Language Modelling as a Multi-Task Problem

Weber, L.; Jumelet, J.; Bruni, E.; Hupkes, D.

DOI
10.18653/v1/2021.eacl-main.176

Publication date
2021

Document Version
Final published version

Published in
The 16th Conference of the European Chapter of the Association for Computational Linguistics

Language Modelling as a Multi-Task Problem

Lucas Weber
DTCL, University Pompeu Fabra
lucas.weber@upf.edu

Jaap Jumelet
ILLC, University of Amsterdam
j.w.d.jumelet@uva.nl

Elia Bruni
IKW, University of Osnabrück
elia.bruni@gmail.com

Dieuwke Hupkes
Facebook AI Research
dieuwkehupkes@fb.com

Abstract

In this paper, we propose to study language modelling as a multi-task problem, bringing together three strands of research: multi-task learning, linguistics, and interpretability. Based on hypotheses derived from linguistic theory, we investigate whether language models adhere to learning principles of multi-task learning during training. To showcase the idea, we analyse the generalisation behaviour of language models as they learn the linguistic concept of Negative Polarity Items (NPIs). Our experiments demonstrate that a multi-task setting naturally emerges *within* the objective of the more general task of language modelling. We argue that this insight is valuable for multi-task learning, linguistics and interpretability research and can lead to exciting new findings in all three domains.

1 Introduction

Humans are optimising their behaviour towards a multitude of objectives to reach their goals in day-to-day life. By learning many things at the same time and exploiting their commonalities, they acquire more general knowledge about the world, which in turn helps them to learn new things quicker (Perkins et al., 1992; Schwartz et al., 2005; Cormier and Hagman, 2014; Luriia, 1976). This idea of finding more general solutions through the diversification of tasks has found its way also to the machine learning community, in the field of multi-task learning (MTL) (Caruana, 1993, 1997). In MTL, multiple tasks are optimised jointly, enabling the transfer of relevant information across tasks. MTL research yields fruitful results in both application (e.g. Collobert and Weston, 2008; Collobert et al., 2011; Zhang et al., 2014; Donahue et al., 2014; Kaiser et al., 2017) and theory (e.g. Baxter, 2000; Maurer, 2006; Ando and Zhang, 2005; Argyriou et al., 2008).

However, deciding on a setup requires making many arbitrary choices. The researcher or engineer has to decide which tasks to train together (e.g. Bingel and Søgaard, 2017; Standley et al., 2020); at which hierarchy-level to allow tasks to interact (e.g. Søgaard and Goldberg, 2016); which degree of parameter sharing to employ (Ruder, 2017); which distribution of training data to employ (e.g. Luong et al., 2016), and so on. Having to make so many arbitrary choices is inconvenient for modellers, but also stands in the way of understanding the learning principles of neural models in multi-task settings. The highly constructed learning scenarios make it difficult to see whether outcomes should be attributed to one of the many a-priori decisions or to inherent properties of the learning process.

In this paper, we propose to study MTL not in a constructed, artificial scenario, but in a more natural setting. To do so, we consider the objective of *language modelling* and exploit the fact that it can be seen as a conglomerate of many different tasks. To give an example: rules of word ordering have to be learned simultaneously to rules of feature agreement and the monotonicity properties of different linguistic environments. These different tasks all need to be learned to achieve the greater goal of producing acceptable sentences, and they have to be optimised in parallel when the language model (LM) is trained. Language modelling is in that sense a *natural* multi-task learning problem with a naturally given *task hierarchy* provided by linguistic theory (see also Figure 1).

Studying language modelling as a multi-task problem has several distinct advantages. From an MTL perspective, it gives us a complete hierarchy of relevant tasks that can freely interact throughout the learning process, unconstrained by prior assumptions. We can make theoretically informed decisions about these tasks, drawing on linguistic theory. We can also deduce from linguistics how these tasks relate to each other (or, in other words, how similar they are), which in MTL is considered to be one of the crucial factors for the learn-

Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics, pages 2049–2060
April 19 - 23, 2021.

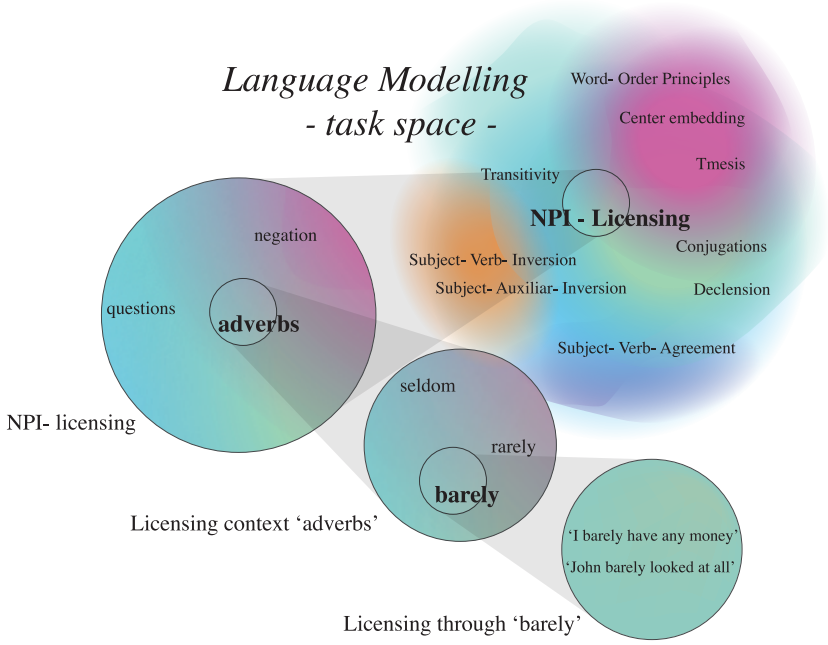

Figure 1: A conceptual visualisation of a language modelling task hierarchy, from language modelling as a whole to single examples, with complex similarities between tasks. Colours indicate task similarities.

