiModel, we have used two configurations. First, the base (MM-B) configuration as described in the paper and second, a light version (MM-L) configuration. The light version has fewer parameters and was used to test the effect of number of parameters on the result. The Transformer, MultiModel base and MultiModel light were trained with a batch size of 4096, 2048 and 1024, while the hidden size of each layer was 512, 512 and 128, and the filter size was 2048, 2084 and 1024. In case of the multi-task models, each training step comprised one batch of input data of one of the selected problems. To speed up the process, we trained the algorithms on four machines. The Transformer model on a machine with 4x Tesla K80, while the MultiModel base version was trained on two machines. The first one was NVIDIA DGX-1 with 8x Tesla V100 and the second one was with 5x Titan XP. The MultiModel light version was trained on a machine with 4x GTX 1080 Ti.

Different combinations of the joint tasks have been tested. For translation, we choose two combinations, a pool combination (jt-pool-5) which consists of the five available German translation pairs "de-en, de-es, de-fr, de-it, de-sv", and a chain combination (jt-chain-7) which consists of a chain of language "cs-de, de-en, en-es, es-fr, fr-it, it-sv". For summarization, we investigated one combination (js-7) which joins all the summarization languages. For multi-labeling, we selected one combination (jl-7) which joins all the multi-labeling languages. Finally, we included an overall combination which comprises tasks across task families with the same language (ja-3). It combines the translation (de-en), summarization and multi-labeling task of the German language together. All of these combinations of the MultiModel were compared with the result of the state of the art models, which is the Transformer for general translation and summarization, and JEX for JRC-Acquis multi-label classification. Finally, due to the time and the number of pages constrains, we only report the result of the German language.

5.2 Metrics

We report our results with common task-dependent metrics. In the following sections, we cover each task metrics.

5.2.1 Translation. The BLEU [16] score was used to evaluate the translation results. It measures the quality of the translation based in the n-grams overlaps between the predicted translation and the target translation.

$$BLEU = min1, \frac{hypothesis\_length}{reference\_length} \prod_{i=1}^{4} precision_i^{\frac{1}{4}} \quad (1)$$

5.2.2 Summarization. A standard metric for evaluating summarization tasks is the ROUGE [11] metric, which we use to evaluate the summaries. We only evaluated the results based on 1-gram, 2-grams and the longest n-gram. For the sake of convenience, we call them ROUGE-1, ROUGE-2 and ROUGE-L.

$$ROUGE_N = \frac{\sum_{S \in reference\_summaries} \sum_{gram_n \in S} count\_matchgram_n}{\sum_{S \in reference\_summaries} \sum_{gram_n \in S} countgram_n} \quad (2)$$

5.2.3 Multi-Label Classification. For multi-label classification, we report precision, recall and F1 score.

$$Precision = \frac{\#TruePositive}{\#TruePositive + \#FalsePositive} \quad (3)$$

$$Recall = \frac{\#TruePositive}{\#TruePositive + \#FalseNegative} \quad (4)$$

$$F_1 = \frac{Precision \times Recall}{Precision + Recall} \quad (5)$$

6 RESULT AND DISCUSSION

Figure 2 shows the translation results. Generally, all models had better results on both dcep and jrc-acquis datasets than the europarl. This might be because these two datasets contain a larger portion of cross references, sentence fragments and enumerations compared to the europarl.

The MultiModel light version (MM-L single, MM-L jt-pool-5, MM-L jt-chain-7 and MML- ja-3) falls behind both the Transformer model (TF-B single) and the MultiModel base version on the three datasets. This is because the number of parameters of the light version is almost half the number of parameters to its competitors. The light version usually produces shorter sentences, however, after manually examining them, we found that semantic meaning remains largely untouched. Another important observation is that the MultiModel light trained on single tasks generally outperforms itself trained on joint tasks. This originates in the limited capacity of this version which prevented it to attach to multiple tasks. Further increasing the number of tasks yielded stepwise decreasing BLEU scores.

The MultiModel base version (MM-B single) outperformed the Transformer model for both dcep and jrc-acquis datasets with BLEU score 54.98 and 67.24 compared to 53.3 and 64.22. However, in the case of the europarl dataset, the BLEU score was a little bit less, 37.15 compared to 37.34. When the model was trained jointly with other translation languages (MM-B jt-pool-5 and MM-B ja-chain-7) the BLEU score falls behind the Transformer.

