One Model to Rule them All

Bjerva, Johannes

CHAPTER 1

Introduction

When learning a new skill, you take advantage of your preexisting skills and knowledge. For instance, if you are a skilled violinist, you will likely have an easier time learning to play cello. Similarly, when learning a new language you take advantage of the languages you already speak. For instance, if your native language is Norwegian and you decide to learn Dutch, the lexical overlap between these two languages will likely benefit your rate of language acquisition. This thesis deals with the intersection of learning multiple tasks and learning multiple languages in the context of Natural Language Processing (NLP), which can be defined as the study of computational processing of human language. Although these two types of learning may seem different on the surface, we will see that they share many similarities.

Traditionally, NLP practitioners have looked at solving a single problem for a single task at a time. For instance, considerable time and effort might be put into engineering a system for part-of-speech (PoS) tagging for English. However, although the focus has been on considering a single task at a time, fact is that many NLP tasks are highly related. For instance, different lexical tag sets will likely exhibit high correlations with each other. As an example, consider the

following sentence annotated with Universal Dependencies (UD) PoS tags (Nivre et al., 2016a), and semantic tags (Bjerva et al., 2016b).[1,2]

(1.1)

We	*must*	*draw*	*attention*	*to*	*the*	*distribution*	*of*
PRON	AUX	VERB	NOUN	ADP	DET	NOUN	ADP
this	*form*	*in*	*those*	*dialects*	.		
DET	NOUN	ADP	DET	NOUN	PUNCT		

(1.2)

We	*must*	*draw*	*attention*	*to*	*the*	*distribution*	*of*
PRO	NEC	EXS	CON	REL	DEF	CON	AND
this	*form*	*in*	*those*	*dialects*	.		
PRX	CON	REL	DST	CON	NIL		

While these tag sets are certainly different, the distinctions they make compared to one another in this example are few, as there are only two apparent systematic differences. Firstly, the semantic tags offer a difference between definite (DEF), proximal (PRX), and distal determiners (DST), whereas UD lumps these together as DET (highlighted in green). Secondly, the semantic tags also differentiate between relations (REL) and conjunctions (AND), which are both represented by the ADP PoS tag, highlighted in blue. Hence, although these tasks are undoubtedly different, there are considerable correlations between the two, as the rest of the tags exhibit a one-to-one mapping in this example. This raises the question of how this fact can be exploited, as it seems like a colossal waste to not take advantage of such inter-task correlations. In this thesis I approach this by exploring multitask learning (MTL, Caruana, 1993; 1997), which has been beneficial for many NLP tasks. In spite of such successes, however, it is not clear *when* or *why* MTL is beneficial.

[1] PMB 01/3421. Original source: Tatoeba. UD tags obtained using UD-Pipe (Straka et al., 2016)

[2] The semantic tag set consists of 72 tags, and is developed for multilingual semantic parsing. The tag set is described further in Chapter 4.

Similarly to how different tag sets correlate with each other, languages also share many commonalities with one another. These resemblances can occur on various levels, with languages sharing, for instance, syntactic, morphological, or lexical features. Such similarities can have many different causes, such as common language ancestry, loan words, or being a result of universals and constraints in the properties of natural language itself (see, e.g., Chomsky,2005, and Hauser et al., 2002). Consider, for instance, the following German translation of the previous English example, annotated with semantic tags.[3]

(1.3)

Wir	*müssen*	*die*	*Verbreitung*	*dieser*	*Form*	*in*	*diesen*
PRO	NEC	DEF	CON	PRX	CON	REL	PRX

Dialekten	*beachten*	*.*
CON	EXS	NIL

Comparing the English and German annotations, there is a high overlap between the semantic tags used, and a high lexical overlap. As in the case of related NLP tasks, this begs the question of how multilinguality can be exploited, as it seems like an equally colossal waste to not consider using, e.g., Norwegian PoS data when training a Swedish PoS tagger. There are several approaches to exploiting multilingual data, such as annotation projection and model transfer, as detailed in Chapter 3. The approach in this thesis is a type of model transfer, in which such inter-language relations are exploited by exploring multilingual word representations, which have also been beneficial for many NLP tasks. As with MTL, in spite of the fact that such approaches have been successful for many NLP tasks, it is not clear in which cases it is an advantage to *go multilingual*.