ing outcomes (e.g. Thrun and O'Sullivan, 1996; Passos et al., 2012). MTL has not yet been studied from this dynamic and unconstrained perspective. Then, somewhat more delicately, the extent to which models can exploit similarities hypothesised by linguistic theory can play a role in confirming or refuting specific linguistic hypotheses. Lastly, when it comes to interpretability research, applying concepts from MTL can be valuable to better understand the learning dynamics of models. By understanding how models are finding solutions, we can infer what these solutions are.

Outline In the remainder of this paper, we will first provide some basic background about MTL (§ 2.1), the subset of linguistic tasks we focus on (*Negative Polarity Items*, where we consider their different *licensing contexts* as tasks, § 2.2) and discuss some related work in interpretability (§ 2.3). Then, in § 3 and § 4, respectively, we present our approach and empirical results that showcase our idea. In § 5, we discuss our results and framework in the light of the three fields mentioned before. We conclude in § 6.

2 Background

In this paper, we aim to bring together three strands of research: MTL, linguistics and interpretability research. As a proof of concept, we focus on one specific complex subset of linguistic tasks: *licensing of Negative Polarity Items (NPIs)*. Below, we give a short overview of the most important characteristics of the three fields of interest.

2.1 Multi-task learning

In MTL, multiple tasks are learned together to enable information transfer from one task to another. If the transfer is successful, the benefits might be threefold: the model learns tasks with less training data (i.e. *more efficient*, Collobert et al., 2011; Benton et al., 2017; Kaiser et al., 2017), up to a higher final accuracy (Collobert and Weston, 2008; Kaiser et al., 2017) and in a way that better generalises to new tasks (Baxter, 2000; Collobert and Weston, 2008).

Caruana (1993, 1997) and Ruder (2017) propose several different – but related – processes that might enable positive transfer: related tasks can provide additional training examples for each other on the features they share (*statistical data amplification*), certain features might be easier to learn through one task than through another, but be useful for both of them (*eavesdropping*), and idiosyncratic features of single tasks can be averaged out, while more general features are reinforced (*attention focusing*)[1].

[1]For a complete list of processes please consult the original publications.

However, positive transfer is not guaranteed; It is also possible that performance *deteriorates* due to interference between different tasks, resulting in negative transfer, (Rosenstein et al., 2005; Pan and Yang, 2010; Wang et al., 2019). Whether transfer is positive depends on the *task similarity* and whether the model is able to exploit this similarity (Rosenstein et al., 2005; Thrun and O'Sullivan, 1996; Passos et al., 2012).

The main goal of MTL so far has been to avoid negative- and promote positive transfer by determining task-similarity and regulate the interactions between tasks based on these similarities. Due to its pivotal role, much research effort was spent on determining similarities of tasks and the regulation of information transfer between them (for an overview, see Zhang and Yang, 2017; Ruder, 2017). The disadvantage of these approaches is that assuming fixed tasks and regulating transfer between them based on fixed task-similarities puts large constraints on possible transfers between tasks, because it neglects the fact that learning processes are dynamic. From the perspective of the model, tasks, as well as their similarities, can change throughout the learning process. Here, we only use predefined tasks and their similarities to *analyse* the learning behaviour of the model, without constraining the learning process in any way.

2.2 Negative Polarity Items

We exemplify our idea by analysing the learning behaviour on a complex subset of linguistic tasks: the licensing of Negative Polarity Items (NPIs). The properties of NPI licensing make it an interesting and adequate subset of tasks to study, as it has a high degree of complexity, has an appropriate frequency within natural language and was previously frequently investigated in neural models.

NPIs are characterised by the property that they can only occur within the scope of certain *licensing contexts*. For instance, in the example below, the NPI '**any**' can occur in sentence (1)a., where it is in the scope of a negation, but not in sentence (1)b., where there is no licensor present.

(1) a. Bill did*n't* buy **any** books that day.
b. * Bill did buy **any** books that day.

(2) a. *Nobody* has **ever** been there.
b. * Somebody has **ever** been there.

Licensing contexts are formed on the basis of semantic properties, such as downward entailment (Fauconnier, 1975; Ladusaw, 1980), non-veridicality (Giannakidou, 2011), or scope marking (Barker, 2018). Common licensing contexts include negation, conditionals, or superlatives, and are often *triggered* by a specific expression, such as 'not' or 'nobody'.

Grasping the phenomenon of NPI licensing requires understanding of three different aspects:

1. *The class of NPIs*: there is a group of expressions that are restricted in their occurrence.
2. *Licensing contexts*: there exists a group of expressions that allow NPIs to occur.
3. *Scope and structure*: the licensing contexts have to stand in a certain structural relationship to the NPIs.

We focus on how LMs learn the second aspect by analysing how different types of licensing contexts interact and generalize throughout training. During learning they should be able to exploit their similarity in the other two aspects.

2.3 Interpretability

Interpretability research on LMs has shown that in pre-trained models, such as BERT (Devlin et al., 2019), hierarchical structure emerges throughout the layers and that this structure demonstrates parallels with linguistic theory (Peters et al., 2018; Liu et al., 2019; Tenney et al., 2019). However, the emergence of this structure has not been explicitly connected to MTL yet.