However, the model (MM-B ja-3) which was trained across task families (translation, summarization and classification) with the same input language (German), outperformed all other models in the dcep dataset with BLEU score 55.11. On the jrc-acquis dataset,

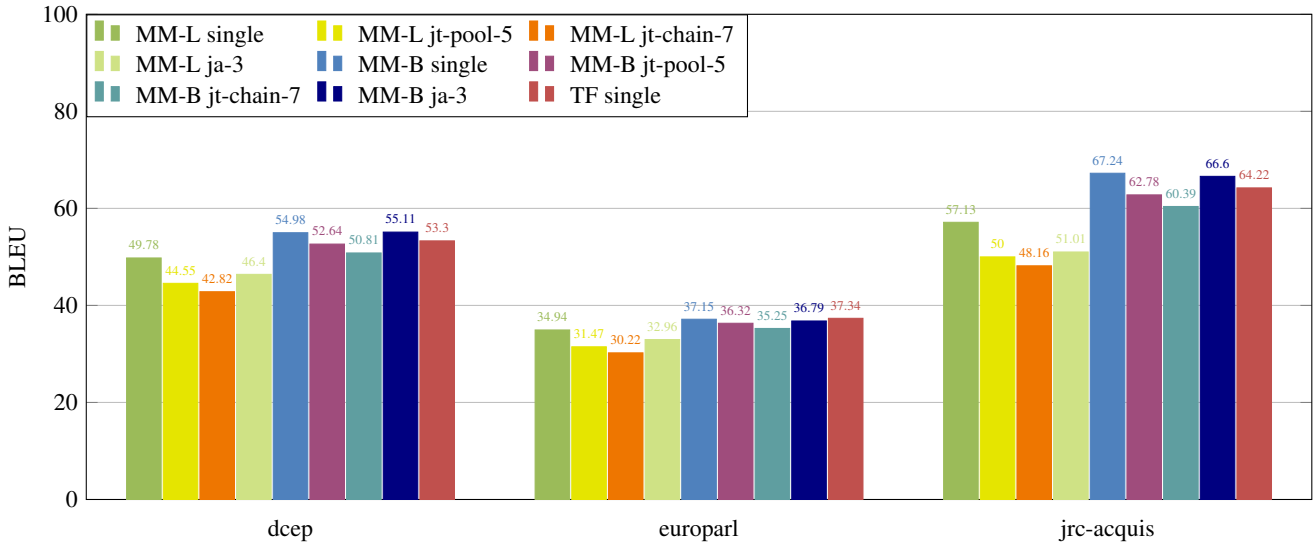

Figure 2: German-to-English translation BLEU score performance for all single-task & multi-task translation combinations trained on the MultiModel Light (MM-L), MultiModel Base (MM-B) and Transformer (TF)

