One Model to Rule them All

Bjerva, Johannes

CHAPTER 5

*Information-theoretic Perspectives on Multitask Learning Effectivity

Abstract | In the previous chapter, we saw that multitask learning improved the performance on POS tagging, when using semantic tagging as an auxiliary task. In fact, multitask learning often improves system performance for various tasks in ML in general, and NLP in particular. 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, it can be argued that this is not sufficient. In this chapter, we will see that information-theoretic measures which consider the joint label distributions of the main and auxiliary tasks offer far more explanatory value. The findings in this chapter are empirically supported by experiments on morphosyntactic tasks on 39 languages, and by experiments on several semantic tasks for English.

*Chapter adapted from: **Bjerva, J.** (2017) Will my auxiliary tagging task help? Estimating Auxiliary Tasks Effectivity in Multi-Task Learning, in Proceedings of the 21st Nordic Conference on Computational Linguistics, NoDaLiDa, 22-24 May 2017, Gothenburg, Sweden, number 131, pages 216–220. Linköping University Electronic Press, Linköpings universitet. Best short-paper award.

5.1 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. As discussed in Chapter 3, a common way of taking advantage of this is to apply multitask learning (MTL, Caruana (1997)). 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; Bjerva et al., 2016b; Ammar et al., 2016; Plank et al., 2016; Martínez Alonso and Plank, 2017; Bingel and Søgaard, 2017). This trend is in part owed to the fact that a specific type of MTL, namely hard parameter sharing in neural networks, is relatively easy to implement and often quite effective. It is, however, unclear *when* an auxiliary task is useful, although previous work has provided some insights (Caruana, 1997; Martínez Alonso and Plank, 2017; Bingel and Søgaard, 2017). For a further overview of MTL, see Chapter 3.

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 chapter the aim is to alleviate this process by providing a means to empirically investigating the potential effectivity of an auxiliary task. We aim to answer the following two research questions, in order to answer **RQ 2**:

RQ 2a Which information-theoretic measures can be used to estimate auxiliary task effectivity?

RQ 2b To what extent do correlations between information-theoretic measures and auxiliary task effectivity generalise across languages and NLP tasks?

Concretely, we apply information-theoretic measures to a collection of data- and tag sets, and investigate correlations between these measures and auxiliary task effectivity. We investigate this both exper-

imentally on a collection of syntactically oriented tasks on 39 languages, as well as on several semantically oriented tasks for English. We take care to structure our experiments so as to generalise across many common real-world situations in which MTL is applied. Concretely, we apply neural multitask learning (see Chapter 5), using a bi-directional GRU (bi-GRU, see Section 2.4), as introduced in Chapter 4, using hard parameter sharing.

5.2 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.[2]

5.2.1 Entropy

The **entropy** of a probability distribution, originally described in Shannon and Weaver (1949), 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

$$H(X) = -\sum_{x \in X} p(x) \log p(x), \tag{5.1}$$

where x is a given tag in tag set X.

5.2.2 Conditional Entropy

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**. This is depicted in Figure 5.1 as $H(X|Y)$ and

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

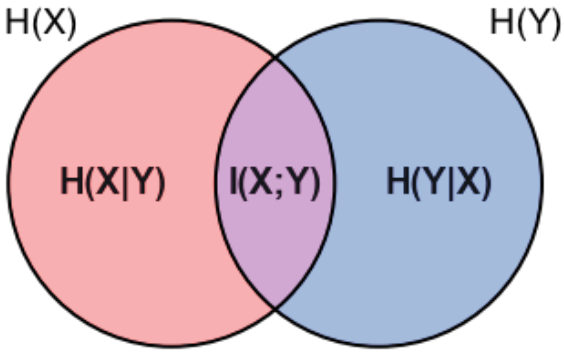


Figure 5.1: Information theory overview. The left circle denotes H(X), and the right circle H(Y). Blue ($H(X|Y)$) indicates the conditional entropy of X given Y, red ($H(Y|X)$) indicates the opposite, and purple ($I(X;Y)$) indicates the mutual information of X and Y.