In recent years, research has shown that LMs are able to understand NPI licensing. Jumelet and Hupkes (2018) evaluate the performance of LMs on data sets containing NPI constructions extracted from large corpora, and Marvin and Linzen (2018); Wilcox et al. (2019); Warstadt and Bowman (2020) test them on artificial data sets containing template-based NPI constructions. In our own experimental setup we will utilise the extensive template-based NPI corpus of Warstadt et al. (2019).

What these approaches have in common is their focus on the performance of pretrained LMs. Our MTL approach sheds light on an unexplored aspect of NPI understanding: the learning dynamics of the model *during* training.

3 Approach

We consider two different types of experiments. First, to understand to which extent models can

understand and use the similarity between different licensing contexts (our *tasks*) during learning, we exploit the effect that frequency of the different contexts has on learning. Second, we manipulate the LMs' training corpus to constrain their ability to leverage information from other licensing contexts during learning. In accordance with the MTL-literature, we expect the LMs to learn tasks more data-efficient and to a higher final accuracy if they can leverage information across contexts. Before we describe our experiments in more detail, we present our model architecture and training, the evaluation procedure of the licensing contexts, and the filter procedure we use to manipulate the training corpus.

3.1 Model

Following previous work in this area, we consider recurrent language models. We focus on unidirectional LSTM models and mirror the hyperparameter setup of Gulordava et al. (2018)[2]. We train the models on the corpus provided by the same authors[3] – a subset of the English Wikipedia – or modified versions of the same for our second experiment (see § 4.2). To track the learning process, we save models every 100 batches of training (371 model-checkpoints per epoch). For all experiments, we average performance across five random seeds.

3.2 Evaluation

To estimate the LMs' understanding of NPIs and their dependence on the different licensing contexts, we adapt the Cloze task of Warstadt et al. (2019), based on the implementation of Jumelet (2020). This task considers nine different types of licensing contexts (a list of the contexts, including examples, can be found in Table 1). For every such context, Warstadt et al. (2019) generated a large number of *minimal pair sentences*, containing correctly and incorrectly licensed NPIs. For instance, for the *adverbs* licensing context:

(3) a. A lady *rarely* **ever** thought that the children saw the boy.

b. * A lady *sometimes* **ever** thought that the children saw the boy.

Following previous work, we quantify an LM's understanding of a particular type of licensing context by computing the percentage of minimal pairs in that context for which the model correctly assigns a higher probability to the NPI in the licensing contexts than in the non-licensing contexts. I.e., in the example above, we would compare the probability the model assigns to the word *ever* in the contexts "A lady rarely" and "A lady sometimes" (see also Figure 2).

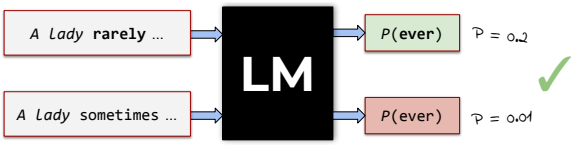

Figure 2: The NPI judgement task that is used for evaluating the LMs. A correct prediction assigns a higher probability to an NPI in a context that licenses it, based on the corpus of Warstadt et al. (2019).

3.3 Identification of NPIs in training corpus

The Warstadt et al. (2019) corpus provides us with a task to evaluate nine different context types that license NPIs. To manipulate the training corpus for our experiments we also need to identify sentences in the training corpus of the model in which these contexts actually licence NPIs. To do so, we need to locate these contexts, as well as establish that they in fact licence an NPI in a particular sentence.

We consider the nine Warstadt et al. context types, and the corresponding list of 30 expressions that are part of these contexts (e.g. the list of adverbs licensing NPIs). As for the NPIs, we consider an extensive list of 160 distinct NPIs[4], based on the collection provided by Hoeksema (2012). We then identify sentences in which an element of our NPI list is preceded by an element from our context list, ensuring that there is a dependency relation between them using the dependency parser of spaCy (Honnibal and Johnson, 2015). When there are multiple potential licensors in a sentence, we use the hierarchical distance between the licensor and the NPI in the parse tree as a heuristic to find the correct licensor. By testing this procedure on a manually labeled set of 200 randomly selected sentences with multiple licensors, we estimate that it identifies the correct among multiple licensors in around 97% of cases. In Table 1, we report examples and frequencies of the different licensing contexts in the training corpus based on this filtering scheme.

[2]Hyperparameters: batch size = 64, BPTT length = 35, dropout = 0.1, adaptive SGD learning rate = 20, layers = 2, hidden and embedding size = 650, epochs = 40.

[3]https://github.com/facebookresearch/colorlessgreenRNNs/tree/master/data

[4]This list can be found in Appendix A.

Context	Example	**Frequency** per 100k sentences
Simple Questions	Did he **ever** do a mean thing*?*	10
Adverbs	In the present political culture, there are *hardly* **any** leaders who would avoid limelight and refuse positions of power.	23
Questions	However, various writers attribute it to Putnam, Stark, Prescott or Gridley, while others question *whether* it was said **at all** .	25
Superlative	[...] and caused the *worst* winter flooding **in decades** for river and stream valleys [...].	32
Only	[...] "Those [students] *only* are supposed to pay **anything** who are abundantly able, or prefer to do so.	85
Conditional	In 1997 Li published a paper attempting to replicate <unk>'s results and showed the effect was very small, *if* it existed **at all**.	127
Quantifier	That's *all* you'll **ever** need.	179
Determiner negation	In spite of the <unk> of the disaster, *no* one was **ever** held accountable.	218
Sentential negation	It is *not* judged under **any** subjective points of view, only the clock.	712

Table 1: The nine types of licensing contexts taken from Warstadt et al. (2019), with an example and the context frequency within the training corpus.