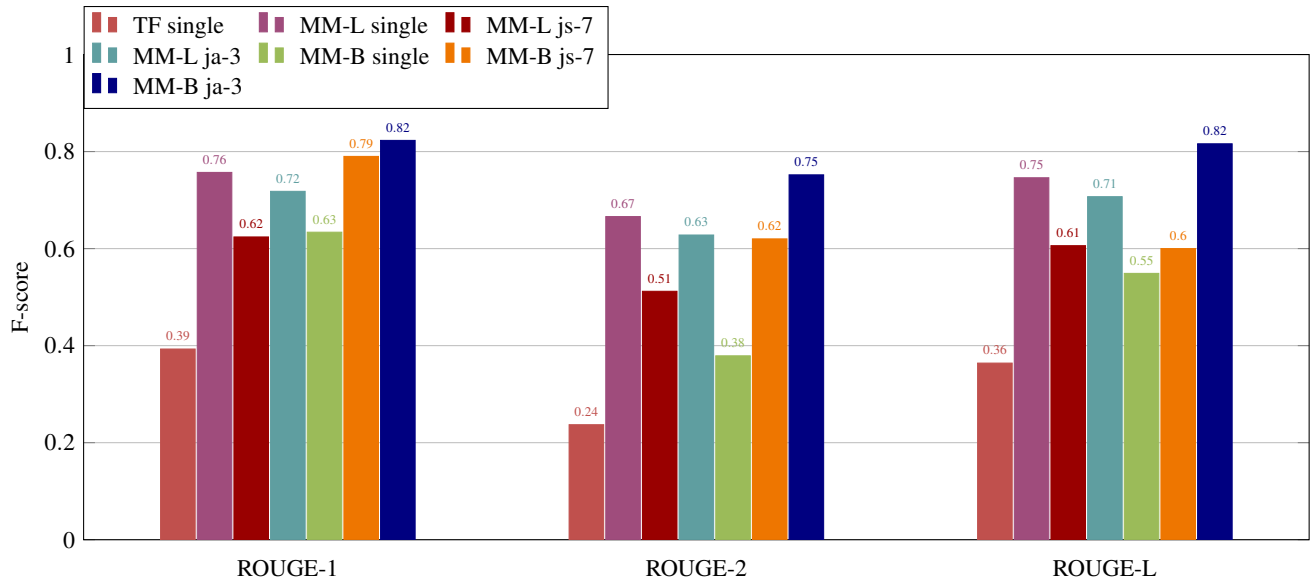

Figure 3: German Summarization performance using ROUGE score for all single-task & multi-task summarization combinations trained on the MultiModel Light (MM-L), MultiModel Base (MM-B) and Transformer (TF)

it was better than the Transformer, but worse than the MultiModel which was trained on a single task (MM-B single) with BLEU score 66.6. For the europarl dataset, it slightly falls behind both the Transformer and the MultiModel base single task. Table 3 shows predicted translation samples from the different models.

Figure 3 shows the summarization results. The Transformer model falls behind the MultiModel for both the light and base versions. It had almost 50% less ROUGE points compared to the MultiModel. The MultiModel light versions had almost always better ROUGE scores than the MultiModel base when it was trained on either single German summarization or multi-language summarization. We see the reason in the size of the dataset which was small relative to the number of parameters for the base model, which lead to fast overfitting even with using regularization techniques. The best ROUGE scores were obtained from the MultiModel (MM-B ja-3) which was trained on the three different tasks jointly with the same input language (German), with ROUGE-1, ROUGE-2 and ROUGE-L of 0.82, 0.75 and 0.82.

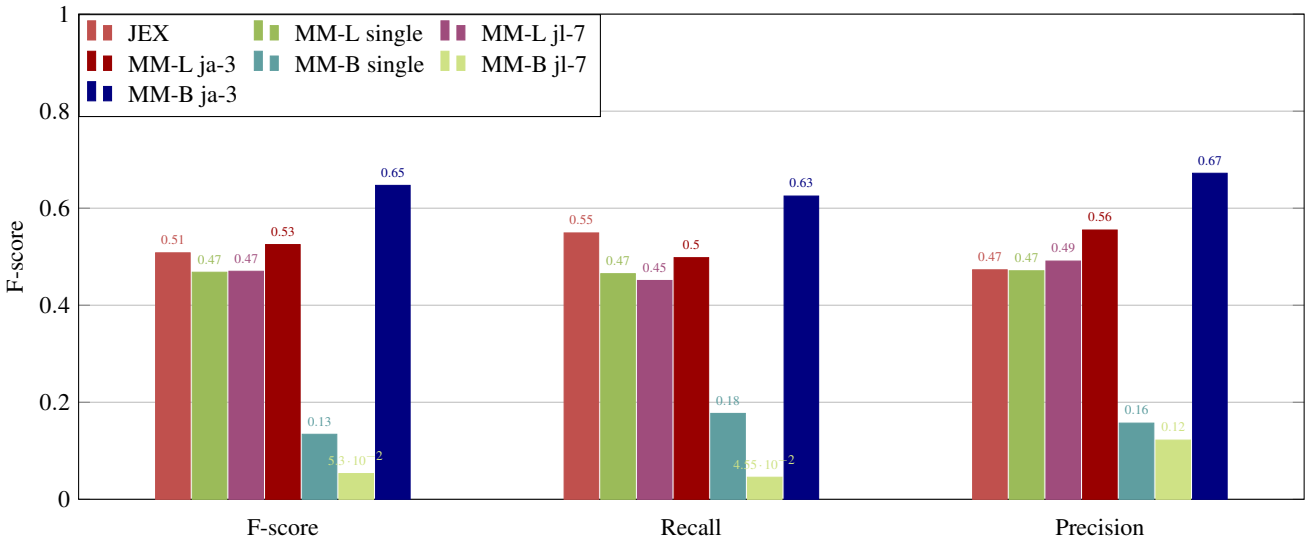

Figure 4: German multi-label classification performance using F-score, Recall and Precision scores for all single-task & multi-task classification combinations trained on the MultiModel Light (MM-L), MultiModel Base (MM-B) and JEX [20]