$H(Y|X)$, with red and blue respectively. Formally, the conditional entropy of a distribution Y given the distribution X is defined as

$$H(Y|X) = \sum_{x \in X} \sum_{y \in Y} p(x, y) \log \frac{p(x)}{p(x, y)}, \tag{5.2}$$

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. In the case of a multitask setup, Y and X are the distributions of the main and auxiliary task tag sets respectively. The variables y and x are specific tags in these tag sets.

5.2.3 Mutual Information

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

$$I(X;Y) = \sum_{x \in X} \sum_{y \in Y} p(x,y) \log \frac{p(x,y)}{p(x)\,p(y)}, \tag{5.3}$$

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. This is depicted in Figure 5.1 as $I(X;Y)$, in purple, which also illustrates the alternative definition of MI, namely expressed in terms of entropy and conditional entropy as

$$I(X;Y) \equiv H(X) - H(X|Y) \equiv H(Y) - H(Y|X). \tag{5.4}$$

In the figure, this is depicted as that subtracting either $H(X)$ from $H(X|Y)$ or $H(Y)$ from $H(Y|X)$ will result in $I(X;Y)$. MI describes how much information is shared between X and Y, and can therefore be considered a measure of 'correlation' between tag sets. Should two tag sets be completely independent from each other, then knowing Y would not give any information about X.

5.2.4 Information Theory and MTL in NLP

Entropy has in the literature been hypothesised to be related to the usefulness of an auxiliary task (Martínez Alonso and Plank, 2017). 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. Such a situation is shown in the following examples:

(5.5) *The quick brown fox jumps over the lazy dog .*
DET ADJ ADJ NOUN VERB ADP DET ADJ NOUN PUNCT

(5.6) *The quick brown fox jumps over the lazy dog .*
ADJ ADJ DET DET NOUN NOUN ADJ ADP VERB PUNCT

The tag distributions in X and X' do not change as a result of a reordering as in the examples, hence the tag set entropies will be the same.[3] However, the tags in X' are now likely to have a vanishingly low correspondence with any sort of natural language signal (as in the second sentence), 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.

5.3 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., 2016a).[4] We experiment with POS tagging as a main task, and various dependency relation classification tasks (as defined in Section 5.3.1) as auxiliary tasks. We also investigate whether our hypothesis fits with recent results in the literature, and train sequence taggers on the collection of semantically oriented tasks presented in Martínez Alonso and Plank (2017), as well as on the semantic tagging task in Bjerva et al. (2016b).

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

[3] Note that we look at the entropy of the marginal distribution of each tag set, as this is what has been hypothesised to be of importance in previous work.

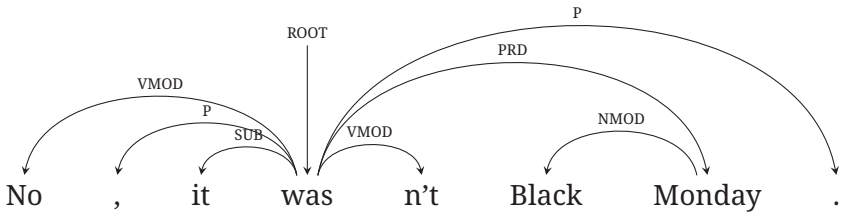
[4] Japanese was excluded due to treebank unavailability.

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.[5,6]

5.3.1 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, differing in the amount of information they encode (Ouchi et al., 2014, 2016). In this work, we use labels designed to range from highly complex and informative, to relatively basic ones.[7] The labelling schemes used are shown in Table 5.1. As an example, consider the following dependency graph:



[5]Since the dependency relation auxiliary task data overlaps with POS tagging data, this allowed us to confirm that the differences between the measures obtained with estimated and gold data in this case are negligible (i.e. $\leq 5\%$).