4 Experiments and results

As a first step, we assess whether the LMs can adequately represent all nine categories of the evaluation task. To do so, we train five models on the regular training corpus, and compute their final accuracy on our nine tasks. All models show adequate performance on most contexts (see Table 2), with the exception of the simple question context. Additionally, we observe that the models achieve their accuracy surprisingly fast: already after two epochs, there are no more substantial changes in empirical error (see Figure 3). In the rest of our experiments, we therefore focus only on these first two epochs.

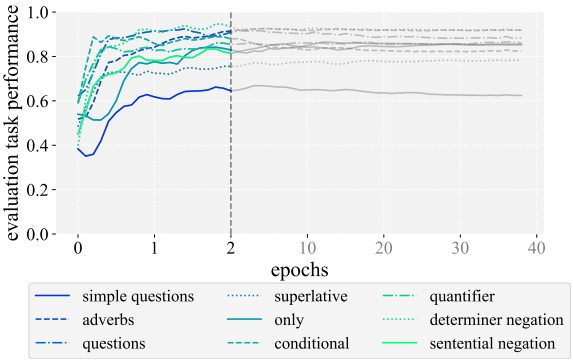

Figure 3: Average evaluation task performance. The performance rises steeply during approximately the first 2 epochs of training and afterwards levels off.

Context	**Accuracy** ± std
Simple Questions	0.62 ± 0.05
Adverbs	0.92 ± 0.01
Questions	0.88 ± 0.03
Superlative	0.78 ± 0.03
Only	0.86 ± 0.04
Conditional	0.82 ± 0.06
Quantifier	0.86 ± 0.04
Determiner negation	0.92 ± 0.05
Sentential negation	0.85 ± 0.03

Table 2: Performance of the LMs on the evaluation task after 40 epochs of training, averaged over 5 runs.

4.1 Frequency vs data efficiency

While some licensing contexts are rather common (e.g. negation), others appear scarcely as a licensor (e.g. adverbs). Therefore, throughout the learning process, the LMs encounter many instances of the more frequent contexts before they see an example of an infrequent context. If LMs were able to leverage information across contexts, less frequent contexts should thus have more prior established NPI-understanding that they can bootstrap from. Consequently, the LMs should require fewer training examples to learn less frequent contexts than they need to learn more frequent contexts. In other words, the LM should be more *data efficient* for these infrequent contexts.

In our first experiment, we use this hypothesised relationship between frequency and data efficiency to assess whether LMs can exploit the similarities between different licensing contexts. To be able to compare across different contexts, we quantify the data efficiency of an LM for a particular context as

the number of examples the LM needs to observe until it reaches 95% of its final accuracy for that context.[5] To make this measure more robust, we first apply a Savitzky–Golay noise-filter to the learning curve (degree of polynomial = 1, window size = 25; Savitzky and Golay 1964).

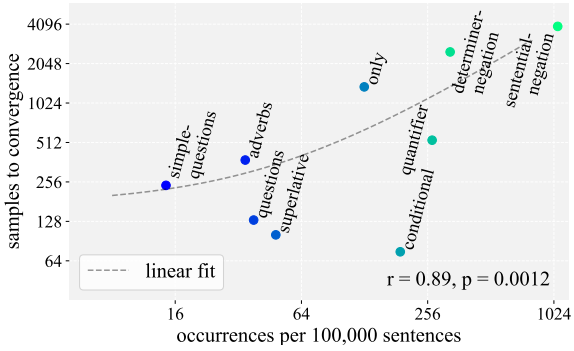

Figure 4: Data efficiency of nine different licensing contexts plotted against their frequency, averaged over five runs. The data efficiency is quantified as number of training examples the model needs to observe to achieve 95% of the trained-out performance.

We compute the data efficiency of the trained LMs for all nine contexts and compute the correlation between a context's frequency and the model's data efficiency with respect to that context. In Figure 4, we plot the average data efficiency of each context against the frequency of that context, as well as the linear fit that relates these two variables. The experiment demonstrates a strong relationship between the data efficiency and frequency of a respective context: $r = .89$, $p < .05$. Hence, the less frequent a licensing context is, the fewer examples are needed for the model to learn it, from which we conclude that the model is indeed able to transfer knowledge from previously acquired knowledge.

4.2 Transfer from general knowledge

While the presented relationship between frequency and data efficiency demonstrates that LMs can leverage previously learned information to learn less frequent licensing contexts, it does not unequivocally show that it leverages information from *other NPI contexts*. After all, when a less frequent context is encountered, the LM has not only had the opportunity to acquire prior knowledge about NPIs, it has also simply seen more language in general. In other words, the LM may meanwhile also have acquired more *general language knowledge*, which may help it to more quickly learn a less frequent licensing context. In our second experiment, we isolate transfer from general language knowledge and transfer from previously observed NPIs by training LMs on *single-context* corpora.

Single-context corpora *Single-context corpora* contain NPIs licensed only by a single context. LMs trained on these corpora can thus not transfer knowledge acquired from other licensing contexts, as these are not present in the training data. By comparing the data efficiency of contexts between LMs trained on all-context and single-context corpora, we can thus infer how much of the increase of data efficiency for lower-frequent contexts is due to leveraging information from other contexts.

