Will my auxiliary tagging task help? Estimating Auxiliary Tasks Effectivity in Multi-Task Learning

Johannes Bjerva
University of Groningen
The Netherlands
j.bjerva@rug.nl

Abstract

Multitask learning often improves system performance for morphosyntactic and semantic tagging tasks. However, the question of *when* and *why* this is the case has yet to be answered satisfactorily. Although previous work has hypothesised that this is linked to the label distributions of the auxiliary task, we argue that this is not sufficient. We show that information-theoretic measures which consider the joint label distributions of the main and auxiliary tasks offer far more explanatory value. Our findings are empirically supported by experiments for morphosyntactic tasks on 39 languages, and are in line with findings in the literature for several semantic tasks.

1 Introduction

When attempting to solve a natural language processing (NLP) task, one can consider the fact that many such tasks are highly related to one another. A common way of taking advantage of this is to apply multitask learning (MTL, Caruana (1998)). MTL has been successfully applied to many linguistic sequence-prediction tasks, both syntactic and semantic in nature (Collobert and Weston, 2008; Cheng et al., 2015; Søgaard and Goldberg, 2016; Martínez Alonso and Plank, 2016; Bjerva et al., 2016; Ammar et al., 2016; Plank et al., 2016). It is, however, unclear *when* an auxiliary task is useful, although previous work has provided some insights (Caruana, 1998; Martínez Alonso and Plank, 2016).

Currently, considerable time and effort need to be employed in order to experimentally investigate the usefulness of any given main task / auxiliary task combination. In this paper we wish to alleviate this process by providing a means to investigating *when* an auxiliary task is helpful, thus also shedding light on *why* this is the case. Concretely, we apply information-theoretic measures to a collection of data- and tag sets, investigate correlations between such measures and auxiliary task effectivity, and show that previous hypotheses do not sufficiently explain this interaction. We investigate this both experimentally on a collection of syntactically oriented tasks on 39 languages, and verify our findings by investigating results found in the literature on semantically oriented tasks.

2 Neural Multitask Learning

Recurrent Neural Networks (RNNs) are at the core of many current approaches to sequence prediction in NLP (Elman, 1990). A bidirectional RNN is an extension which incorporates both preceding and proceeding contexts in the learning process (Graves and Schmidhuber, 2005). Recent approaches frequently use either (bi-)LSTMs (Long Short-Term Memory) or (bi-)GRUs (Gated Recurrent Unit), which have the advantage that they can deal with longer input sequences (Hochreiter and Schmidhuber, 1997; Chung et al., 2014).

The intuition behind MTL is to improve performance by taking advantage of the fact that related tasks will benefit from similar internal representations (Caruana, 1998). MTL is commonly framed such that all hidden layers are shared, whereas there is one output layer per task. An RNN can thus be trained to solve one *main* task (e.g. parsing), while also learning some other *auxiliary* task (e.g. POS tagging).

3 Information-theoretic Measures

We wish to give an information-theoretic perspective on when an auxiliary task will be useful for a given main task. For this purpose, we introduce some common information-theoretic measures which will be used throughout this work.[1]

[1] See Cover and Thomas (2012) for an in-depth overview.

Proceedings of the 21st Nordic Conference of Computational Linguistics, pages 216–220,
Gothenburg, Sweden, 23-24 May 2017.

The **entropy** of a probability distribution is a measure of its unpredictability. That is to say, high entropy indicates a uniformly distributed tag set, while low entropy indicates a more skewed distribution. Formally, the entropy of a tag set can be defined as

$$H(X) = -\sum_{x \in X} p(x) \log p(x), \quad (1)$$

where x is a given tag in tag set X.

It may be more informative to take the joint probabilities of the main and auxiliary tag sets in question into account, for instance using **conditional entropy**. Formally, the conditional entropy of a distribution Y given the distribution X is defined as

$$H(Y|X) = \sum_{x \in X} \sum_{y \in Y} p(x,y) \log \frac{p(x)}{p(x,y)}, \quad (2)$$

where x and y are all variables in the given distributions, $p(x,y)$ is the joint probability of variable x cooccurring with variable y, and $p(x)$ is the probability of variable x occurring at all. That is to say, if the auxiliary tag of a word is known, this is highly informative when deciding what the main tag should be.