[6]The POS-tagger used is the deep bi-GRU ResNet-tagger described in Chapter 4.

[7]Labels are automatically derived from the UD dependency annotations.

Table 5.1: Dependency relation labels used in this work, with entropy in bits (H) measured on English. The labels differ in the granularity and/or inclusion of the category and/or directionality.

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 5.2 shows examples of dependency relation labels based on this graph. The labels encode the head of each word, as well as the relative position, or direction. For instance, the word *it* is the subject of a word on its right. The more complex tags, for instance *ROOT+SUB/L_PRD/R*, include information regarding the dependents of each word. In this case, the word *was* has an obligatory *sub* dependent to the left, and an obligatory *prd* dependent on the right.

Table 5.2: Examples of some dependency relation instantiations in context. The columns indicate the granularity used (category/directionality).

Word	**Full/Full**	**Simple/Simple**	**Simple/None**	**None/Simple**
No	VMOD/R	VMOD/R	VMOD	R
,	P/R	P/R	P	R
it	SUB/R	SUB/R	SUB	R
was	ROOT+SUB/L_PRD/R	ROOT	ROOT	
n't	VMOD/L	VMOD/L	VMOD	L
Black	NMOD/R	NMOD/R	NMOD	R
Monday	PRD/L+L	PRD/L	PRD	L
.	P/L	P/L	P	L

Table 5.3: Data splitting scheme. The training set is split into three equal parts. The annotations in each part differ per condition.

Condition	Part I	Part II	Part III
Identity	PoS	PoS $\wedge$ DepRel	n/a
Overlap	PoS	PoS $\wedge$ DepRel	DepRel
Disjoint	PoS	PoS	DepRel

The systems in the syntactic experiments are trained on main task data (D_{main}), and on auxiliary task data (D_{aux}). Generally, the amount of overlap between such pairs of data sets differs, and can roughly be divided into three categories: i) identity (identical data sets); ii) overlap (some overlap between data sets); 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 (D_{main}, 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. Hence, the system always sees the exact same POS tagging data, whereas the amount of dependency relation data and overlap differs between conditions. Each dependency relation instantiation is used in our experiments, paired with PoS tagging. The data splitting scheme is shown in Table 5.3.

5.3.2 Semantic Tasks

Martínez Alonso and Plank (2017) experiment with using POS tagging, chunking, dependency relation tagging, and a frequency based measure as auxiliary tasks, with main tasks based on several seman-

tically oriented tasks. In this chapter, we limit ourselves to considering the PoS tagging auxiliary task, for the following semantic main tasks.

Named Entity Recognition

For NER, we use the CONLL2003 shared-task data (e.g. Person, Loc, etc., Tjong Kim Sang and De Meulder (2003)).

Frames

We use FrameNet 1.5 (Baker et al., 1998) with the same data splits as Das et al. (2014) and Hermann et al. (2014). This data set is annotated for the joint task of frame detection and identification. As in Martínez Alonso and Plank (2017), we approach this task as a standard sequence prediction task.

Supersenses

We experiment with the supersense version of SemCor (Miller et al., 1993) from Ciaramita and Altun (2006), using course-grained semantic labels (e.g. noun.person).

Semtraits

We use the conversions of Martínez Alonso and Plank (2017) of supersenses into coarser semantic traits (e.g. Animate, UnboundedEvent, etc.), for which they used the EuroWordNet list of ontological types for senses from Vossen et al. (1998).

Multi-Perspective Question Answering

We also consider the Multi-Perspective Question Answering (MPQA) corpus, using the coarse level of annotation (Deng and Wiebe, 2015).

Semantic Tags

We also investigate the semantic tagging task of Chapter 4, using the same data splits (Bjerva et al., 2016b).

We use the same setup as for our syntactic experiments, by using these semantic tasks as auxiliary tasks with POS tagging as the main task.