To create our nine single-context corpora, we use the procedure described in § 3.3 to identify all sentences containing NPIs licensed by our nine contexts. For every context, we then create a corpus in which all sentences containing other contexts licensing NPIs are replaced by a neutral sentence of the same length, sampled from the rest of the corpus. During this replacement procedure, the ordering and composition of the corpus remained otherwise intact.

When we compare the learning of single-context with all-context models, we cannot rely on the previously used data-efficiency metric from Experiment 4.1. The data-efficiency measure is bound to how quickly the model reaches its final accuracy and accordingly benefits when its final accuracy decreases. As we expect the final accuracy to be lower in the single context models, comparing only data-efficiencies between models is likely to be uninformative.[6]. In this experiment, as explained below, we instead consider the area between the curves (AbC).

Area between Curves (AbC) *Area between Curves* (AbC) incorporates both data efficiency and accuracy: for every context, we calculate the area between the all-contexts and single-context learning curves until the point in time where they both have reached 95% of their final accuracy. The larger this area is, the more impactful it is to remove

[5] The *more* data efficient, the *lower* this number thus is.

[6] Consider, for instance, the extreme case in which an LM does not learn a particular context at all anymore in the single-context condition, as indicated by a chance accuracy of 0.5. Because it is not learning anything, the model would arrive at its maximum accuracy before having seen any examples, resulting in a data efficiency of 0.

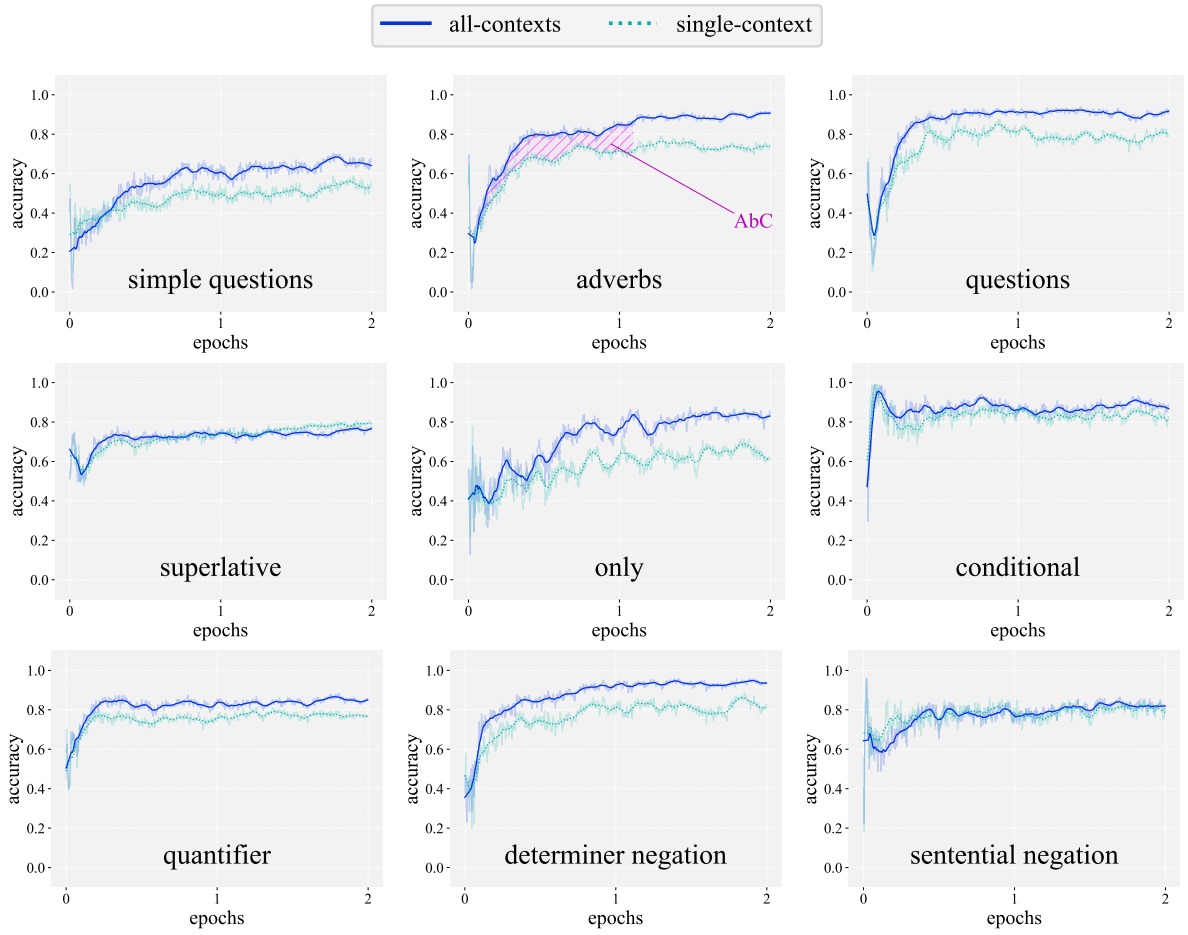

Figure 5: The LMs performance on different licensing contexts for the first two epochs of training. We obtained these curves by evaluating all models at all 730 training-checkpoints on the evaluation task.

all other NPI contexts, and the more the model leveraged from these contexts. The learning curves of all contexts, along with an illustration of the AbC-measure, can be found in Figure 5.

As a first interesting observation, we see that for seven of the nine contexts, the all-contexts model learns faster and achieves higher final performance.[7] Both frequent and infrequent contexts thus benefit from information acquired by other licensing contexts, in terms of both data-efficiency and final accuracy.

