One Model to Rule them All

Bjerva, Johannes

PART V

Conclusions

CHAPTER 9

Conclusions

While traditional NLP approaches consider a single task or language at a time, the aim of this thesis was to answer several research questions dealing with pushing past this boundary. In doing so, the hope is that in the long term, low-resource languages can benefit from the advances made in NLP which are currently to a large extent reserved for high-resource languages. This, in turn, may then have positive consequences for, e.g., language preservation, as speakers of minority languages will have a lower degree of pressure to using high-resource languages. In the short term, answering the specific research questions posed should be of use to NLP researchers working towards the same goal. We will now see the conclusions which can be drawn from each research part of this thesis.

9.1 Part II - Multitask Learning

In the first research part of the thesis, we began by exploring the following research question in Chapter 4.

> *'To what extent can a semantic tagging task be informative for other NLP tasks?'*
> **–RQ 1**

We found that semantic tags are informative for the task of PoS tagging. Furthermore, the results obtained when exploiting this were state-of-the-art results at the time. Additionally, we found that using coarse-grained semantic tags was not informative for semantic tagging. This then raised more questions. Why were the semantic tags useful for PoS tags, while coarse-grained semantic tags were not useful for semantic tagging? A look at correlations between the tag sets showed that these were high in both cases. Coarse-grained semantic tags have a one-to-one mapping with semantic tags. Semantic tags do not have a one-to-one mapping with PoS tags, but still exhibit large correlations. The idea was, then, that such high correlations between tag sets might correlate with auxiliary task effectivity, given differing data sets. Thus, we aimed at answering the next research question in Chapter 5:

> *'How can multitask learning effectivity in NLP be quantified?'*
> **–RQ 2**

Taking an information-theoretic perspective, we found that these correlations could be quantified fairly well by using mutual information. Running experiments in various data overlap settings, on a large selection of languages and tasks, showed that the hypothesis was supported. That is to say, providing the model with different data including annotations which correlate highly with the main task yields gains in performance. However, providing the model with the same data with such highly correlated auxiliary annotations, does not yield any increase at all. Intuitively, this makes sense if one thinks about it as follows. The model has already seen sentence x with some annotation. Giving it the same sentence x with highly correlated annotation does not give the model anything more to learn from – after all, this example has practically already been observed!

However, giving the model a different sentence y with highly correlated annotation essentially entails giving the model an extra training example.

9.2 Part III - Multilingual Learning

In the second content part of the thesis, the aim was to investigate similar research questions to Part I, focussing on similarities between *languages* rather than between tasks. We began by asking the following research question.

> *'To what extent can multilingual word representations be used to enable zero-shot learning in semantic textual similarity?'*
> –**RQ 3**

In Chapter 6 we found that a simple language-agnostic feed-forward neural network using multilingual word representations was able to solve the task of semantic textual similarity assessment to some extent. Although results were below the current state-of-the-art for this task, some useful insights were gained. Mainly, we found that languages which are more similar to one another are more suited for this approach, indicating that language similarity is important for the effectivity of model multilinguality. This is similar to the case in MTL, where task relatedness is an important factor, and raised the following research question which we approached in Chapter 7.

> *'In which way can language similarities be quantified, and what correlations can we find between multilingual model performance and language similarities?'*
> –**RQ 4**

We looked at correlations between language similarity and multilingual model effectivity in two sequence prediction tasks, namely se-

mantic tagging and PoS tagging, as well as in a sequence-to-sequence task, namely morphological inflection. The overall results indicate that both measures of language similarity under consideration offer some explanatory value. One interesting finding in the case of semantic tagging, was the fact that English, Dutch, and German benefitted from having their input representations updated during joint training. Combining these languages with Italian and updated embeddings, however, resulted in a serious drop in performance. A potential reason for this is that language relatedness plays a large role in maintaining the quality of the multilingual embedding space in such a context. In future work, it would therefore be interesting to observe the resulting word-space after updating word representations in such a setting.

9.3 Part IV - Combining Multitask Learning and Multilinguality

In the final research part of the thesis, the aim was to probe the possibilities of combining the paradigms of multitask learning and multilingual learning. Chapter 7 aimed at providing an answer to the following research question.

> *'Can a multitask and multilingual approach be combined to generalise across languages and tasks simultaneously?'*
> **–RQ 5**

We looked at predicting labels for an unseen task–language combination, by taking advantage of other task–language combinations. In the admittedly somewhat artificial setup, the target task was PoS tagging for three languages offering some typological diversity, namely Finnish, Italian, and Slovene. In a high-resource scenario, assuming access to parallel text similar to Europarl, sensible tags could be produced for the target languages without seeing any annotated data for that target language. In the low-resource scenario, assuming access

to parallel text similar to the New Testament, similar results have the additional requirement of also having access to target-language annotations of some sort. Finally, access to the high-resource scenario as well as target-language annotations yielded results on par with a monolingual monotask PoS tagger for the target language – and that without seeing a single PoS tag for the target language.

9.4 Final words

A large part of this thesis was motivated by the intuition that similarities between tasks and languages is one of the most important factors when considering a multitask or a multilingual approach. Even though some correlations were found in experiments, attempting to correlate measures of task and language similarities with change in model performance, much of the change that is observed is left unaccounted for. This highlights the case that even if such similarities are important, the situation is more complex than what can be explained purely by measures of correlation.

The successful experiments dealing with the combination of multitask and multilingual learning show the most potential for future research based on this thesis. A plethora of new studies based on this idea can be imagined. One could take advantage of morphological similarities by looking at character-level representations, investigating to what extent an architecture such as sluice networks can learn to share parameters for similar task–language combinations. Another option is to probe into how much annotation is needed in order to bootstrap off of other languages than the target language at hand in order to predict reasonable labels for the target language.

A concrete proposal toward *One Model to rule them all* at a larger scale, involving more languages and tasks, is to model this in a sluice network (Ruder et al., 2017). In this recently proposed architecture, the sharing of layers itself is learned by the network. Combining a

large amount of tasks in such a network should therefore allow for taking advantage of relevant similarities between tasks, while not sharing parameters in the case of dissimilarities which may lead to negative transfer. Taking this one step further, by also involving multilingual learning as in this thesis, could also allow for learning between which languages to share parameters. For instance, this architecture might be able to learn which parts of a character-RNN to share between which languages, for instance learning to only share these parameters between closely related languages, thus avoiding any negative transfer in this setting. Further combining this approach with language vectors (Östling and Tiedemann, 2017; Malaviya et al., 2017) might facilitate exploitation of language similarities. This approach might therefore alleviate many of the problems with hard parameter sharing, by allowing the model to only utilise parameter sharing for similarities between languages, while learning separate parameters for language-specific features. As the amount of both unannotated parallel data, and annotated data with various universal annotation schemes increases, it is only a matter of choosing the right approach in order to arrive at *One Model to rule them all*.