5.4 Method

5.4.1 Architecture and Hyperparameters

We apply a deep neural network with the exact same hyperparameter settings in each syntactic experiment, with reasonably default parameter settings, similar to what was used in Chapter 4. Our system consists of a two layer deep bi-GRU (100 dimensions per layer), taking an embedded word representation (64 dimensions) as input (see Figure 5.2). 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 hyper-parameters 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.

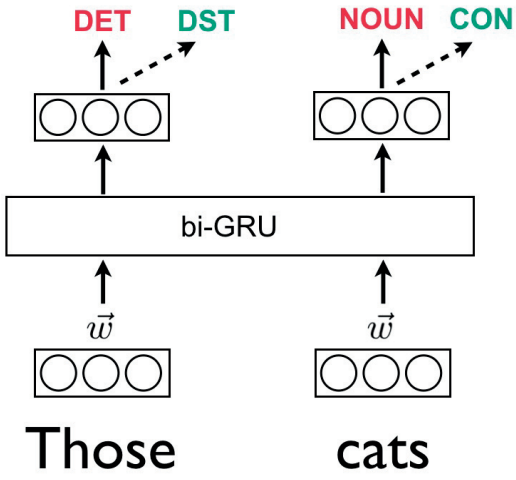


Figure 5.2: System architecture used in the multitask learning experiments.

5.4.2 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.

5.4.3 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.[8] To ensure reproducibility, we release a collection including: i) A Docker file containing all code and depen-

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

dencies 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.[9]

5.5 Results and Analysis

Table 5.4: Morphosyntactic tasks. Correlation scores and associated p-values, between change in accuracy (Δ_{acc}) and entropy ($H(Y)$), conditional entropy ($H(X|Y)$, $H(Y|X)$), and mutual information ($I(X;Y)$), calculated with Spearman's ρ, across all languages and label instantiations. Bold indicates the strongest significant correlations.

Condition	$\rho(\Delta_{acc}, H(Y))$	$\rho(\Delta_{acc}, H(X\|Y))$	$\rho(\Delta_{acc}, H(Y\|X))$	$\rho(\Delta_{acc}, I(X;Y))$
Identity	$-$0.06 (p=0.214)	0.10 (p=0.020)	0.12 (p=0.013)	0.08 (p=0.114)
Overlap	0.07 (p=0.127)	0.23 (p<0.001)	0.27 (p<0.001)	**0.43 (p$\ll$0.001)**
Disjoint	0.08 (p=0.101)	0.26 (p<0.001)	0.25 (p<0.001)	**0.41 (p$\ll$0.001)**

Table 5.5: Change in accuracy, and information theoretic measures, for the semantic tasks.

Auxiliary task	Δ_{acc}	$H(Y)$	$H(X\|Y)$	$H(Y\|X)$	$I(X;Y)$
Frames	-14.64	1.6	2.7	1.4	0.2
MPQA	-5.62	1.1	2.6	1.0	0.1
Supersenses	-2.86	1.8	2.7	1.6	0.2
NER	-1.36	0.8	2.6	0.7	0.1
Semtraits	0.67	1.3	3.0	0.8	0.5
Semtagging	0.79	3.0	2.0	1.5	1.5

5.5.1 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.

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

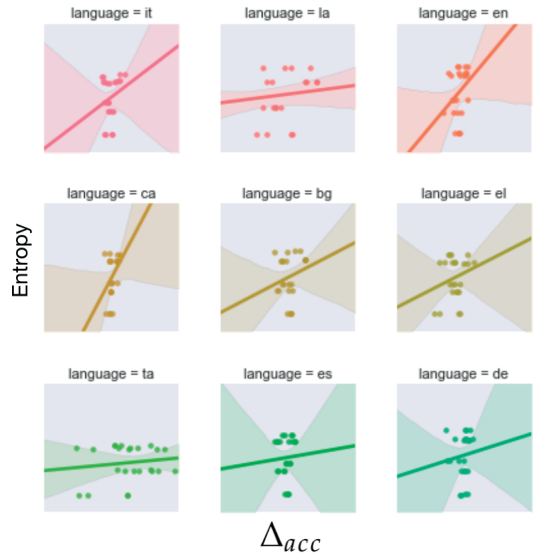


Figure 5.3: Correlations between Δ_{acc} and entropy. Each data point represents a single experiment run.