	BLEU	Example
Input	-	9 . Argentinien gewährleistet die Einhaltung dieser Vereinbarung insbesondere dadurch , daß es innerhalb der in dieser Vereinbarung festgelegten Mengen Ausfuhrlizenzen für die unter Nummer 1 genannten Erzeugnisse erteilt .
MM-L single	17.61	9. Argentina shall ensure compliance with this Agreement by granting the export licences referred to in point 1 within the quantities laid down in this Agreement.
MM-L jt-pool-5	27.57	9. Argentina shall ensure compliance with this Agreement, in particular by providing for export licences for the products referred to in point 1 within the quantities set out in this Agreement.
MM-L jt-chain-7	34.01	9. Argentina shall ensure compliance with this Agreement, in particular by granting export licences for products referred to in paragraph 1 within the quantities set out in this Agreement.
MM-L ja-3	25.72	9. Argentina shall ensure compliance with this Agreement in particular by granting export licences for the products referred to in point 1 within the quantities laid down in this Agreement.
MM-B single	29.63	9. Argentina shall ensure compliance with this Agreement, in particular by issuing export licences for the products referred to in point 1 within the quantities specified in this Agreement.
MM-B jt-pool-5	30.02	9. Argentina shall ensure compliance with this Agreement, in particular by granting it export licences for the products referred to in point 1 within the limits laid down in this Agreement.
MM-B jt-chain-7	29.71	9. Argentina shall ensure compliance with this Agreement, in particular by issuing export licences for the products referred to in point 1 within the quantities set out in this Agreement.
MM-B ja-3	50.62	9. Argentina shall ensure compliance with this Agreement, in particular by issuing export licences for the products referred to in paragraph 1 within the limits of the quantities fixed in this Agreement.
TF single	40.09	9. Argentina shall ensure compliance with this Agreement in particular by issuing export licences for the products referred to in point 1 within the limits of the quantities laid down in this Agreement.
Reference	-	9. Argentina shall ensure that this arrangement is observed, in particular, by issuing export certificates covering the products referred to in paragraph 1 within the limits of the quantities covered by this arrangement.

Table 3: German-to-English translation examples of the jrc-acquis for all trained combinations

Figure 4 shows the multi-label classification results. The JEX model outperformed the MultiModel light (MM-L single, MM-L jl-7 and ML-L ja-3) on both F-score and recall, against having lower precision. The MultiModel base which was trained on single (MM-B single) and all classification languages (MM-B jl-7) performs not well enough to provide any good classification results. The reason lies in the multi-labeling dataset, which is very small compared to the model capacity, which made the model to overfit quickly. The best result that outperformed JEX, the state of the art model, was

obtained by combining the translation, summarization and multi-label classification task within the same language (MM-B ja-3) with F-score, recall and precision of 0.65, 0.63 and 0.67 compared to 0.51, 0.55 and 0.47.

The previous experiments lead to three important points, which answer the three research questions. First, multi-task deep learning outperforms the single task state of the art models, when it is combined with different tasks of the same input language and one of these tasks has a large number of samples. This allows to transfer the knowledge the algorithm learns between these tasks. Second, the greater the number of tasks in a joint task the greater is the impact on performance compared to the relatedness or diversity of the joined tasks. Third, the capacity of multi-task models must be adopted depending on dataset sizes [6].

7 CONCLUSIONS & FUTURE WORK

We proved that multi-task deep learning can be useful in the legal domain. Of course, the type, the amount of joined tasks and the capacity of the multi-task model are major influential factors on the result. However, it is an effective approach to solve the data scarcity problem through transfer learning. Using this approach makes it possible to outperform current state of the art results, and also facilitates the application of deep learning technology in the legal domain. Our work is a base for further research on the effectiveness and usage of multi-task in the legal domain. However, more experiments are required to test it on other tasks, datasets, languages and training combinations. The provided datasets could be used to test the approach on the remaining languages.

8 ACKNOWLEDGEMENTS

We gratefully acknowledge the support of Leibniz-Rechenzentrum, Microsoft Corporation and NVIDIA Corporation for the hardware which was used for this research.

REFERENCES

[1] Guido Boella, Luigi Di Caro, Daniele Rispoli, and Livio Robaldo. 2013. A system for classifying multi-label text into EuroVoc. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Law*. ACM, 239–240.

[2] Ronan Collobert and Jason Weston. 2008. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th international conference on Machine learning*. ACM, 160–167.

[3] Claire Grover, Ben Hachey, and Ian Hughson. 2004. The HOLJ Corpus. Supporting Summarisation of Legal Texts. In *Proceedings of the 5th International Workshop on Linguistically Interpreted Corpora*.