The **mutual information** (MI) of two tag sets is a measure of the amount of information that is obtained of one tag set, given the other tag set. MI can be defined as

$$I(X;Y) = \sum_{x \in X} \sum_{y \in Y} p(x,y) \log \frac{p(x,y)}{p(x)\,p(y)}, \quad (3)$$

where x and y are all variables in the given distributions, $p(x,y)$ is the joint probability of variable x cooccurring with variable y, and $p(x)$ is the probability of variable x occurring at all. MI describes how much information is shared between X and Y, and can therefore be considered a measure of 'correlation' between tag sets.

3.1 Information Theory and MTL

Entropy has in the literature been hypothesised to be related to the usefulness of an auxiliary task (Martínez Alonso and Plank, 2016). We argue that this explanation is not entirely sufficient. Take, for instance, two tag sets X and X', applied to the same corpus and containing the same tags. Consider the case where the annotations differ in that the labels in every sentence using X' have been randomly reordered. The tag distributions in X and X' do not change as a result of this operation, hence their entropies will be the same. However, the tags in X' are now likely to have a very low correspondence with any sort of natural language signal, hence X' is highly unlikely to be a useful auxiliary task for X. Measures taking joint probabilities into account will capture this lack of correlation between X and X'. In this work we show that measures such as conditional entropy and MI are much more informative for the effectivity of an auxiliary task than entropy.

4 Data

For our syntactic experiments, we use the Universal Dependencies (UD) treebanks on 39 out of the 40 languages found in version 1.3 (Nivre et al., 2016).[2] We experiment with POS tagging as a main task, and various dependency relation classification tasks as auxiliary tasks. We also investigate whether our hypothesis fits with recent results in the literature, by applying our information-theoretic measures to the semantically oriented tasks in Martínez Alonso and Plank (2016), as well as the semantic tagging task in Bjerva et al. (2016).

Although calculation of joint probabilities requires jointly labelled data, this issue can be bypassed without losing much accuracy. Assuming that (at least) one of the tasks under consideration can be completed automatically with high accuracy, we find that the estimates of joint probabilities are very close to actual joint probabilities on gold standard data. In this work, we estimate joint probabilities by tagging the auxiliary task data sets with a state-of-the-art POS tagger.

4.1 Morphosyntactic Tasks

Dependency Relation Classification is the task of predicting the dependency tag (and its direction) for a given token. This is a task that has not received much attention, although it has been shown to be a useful feature for parsing (Ouchi et al., 2014). We choose to look at several instantiations of this task, as it allows for a controlled setup under a number of conditions for MTL, and since data is available for a large number of typologically varied languages.

Previous work has suggested various possible instantiations of dependency relation classification labels (Ouchi et al., 2016). In this work, we use labels designed to range from highly complex and informative, to very basic ones.[3] The labelling schemes used are shown in Table 1.

[2] Japanese was excluded due to treebank unavailability.

[3] Labels are automatically derived from UD.

Category	Directionality	Example	H
Full	Full	nmod:poss/R_L	3.77
Full	Simple	nmod:poss/R	3.35
Simple	Full	nmod/R_L	3.00
Simple	None	nmod	2.03
None	Full	R_L	1.54
None	Simple	R	0.72

Table 1: Dependency relation labels used in this work, with entropy in bytes (H) measured on English. The labels differ in the granularity and/or inclusion of the category and/or directionality.

The systems in the syntactic experiments are trained on main task data ($\mathbb{D}_{main}$), and on auxiliary task data ($\mathbb{D}_{aux}$). Generally, the amount of overlap between such pairs of data sets differs, and can roughly be divided into three categories: i) identity; ii) overlap; and iii) disjoint (no overlap between data sets). To ensure that we cover several possible experimental situations, we experiment using all three categories. We generate ($\mathbb{D}_{main}$, $\mathbb{D}_{aux}$) pairs by splitting each UD training set into three portions. The first and second portions always contain POS labels. In the identity condition, the second portion contains dependency relations. In the overlap condition, the second and final portions contain dependency relations. In the disjoint condition, the final portion contains dependency relations.

4.2 Semantic Tasks

Martínez Alonso and Plank (2016) experiment with using, i.a., POS tagging as an auxiliary task, with main tasks based on several semantically oriented tasks: Frame detection/identification, NER, supersense annotation and MPQA. Bjerva et al. (2016) investigate using a semantic tagging task as an auxiliary task for POS tagging. We do not train systems for these data sets. Rather, we directly investigate whether changes in accuracy with the main/auxiliary tasks used in these papers are correctly predicted by any of the information-theoretic measures under consideration here.