This positive transfer can also be seen in Figure 6, where we plot the AbC for all licensing contexts against their frequency. This plot also confirms the relationship found in our previous experiments: the less frequent a context is, the more it benefits from other NPIs ($r = .76$, $p < .05$).

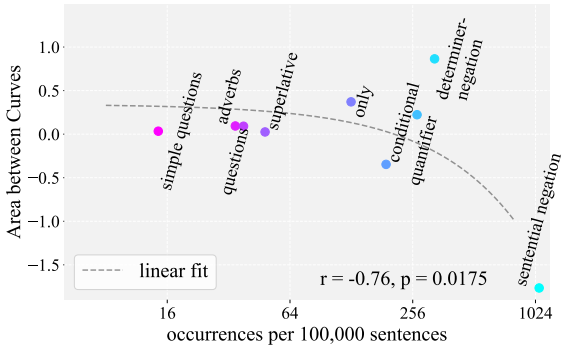

Figure 6: Normalised AbC for all licensing contexts until convergence of both contexts to 95% accuracy. AbC > 0 indicates a better performance of the all-context model and vice versa.

5 Discussion

In this paper, we studied language modelling as a multi-task problem. We show that neural language models can find and exploit similarity between the different language construction rules that we deduced from linguistic theory and that their transfer

[7] A one-sided Welch's test confirms that the calculated AbCs are overall different from zero: $t = 2.61$, $p < .05$.

behaviour mirrors the generalisation behaviour in traditionally constructed MTL settings. In this section, we now reflect on how our setup and results contribute to the three different areas that we mentioned in the introduction: MTL, linguistics and interpretability research.

5.1 Multi-task learning research

Studying LMs as multi-task learners, we observe several phenomena known from traditional MTL: when trained in parallel, similar (sub)tasks are learned more efficiently (compare Collobert et al., 2011; Kaiser et al., 2017), and with higher accuracy (Collobert and Weston, 2008; Kaiser et al., 2017), and this effect is stronger for less frequent tasks (Benton et al., 2017; Kaiser et al., 2017).

Our study differs in one crucial aspect from previous research on MTL: it looks at learning dynamics *within* one, larger, natural task instead of between tasks defined by the modeller. As a consequence, the learning process itself is not constrained through a priori decisions concerning task selection, or how tasks should be optimised together. In our scenario, contrary to traditional MTL, we use tasks and their hypothesised similarity only to *analyse* the learning process of the language model, not to inform its training. As such, our natural setting allows to study traditional MTL phenomena, such as data amplification, eavesdropping, and attention focusing, independent of arbitrary decisions regarding task selection and optimisation. This knowledge can then be transferred to scenarios in which more control over the selection of tasks may be required.

5.2 Interpretability research

A second field where we believe studying language models as multi-task learners can contribute, is the field of interpretability. On a more basic level, our paper confirms previous findings in interpretability that LMs are able to adequately model NPIs (Jumelet and Hupkes, 2018; Wilcox et al., 2019; Marvin and Linzen, 2018). We add to this literature by *explicitly* showing that LMs are connecting different types of contexts together through their learning behaviour. Contrary to previous work, we are tapping the learning process itself as a source of information to better understand the inner workings of these models.

Traditional concepts from MTL, such as the earlier mentioned explanations of Caruana (1993) and Ruder (2017) (§ 2.1) are valuable to better understanding what models are learning and how. For instance, when we observe that the solution of models improves when more varied NPI material is presented (our single- versus all-context experiment), MTL can aid to formulate concrete hypotheses about *why* this is the case. This, in turn, can help us improve our understanding of the solutions that are learned by the model. For instance, we find that the single-context models usually level-off on a lower accuracy-level than the all-context model (see Figure 5). This is not merely explainable by the amount of data, as we continue to add training examples in either case. The difference between models instead appears to be due to the variety of the training data. The idea of *attention focusing* (Caruana, 1993, 1997; Ruder, 2017) helps us to understand what is going on: by being trained on more varied NPI material, the model can better sort out which features are relevant and which ones are instead idiosyncrasies correlated with specific contexts. Such hypotheses can then help inform further experiments, that investigate – for example – which features specifically are better learned through attention focusing.

5.3 Linguistics research

Finally, we believe that studying language models as multi-task learners can also contribute to the field of linguistics. In our study, we show that LMs can find and exploit similarities between linguistically defined concepts. Turning things around, this generalisation behaviour of models can also be seen as a confirmation of the linguistic task hierarchy that we assumed from the start. The language modelling objective is unconstrained by linguistic theory and therefore does not necessarily have to find the same solutions as linguistics. Similarity derived from the learning behaviour of language models might therefore be used as a tool to work on more disputed ideas in linguistics and to form new hypotheses in linguistic theory. While the linguistic insights that can be drawn from the current study are relatively limited, they do provide a proof of concept for future work: we show that domain knowledge and learning behaviour of neural models can be connected.

6 Conclusion

In the current study we explored the possibility to use multi-task learning as a framework to study learning behaviour *within* a task. To this end we

considered LMs as multi-task learners and investigated how they learn the task-cluster of NPI-licensing. We find that LMs pick up on similarities that we assume from linguistic theory and exploit them to learn similar language constructions with less data and to a higher accuracy. Especially less frequent tasks benefit from this effect.