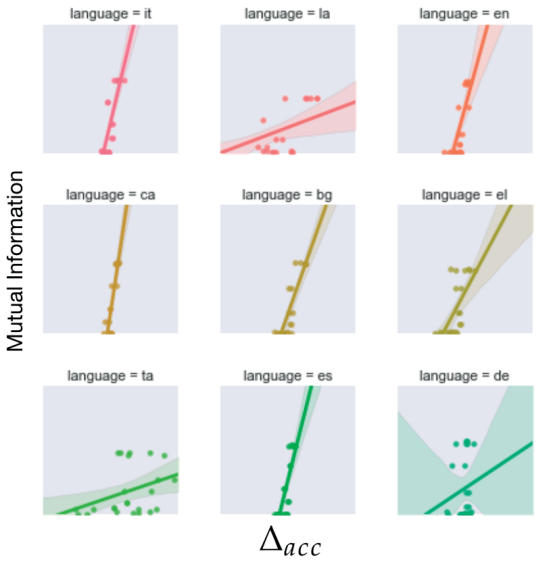


Figure 5.4: Correlations between Δ_{acc} and mutual information. Each data point represents a single experiment run.

(2014), we set our p cut-off value to $p < 0.0025$. Table 5.4 shows that MI correlates significantly with auxiliary task effectivity in the most commonly used settings (overlap and disjoint). The fact that no correlation is found in the identity condition between Δ_{acc} and any information-theoretic measure yields some interesting insights into the issue at hand. Intuitively, this makes sense considering that the lack of any extra data in the identity setting does not offer much opportunity for the network to learn from the auxiliary data. In other words, if the model is allowed to train on the same data with essentially the same label (i.e., if MI is high), this does not allow the model to learn anything new. This is supported by the significant positive correlation between MI and Δ_{acc} in the overlap/disjoint conditions, in which the model does have access to more data. This further suggests that in some cases, one of the most effective auxiliary tasks is simply more data for the same task (i.e., the highest MI achievable). Additionally, this shows that high MI between tag sets in identical data is not necessarily helpful, and that in such a setting it may even be advantageous to have a less similar auxiliary task.

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. For a sample of the languages, the correlations between Δ_{acc} and Entropy are shown in Figure 5.3, and the correlations between Δ_{acc} and Mutual Information are shown in Figure 5.4.[10] 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 (Efron and Tibshirani, 1994). We observe a significant relationship ($p < 0.001$) for MI. We also observe a significant relationship for

[10]These figures only show a subset of the languages under evaluation. See Appendix A for a complete overview.

conditional entropy ($p < 0.001$), and again find no significant difference for entropy ($p \geq 0.07$).

5.5.2 Language-dependent results

Results per language are shown in Table 5.6. While not all languages exhibit correlations below our selected α-level, the non-significant languages still exhibit interesting trends in the same direction. Note that there are two cases where Entropy is a fair predictor of Δ_{acc}, namely for Latvian and Turkish. However, in both of these cases the correlation is stronger still with MI. Furthermore, the correlations between Δ_{acc} and entropy vary wildly between languages, sometimes exhibiting negative correlations.