5 Method

5.1 Architecture and Hyperparameters

We apply a deep neural network with the exact same settings in each syntactic experiment. Our system consists of a two layer deep bi-GRU (100 dimensions per layer), taking an embedded word representation (64 dimensions) as input. We apply dropout ($p = 0.4$) between each layer in our network (Srivastava et al., 2014). The output of the final bi-GRU layer, is connected to two output layers – one per task. Both tasks are always weighted equally. Optimisation is done using the Adam algorithm (Kingma and Ba, 2014), with the categorical cross-entropy loss function. We use a batch size of 100 sentences, training over a maximum of 50 epochs, using early stopping and monitoring validation loss on the main task.

We do not use pre-trained embeddings. We also do not use any task-specific features, similarly to Collobert et al. (2011), and we do not optimise any hyperparameters with regard to the task(s) at hand. Although these choices are likely to affect the overall accuracy of our systems negatively, the goal of our experiments is to investigate the effect in *change* in accuracy when adding an auxiliary task - not accuracy in itself.

5.2 Experimental Overview

In the syntactic experiments, we train one system per language, dependency label category, and split condition. For sentences where only one tag set is available, we do not update weights based on the loss for the absent task. Averaged results over all languages and dependency relation instantiations, per category, are shown in Table 2.

5.3 Replicability and Reproducibility

In order to facilitate the replicability and reproducibility of our results, we take two methodological steps. To ensure replicability, we run all experiments 10 times, in order to mitigate the effect of random processes on our results.[4] To ensure reproducibility, we release a collection including: i) A Docker file containing all code and dependencies required to obtain all data and run our experiments used in this work; and ii) a notebook containing all code for the statistical analyses performed in this work.[5]

6 Results and Analysis

6.1 Morphosyntactic Tasks

We use Spearman's ρ in order to calculate correlation between auxiliary task effectivity (as measured using Δ_{acc}) and the information-theoretic measures. Following the recommendations in Søgaard et al. (2014), we set our p cut-off value

[4] Approximately 10,000 runs using 400,000 CPU hours.

[5] `https://github.com/bjerva/mtl-cond-entropy`

Auxiliary task	$\rho(\Delta_{acc}, H(Y))$	$\rho(\Delta_{acc}, H(Y\|X))$	$\rho(\Delta_{acc}, I(X;Y))$
Dependency Relations (Identity)	−0.06 (p=0.214)	0.12 (p=0.013)	0.08 (p=0.114)
Dependency Relations (Overlap)	0.07 (p=0.127)	0.27 (p<0.001)	**0.43 (p≪0.001)**
Dependency Relations (Disjoint)	0.08 (p=0.101)	0.25 (p<0.001)	**0.41 (p≪0.001)**

Table 2: Correlation scores and associated p-values, between change in accuracy (Δ_{acc}) and entropy ($H(Y)$), conditional entropy ($H(Y|X)$), and mutual information ($I(X;Y)$), calculated with Spearman's ρ, across all languages and label instantiations. Bold indicates the strongest significant correlations.

to $p < 0.0025$. Table 2 shows that MI correlates significantly with auxiliary task effectivity in the most commonly used settings (overlap and disjoint). As hypothesised, entropy has no significant correlation with auxiliary task effectivity, whereas conditional entropy offers some explanation. We further observe that these results hold for almost all languages, although the correlation is weaker for some languages, indicating that there are some other effects at play here. We also analyse whether significant differences can be found with respect to whether or not we have a positive Δ_{acc}, using a bootstrap sample test with 10,000 iterations. We observe a significant relationship ($p < 0.001$) for MI. We also observe a significant relationship for conditional entropy ($p < 0.001$), and again find no significant difference for entropy ($p \geq 0.07$).

Interestingly, no correlation is found in the identity condition between Δ_{acc} and any information-theoretic measure. This is not surprising, as the most effective auxiliary task is simply more data for a task with the highest possible MI. Hence, in the overlap/disjoint conditions, high MI is highly correlated with Δ_{acc}, while in the identity condition, there is no extra data. It is evident that tag set correlations in identical data is not helpful.