These results resemble positive transfer in 'traditional' MTL. We lined out the possible benefits that our study may have for MTL research, interpretability and linguistics. From here there are many directions for future work: targeting less comprehensively researched areas in linguistics to add empirical data to otherwise usually theoretical linguistic discussions, investigating the change of internal representations in place of the behavioural measure used here to more precisely describe the learning process, or applying the approach to other high-level tasks in other modalities obeying other knowledge domains are just few of theses possibilities.

Acknowledgments

We thank the anonymous reviewers; the COLT-group at UPF for the discussions and their useful feedback; the participants of the EviL-seminars, and especially Emmanuel Dupoux, for inspiration; and Laura Castro Moreno for her help with the graphic designs. Further, LW thanks the Department of Translation and Language Sciences at the University Pompeu Fabra for funding.

References

Rie Kubota Ando and Tong Zhang. 2005. A framework for learning predictive structures from multiple tasks and unlabeled data. *Journal of Machine Learning Research*, 6(Nov):1817–1853.

Andreas Argyriou, Massimiliano Pontil, Yiming Ying, and Charles A Micchelli. 2008. A spectral regularization framework for multi-task structure learning. In *Advances in neural information processing systems*, pages 25–32.

Chris Barker. 2018. Negative polarity as scope marking. *Linguistics and Philosophy*, pages 1–28.

Jonathan Baxter. 2000. A model of inductive bias learning. *Journal of artificial intelligence research*, 12:149–198.

Adrian Benton, Margaret Mitchell, and Dirk Hovy. 2017. Multitask learning for mental health conditions with limited social media data. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 152–162, Valencia, Spain. Association for Computational Linguistics.

Joachim Bingel and Anders Søgaard. 2017. Identifying beneficial task relations for multi-task learning in deep neural networks. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 164–169, Valencia, Spain. Association for Computational Linguistics.

Rich Caruana. 1997. Multitask Learning. *Machine Learning*, 28(1):41–75.

Richard Caruana. 1993. Multitask learning: A knowledge-based source of inductive bias. In *Proceedings of the Tenth International Conference on Machine Learning*, pages 41–48. Morgan Kaufmann.

Ronan Collobert and Jason Weston. 2008. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *ICML '08: Proceedings of the 25th international conference on Machine Learning*, pages 160–167.

Ronan Collobert, Jason Weston, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural Language Processing (Almost) from Scratch. *Journal of Machine Learning Research*, 12:2493–2537.

Stephen M Cormier and Joseph D Hagman. 2014. *Transfer of learning: Contemporary research and applications*. Academic Press.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Jeff Donahue, Yangqing Jia, Oriol Vinyals, Judy Hoffman, Ning Zhang, Eric Tzeng, and Trevor Darrell. 2014. Decaf: A deep convolutional activation feature for generic visual recognition. In *International conference on machine learning*, pages 647–655.

Gilles Fauconnier. 1975. Polarity and the scale principle. *Chicago Linguistics Society*, 11:188–199.

Anastasia Giannakidou. 2011. Negative and positive polarity items: Variation, licensing, and compositionality. *Semantics: An International Handbook of Natural Language Meaning*, pages 1660–1712.

Kristina Gulordava, Piotr Bojanowski, Edouard Grave, Tal Linzen, and Marco Baroni. 2018. Colorless green recurrent networks dream hierarchically. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1195–1205, New

Orleans, Louisiana. Association for Computational Linguistics.

Jack Hoeksema. 2012. On the Natural History of Negative Polarity Items Syntax View project Morphology View project. *Linguistic Analysis*, 44(2):3–3–3.

Matthew Honnibal and Mark Johnson. 2015. An improved non-monotonic transition system for dependency parsing. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1373–1378, Lisbon, Portugal. Association for Computational Linguistics.

Jaap Jumelet. 2020. diagNNose: A library for neural activation analysis. In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 342–350, Online. Association for Computational Linguistics.

Jaap Jumelet and Dieuwke Hupkes. 2018. Do language models understand anything? on the ability of LSTMs to understand negative polarity items. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 222–231, Brussels, Belgium. Association for Computational Linguistics.

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*.

William A. Ladusaw. 1980. *Polarity Sensitivity as Inherent Scope Relations*. Ph.D. thesis, University of Texas, Austin.

Nelson F Liu, Matt Gardner, Yonatan Belinkov, Matthew E Peters, and Noah A Smith. 2019. Linguistic knowledge and transferability of contextual representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1073–1094.

Minh-Thang Luong, Quoc V. Le, Ilya Sutskever, Oriol Vinyals, and Lukasz Kaiser. 2016. Multi-task sequence to sequence learning. In *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*.

Aleksandr Romanovich Luriia. 1976. *Cognitive development: Its cultural and social foundations*. Harvard university press.

Rebecca Marvin and Tal Linzen. 2018. Targeted syntactic evaluation of language models. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1192–1202, Brussels, Belgium. Association for Computational Linguistics.

Andreas Maurer. 2006. Bounds for linear multi-task learning. *Journal of Machine Learning Research*, 7(Jan):117–139.

Sinno J. Pan and Qiang Yang. 2010. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359.

Alexandre Passos, Piyush Rai, Jacques Wainer, and Hal Daume. 2012. Flexible modeling of latent task structures in multitask learning. In *Proceedings of the 29th International Conference on Machine Learning, ICML 2012*, volume 2, page 1283–1290.

David N Perkins, Gavriel Salomon, et al. 1992. Transfer of learning. *International encyclopedia of education*, 2:6452–6457.

Matthew Peters, Mark Neumann, Luke Zettlemoyer, and Wen-tau Yih. 2018. Dissecting contextual word embeddings: Architecture and representation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1499–1509.