5.5.3 Semantic Tasks

We do not have access to sufficient data points to run statistical analyses on the results obtained by Martínez Alonso and Plank (2017), or by Bjerva et al. (2016b), and the results in Table 5.5 do not reveal any obvious patterns. A grouping of these results by whether or not Δ_{acc} was positive can be seen in Figure 5.5, which offers some support to the results from the morphosyntactic tasks. However, the lack of a clear pattern when looking at individual results per dataset serves to highlight the issue at hand, namely that even though MI offers some explanatory value, the interactions behind the workings of MTL are more complex than what can be explained purely by comparing joint distributions of tag sets.

5.6 Conclusions

We have examined the relation between auxiliary task effectivity and three information-theoretic measures. The first research question which we aimed to answer in this chapter (**RQ 2a**) was related

Table 5.6: 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. Bold indicates the strongest significant correlations per row.

Group	**Language**	$\rho(\Delta_{acc}, H(Y))$	$\rho(\Delta_{acc}, H(Y\|X))$	$\rho(\Delta_{acc}, I(X;Y))$
Germanic	Danish	0.27 (p=0.116)	0.42 (p=0.011)	**0.78 (p≪0.001)**
	Dutch	0.31 (p=0.070)	0.16 (p=0.337)	**0.55 (p<0.001)**
	English	0.30 (p=0.076)	0.19 (p=0.280)	**0.58 (p<0.001)**
	German	0.03 (p=0.849)	0.13 (p=0.448)	0.18 (p=0.293)
	Norwegian	-0.03 (p=0.858)	0.23 (p=0.183)	0.23 (p=0.177)
	Swedish	-0.03 (p=0.843)	0.29 (p=0.091)	0.31 (p=0.068)
Romance	Catalan	0.34 (p=0.042)	0.33 (p=0.047)	**0.72 (p≪0.001)**
	French	0.06 (p=0.734)	0.38 (p=0.023)	0.48 (p=0.003)
	Galician	0.10 (p=0.574)	0.18 (p=0.304)	0.28 (p=0.099)
	Italian	0.12 (p=0.503)	0.52 (p=0.001)	**0.67 (p≪0.001)**
	Portuguese	-0.02 (p=0.921)	0.61 (p<0.001)	**0.66 (p<0.001)**
	Romanian	-0.31 (p=0.067)	0.34 (p=0.040)	0.04 (p=0.825)
	Spanish	0.02 (p=0.890)	0.60 (p<0.001)	**0.70 (p≪0.001)**
Slavic	Bulgarian	0.20 (p=0.242)	0.50 (p=0.002)	**0.76 (p≪0.001)**
	Croatian	-0.24 (p=0.159)	0.43 (p=0.009)	0.22 (p=0.189)
	Czech	-0.15 (p=0.376)	0.49 (p=0.002)	0.39 (p=0.017)
	O.C. Slavonic	-0.08 (p=0.634)	0.34 (p=0.044)	0.35 (p=0.038)
	Polish	0.13 (p=0.437)	0.40 (p=0.015)	**0.59 (p<0.001)**
	Russian	0.29 (p=0.086)	0.40 (p=0.015)	**0.81 (p≪0.001)**
	Slovene	-0.24 (p=0.156)	0.41 (p=0.014)	0.19 (p=0.259)
Turkic	Kazakh	0.23 (p=0.172)	0.04 (p=0.817)	0.36 (p=0.030)
	Turkish	0.50 (p=0.002)	-0.17 (p=0.317)	0.43 (p=0.008)
Uralic	Estonian	0.45 (p=0.006)	-0.14 (p=0.430)	0.39 (p=0.017)
	Finnish	0.02 (p=0.924)	0.37 (p=0.025)	**0.50 (p=0.002)**
	Hungarian	0.14 (p=0.413)	0.09 (p=0.594)	0.27 (p=0.116)
Other	Arabic	-0.16 (p=0.362)	0.53 (p<0.001)	0.47 (p=0.004)
	Basque	0.41 (p=0.014)	-0.01 (p=0.952)	**0.49 (p=0.002)**
	Chinese	-0.15 (p=0.399)	0.46 (p=0.005)	0.41 (p=0.012)
	Farsi	0.20 (p=0.244)	0.41 (p=0.012)	**0.75 (p≪0.001)**
	Greek	0.20 (p=0.248)	0.19 (p=0.264)	0.44 (p=0.007)
	Hebrew	0.06 (p=0.724)	0.37 (p=0.028)	**0.52 (p=0.001)**
	Hindi	-0.26 (p=0.121)	0.24 (p=0.161)	0.00 (p=0.979)
	Irish	-0.24 (p=0.150)	0.54 (p<0.001)	0.35 (p=0.034)
	Indonesian	-0.42 (p=0.011)	0.51 (p=0.001)	0.11 (p=0.510)
	Latin	0.19 (p=0.271)	0.16 (p=0.362)	0.47 (p=0.004)
	Latvian	0.64 (p<0.001)	-0.23 (p=0.171)	**0.53 (p<0.001)**
	Tamil	0.16 (p=0.337)	0.12 (p=0.482)	0.31 (p=0.067)