6.2 Semantic Tasks

Although we do not have access to sufficient data points to run statistical analyses on the results obtained by Martínez Alonso and Plank (2016), or by Bjerva et al. (2016), we do observe that the mean MI for the conditions in which an auxiliary task is helpful is higher than in the cases where an auxiliary task is not helpful.

7 Conclusions

We have examined the relation between auxiliary task effectivity and three information-theoretic measures. While previous research hypothesises that entropy plays a central role, we show experimentally that conditional entropy is a better predictor, and MI an even better predictor. This claim is corroborated when we correlate MI and change in accuracy with results found in the literature. It is especially interesting that MI is a better predictor than conditional entropy, since MI does not consider the order between main and auxiliary tasks. Our findings should prove helpful for researchers when considering which auxiliary tasks might be helpful for a given main task. Furthermore, it provides an explanation for the fact that there is no universally effective auxiliary task, as a purely entropy-based hypothesis would predict.

The fact that MI is informative when determining the effectivity of an auxiliary task can be explained by considering an auxiliary task to be similar to adding a feature. That is to say, useful features are likely to be useful auxiliary tasks. Interestingly, however, the gains of adding an auxiliary task are visible at test time for the main task, when no explicit auxiliary label information is available.

We tested our hypothesis on 39 languages, representing a wide typological range, as well as a wide range of data sizes. Our experiments were run on syntactically oriented tasks of various granularities. We also corroborated our findings with results from semantically oriented tasks in the literature. Hence our results generalise both across a range of languages, data sizes, and NLP tasks.

Acknowledgments

This work was funded by the NWO-VICI grant "Lost in Translation – Found in Meaning" (288-89-003). We would like to thank Barbara Plank, Robert Östling, Johan Sjons, and the anonymous reviewers for their comments on previous versions of this manuscript. We would also like to thank the Center for Information Technology of the University of Groningen for their support and for providing access to the Peregrine high performance computing cluster.

References

Waleed Ammar, George Mulcaire, Miguel Ballesteros, Chris Dyer, and Noah Smith. 2016. Many languages, one parser. *Transactions of the Association for Computational Linguistics*, 4:431–444.

Johannes Bjerva, Barbara Plank, and Johan Bos. 2016. Semantic tagging with deep residual networks. In *Proceedings of COLING 2016*, page 35313541, Osaka, Japan.

Rich Caruana. 1998. *Multitask learning*. Ph.D. thesis, Carnegie Mellon University.

Hao Cheng, Hao Fang, and Mari Ostendorf. 2015. Open-domain name error detection using a multitask rnn. In *EMNLP*.

Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*.

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*, pages 160–167. ACM.

Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, 12(Aug):2493–2537.

Thomas M Cover and Joy A Thomas. 2012. *Elements of information theory*. John Wiley & Sons.

Jeffrey L Elman. 1990. Finding structure in time. *Cognitive science*, 14(2):179–211.

Alex Graves and Jürgen Schmidhuber. 2005. Framewise phoneme classification with bidirectional lstm and other neural network architectures. *Neural Networks*, 18(5):602–610.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Diederik Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Héctor Martínez Alonso and Barbara Plank. 2016. Multitask learning for semantic sequence prediction under varying data conditions. In *arXiv preprint, to appear at EACL 2017 (long paper)*.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, Christopher D Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, et al. 2016. Universal dependencies v1: A multilingual treebank collection. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*.

Hiroki Ouchi, Kevin Duh, and Yuji Matsumoto. 2014. Improving dependency parsers with supertags. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, volume 2: Short Papers*, pages 154–158. Association for Computational Linguistics.

Hiroki Ouchi, Kevin Duh, Hiroyuki Shindo, and Yuji Matsumoto. 2016. Transition-Based Dependency Parsing Exploiting Supertags. In *IEEE/ACM Transactions on Audio, Speech and Language Processing*, volume 24.

Barbara Plank, Anders Søgaard, and Yoav Goldberg. 2016. Multilingual part-of-speech tagging with bidirectional long short-term memory models and auxiliary loss. In *Proceedings of ACL 2016*.

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, pages 231–235. Association for Computational Linguistics.

Anders Søgaard, Anders Johannsen, Barbara Plank, Dirk Hovy, and Hector Martinez. 2014. Whats in a p-value in NLP? In *CoNLL-2014*.

Nitish Srivastava, Geoffrey E Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(1):1929–1958.