Michael T. Rosenstein, Zvika Marx, Leslie Pack Kaelbling, and Thomas G. Dietterich. 2005. To transfer or not to transfer. In *In NIPS'05 Workshop, Inductive Transfer: 10 Years Later*.

Sebastian Ruder. 2017. An overview of multi-task learning in deep neural networks. *CoRR*, abs/1706.05098.

Abraham Savitzky and Marcel J.E. Golay. 1964. Smoothing and Differentiation of Data by Simplified Least Squares Procedures. *Analytical Chemistry*, 36(8):1627–1639.

Daniel L Schwartz, John D Bransford, David Sears, et al. 2005. Efficiency and innovation in transfer. *Transfer of learning from a modern multidisciplinary perspective*, 3:1–51.

Anders Søgaard and Yoav Goldberg. 2016. Deep multi-task learning with low level tasks supervised at lower layers. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 231–235, Berlin, Germany. Association for Computational Linguistics.

Trevor Standley, Amir Zamir, Dawn Chen, Leonidas Guibas, Jitendra Malik, and Silvio Savarese. 2020. Which tasks should be learned together in multi-task learning? In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 9120–9132. PMLR.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601, Florence, Italy. Association for Computational Linguistics.

Sebastian Thrun and Joseph O'Sullivan. 1996. Discovering structure in multiple learning tasks: The tc algorithm. In *Proceedings of the Thirteenth International Conference on Machine Learning*, pages 489–497. Morgan Kaufmann.

Zirui Wang, Zihang Dai, Barnabás Póczos, and Jaime Carbonell. 2019. Characterizing and avoiding negative transfer. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11285–11294.

Alex Warstadt and Samuel R Bowman. 2020. Can neural networks acquire a structural bias from raw linguistic data? *arXiv preprint arXiv:2007.06761*.

Alex Warstadt, Yu Cao, Ioana Grosu, Wei Peng, Hagen Blix, Yining Nie, Anna Alsop, Shikha Bordia, Haokun Liu, Alicia Parrish, Sheng-Fu Wang, Jason Phang, Anhad Mohananey, Phu Mon Htut, Paloma Jeretic, and Samuel R. Bowman. 2019. Investigating BERT's knowledge of language: Five analysis methods with NPIs. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2877–2887, Hong Kong, China. Association for Computational Linguistics.

Ethan Wilcox, Peng Qian, Richard Futrell, Miguel Ballesteros, and Roger Levy. 2019. Structural Supervision Improves Learning of Non-Local Grammatical Dependencies. In *Proceedings of North American Association for Computational Linguistics (NAACL)*, pages 3302–3312.

Yu Zhang and Qiang Yang. 2017. A survey on multi-task learning. *CoRR*, abs/1707.08114.

Zhanpeng Zhang, Ping Luo, Chen Change Loy, and Xiaoou Tang. 2014. Facial landmark detection by deep multi-task learning. In *Computer Vision – ECCV 2014*, pages 94–108, Cham. Springer International Publishing.

A List of NPIs

We here present the full list of 160 NPIs that has been used for modifying the corpora:

- a bed of roses
- a care in the world
- a chance in hell
- a damn
- a damn thing
- a day goes by
- a day over
- a ghost of a
- a hair out of place
- a living soul
- a moment of your time
- a moment too soon
- a shadow of a doubt
- a single soul
- all that much
- all that many
- any
- any longer
- any old
- any time soon
- anybody
- anymore
- anyone
- anything
- anything like
- anytime soon
- anywhere
- anywhere close
- anywhere near
- as of yet
- as yet
- at all
- avail
- bat an eye
- be any time
- be anything like
- beat around the bush
- by a long sho
- by any chance
- by any means
- by any stretch
- by miles
- by much
- can be bothered
- can compare to
- can hold a candle to
- can make of
- can possibly
- chance in hell
- come at a worse time
- come cheap
- could care less
- could possibly
- cut the mustard
- even once
- ever
- far wrong
- for much longer
- for shit
- for the life of
- for the soul of
- give a crap
- give a damn
- give a fuck
- give a shit
- half a chance
- half bad
- have a clue
- have any of
- hold a candle to
- hold water
- in a blue moon
- in a hundred years
- in a long time
- in a million years
- in ages
- in all of history
- in any
- in any manner
- in any way
- in centuries
- in days
- in decades
- in his right mind
- in hours
- in living memory
- in minutes
- in months
- in recent memory
- in the least
- in the least bit
- in the slightest
- in weeks
- in years
- just any
- just yet
- know the first thing
- know the first thing about
- know the half of it
- least of all
- let alone
- lift a finger
- make a sound
- make head or tail of
- make much difference
- mean a thing
- mean feat
- miss a beat
- much care
- much help
- much of a
- much of anything
- much to look at
- much to lose
- nor
- on speaking terms
- on your life
- one single thing
- or anything
- rhyme or reason
- say much
- see eye to eye
- set foot
- set foot in
- set foot on
- sit right with
- sit well
- sit well with
- small feat
- so much as
- square with
- squat
- stand a chance
- strong suit
- such thing
- sweat it
- take his eyes off
- take kindly to
- take lightly
- take no for an answer
- that many
- that much
- that often
- the ghost of
- the half of
- the half of it
- the least bit
- the like of which
- the likes of which
- the slightest
- the slightest bit
- think much of
- to be taken lightly
- whatever
- whatsoever
- with a barge pole
- worth a damn
- worth his salt
- worth its salt
- yet