[4] Ben Hachey and Claire Grover. 2005. Automatic legal text summarisation: experiments with summary structuring. In *Proceedings of the 10th international conference on Artificial intelligence and law*. ACM, 75–84.

[5] Najeh Hajlaoui, David Kolovratnik, Jaakko Väyrynen, Ralf Steinberger, and Daniel Varga. 2014. DCEP-Digital Corpus of the European Parliament.. In *LREC*. 3164–3171.

[6] K He, X Zhang, S Ren, and J Sun. 2017. Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification. 2015: 1026-1034.

[7] Lukasz Kaiser, Aidan N Gomez, Noam Shazeer, Ashish Vaswani, Niki Parmar, Llion Jones, and Jakob Uszkoreit. 2017. One model to learn them all. *arXiv preprint arXiv:1706.05137* (2017).

[8] Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *MT summit*, Vol. 5. 79–86.

[9] Philipp Koehn, Alexandra Birch, and Ralf Steinberger. 2009. 462 machine translation systems for europe. *Proceedings of MT Summit XII* (2009), 65–72.

[10] Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, et al. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th annual meeting of the ACL on interactive poster and demonstration sessions*. Association for Computational Linguistics, 177–180.

[11] Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*. 74–81.

[12] Linqing Liu, Yao Lu, Min Yang, Qiang Qu, Jia Zhu, and Hongyan Li. 2017. Generative Adversarial Network for Abstractive Text Summarization. *arXiv preprint arXiv:1711.09357* (2017).

[13] Pengfei Liu, Xipeng Qiu, and Xuanjing Huang. 2016. Recurrent neural network for text classification with multi-task learning. *arXiv preprint arXiv:1605.05101* (2016).

[14] Xiaodong Liu, Jianfeng Gao, Xiaodong He, Li Deng, Kevin Duh, and Ye-Yi Wang. 2015. Representation learning using multi-task deep neural networks for semantic classification and information retrieval. (2015).

[15] Eneldo Loza Mencía and Johannes Fürnkranz. 2010. Efficient multilabel classification algorithms for large-scale problems in the legal domain. In *Semantic Processing of Legal Texts*. Springer, 192–215.

[16] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*. Association for Computational Linguistics, 311–318.

[17] Pranav Rajpurkar, Jeremy Irvin, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Ding, Aarti Bagul, Curtis Langlotz, Katie Shpanskaya, et al. 2017. Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. *arXiv preprint arXiv:1711.05225* (2017).

[18] Sebastian Ruder. 2017. An overview of multi-task learning in deep neural networks. *arXiv preprint arXiv:1706.05098* (2017).

[19] Alexander M Rush, Sumit Chopra, and Jason Weston. 2015. A neural attention model for abstractive sentence summarization. *arXiv preprint arXiv:1509.00685* (2015).

[20] Ralf Steinberger, Mohamed Ebrahim, and Marco Turchi. 2013. JRC EuroVoc Indexer JEX - A freely available multi-label categorisation tool. *CoRR* abs/1309.5223 (2013). arXiv:1309.5223 http://arxiv.org/abs/1309.5223

[21] Ralf Steinberger, Mohamed Ebrahim, and Marco Turchi. 2013. JRC EuroVoc Indexer JEX-A freely available multi-label categorisation tool. *arXiv preprint arXiv:1309.5223* (2013).

[22] Ralf Steinberger, Bruno Pouliquen, Anna Widiger, Camelia Ignat, Tomaz Erjavec, Dan Tufis, and Dániel Varga. 2006. The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. *arXiv preprint cs/0609058* (2006).

[23] Octavia-Maria Şulea, Marcos Zampieri, Shervin Malmasi, Mihaela Vela, Liviu P Dinu, and Josef van Genabith. 2017. Exploring the use of text classification in the legal domain. (2017).

[24] 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 *Advances in Neural Information Processing Systems*. 5998–6008.

[25] Adams Wei Yu, David Dohan, Minh-Thang Luong, Rui Zhao, Kai Chen, Mohammad Norouzi, and Quoc V Le. 2018. QANet: Combining Local Convolution with Global Self-Attention for Reading Comprehension. *arXiv preprint arXiv:1804.09541* (2018).

[26] Honglun Zhang, Liqiang Xiao, Yongkun Wang, and Yaohui Jin. 2017. A generalized recurrent neural architecture for text classification with multi-task learning. *arXiv preprint arXiv:1707.02892* (2017).

[6]. The output of the translation, summarization and classification tasks with the different models can be downloaded from 1^0, 2^0 and 3^0 in the decodes folder.