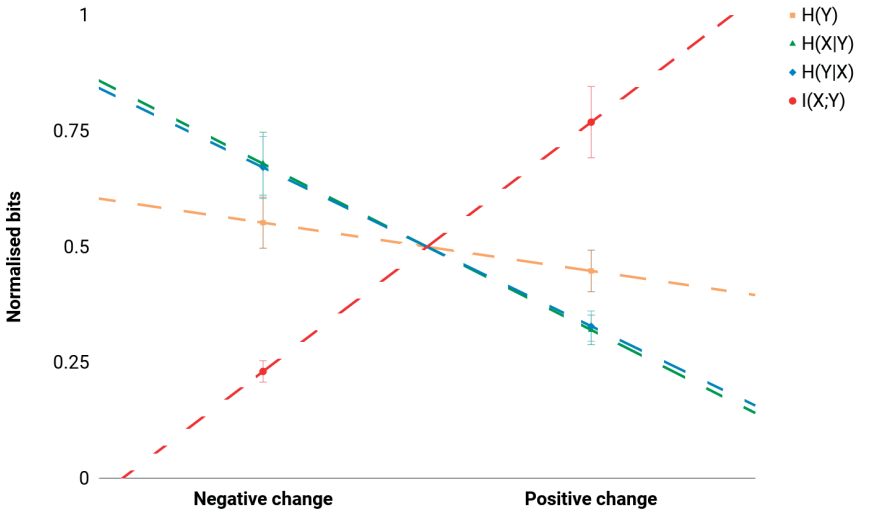


Figure 5.5: Comparison of information theoretic measures and change in accuracy for the semantic tasks. Results are grouped by negative and positive change in Δ_{acc}. Dashed lines are included for clarity, and are not meant to imply a linear correlation.

to which information-theoretic measures are useful for estimation of auxiliary task effectivity. While previous research hypothesises that entropy plays a central role, we show experimentally that this is not sufficient, and that conditional entropy is a somewhat better predictor, and that MI is the best predictor under consideration here. 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 is a symmetric measure, as it does not consider the order between main and auxiliary tasks. For conditional entropy itself, the results for the two directionalities did not differ to a large extent. Our find-

ings 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 assumes.

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.

The second research question which we aimed to answer in this chapter (**RQ 2b**), related to the generalisation ability of the information-theoretic measures as a measure of auxiliary task effectivity, across languages and NLP tasks. 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.

While the correlations between MI and Δ_{acc} were higher than for other information-theoretic measures, it is by no means a perfect predictor. This highlights the fact that the interactions between tasks is more complex than simply the joint distribution between the tag sets at hand. One possibility for future work is to take distributions over tags and words into account simultaneously.

Having considered interactions between tasks in MTL, and finding that task similarity is to some extent predictive of MTL effectivity, this raises the question of what the situation is in the case of multilingual learning. This is explored further in the next part of this thesis.