Given the large amount of data available for many languages in different annotations, it is tempting to investigate possibilities of com-

[3] PMB 01/3421. Original source: Tatoeba.

bining the paradigms of multitask learning and multilingual learning in order to take full advantage of this data. Hence, as the title of the thesis suggests, the final effort in this thesis is to arrive at *One Model to rule them all*.

This thesis approaches these two related aspects of NLP by experimenting with deep neural networks, which represent a family of learning architectures which are exceptionally well suited for the aforementioned purposes (described in Chapter 2). For one, it is fairly straightforward to implement the sharing of parameters between tasks, thus enabling multitask learning (discussed in Chapter 3). Additionally, providing such an architecture with multilingual input representations is also straightforward (discussed in Chapter 3). Experiments in this thesis are run on a large collection of tasks, both semantic and morphosyntactic in nature, and a total of 60 languages are considered, depending on the task at hand.

1.1 Chapter guide

The thesis is divided into five parts, totalling 9 chapters, aiming to provide answers to the following general research questions (**RQ**s):

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

RQ 2 How can multitask learning effectivity in NLP be quantified?

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

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

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

Part I – Background

The goal of Part I is to provide the reader with sufficient background knowledge to understand the material in this thesis. Chapter 2 contains a crash-course in neural networks, introducing the main concepts and architectures used in NLP. Chapter 3 provides an introduction to multitask learning and multilingual learning, which are the two central topics of this work.

Part II – Multitask Learning

In Part II, the goal is to investigate multitask learning (MTL), in particular by looking at the effects of this paradigm in NLP sequence prediction tasks. In Chapter 4 we present a semantic tag set tailored for multilingual semantic parsing. We attempt to use this tag set as an auxiliary task for PoS tagging, observing what effects this yields, answering **RQ 1**. Chapter 5 then delves deeper into MTL, attempting to find when MTL is effective, and how this effectiveness can be predicted by using information-theoretic measures (**RQ 2**).

Part III – Multilingual Learning

Having looked at similarities between tasks, we turn to similarities between languages in Part III. In Chapter 6, we attempt to make a language-agnostic solution for semantic textual similarity, by exploiting multilingual word representations, thus answering **RQ 3**. Having seen the results of combining related languages in this task, we try to quantify these effects in Chapter 7, aiming to answer **RQ 4**.

Part IV – Combining Multitask and Multilingual Learning

In Part IV we want to combine the paradigms of multitask learning and multilingual learning in order to make *One Model to rule them all*. Chapter 8 presents a pilot study taking a step in this direction, looking

at predicting labels for an unseen task–language combination while exploiting other task–language combinations.

Part V – Conclusions

Finally, Chapter 9 contains an overview of the conclusions from this thesis. In addition to this, we provide an outlook for future work in this direction, in particular focussing on the combined multitask–multilingual paradigm.

1.2 Publications

This thesis is based on the following publications:

1. Bjerva, J., Plank, B., and Bos, J. (2016b). Semantic tagging with deep residual networks. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3531–3541

2. Bjerva, J. (2017b). 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.

3. Bjerva, J. and Östling, R. (2017a). Cross-lingual Learning of Semantic Textual Similarity with Multilingual Word Representations. In *Proceedings of the 21st Nordic Conference on Computational Linguistics, NoDaLiDa, 22-24 May 2017, Gothenburg, Sweden*, number 131, pages 211–215. Linköping University Electronic Press, Linköpings universitet

4. Bjerva, J. and Östling, R. (2017b). Multilingual word representations for semantic textual similarity. In *Proceedings of SemEval 2017: International Workshop on Semantic Evaluation*

5. Bjerva, J. (2017a). Quantifying the Effects of Multilinguality in NLP Sequence Prediction Tasks. *Under review*

Some parts of the thesis may also refer to the following peer-reviewed publications completed in the course of the PhD:

6. Bjerva, J. (2014). Multi-class animacy classification with semantic features. In *Proceedings of the Student Research Workshop at the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 65–75

7. Bjerva, J., Bos, J., Van der Goot, R., and Nissim, M. (2014). The meaning factory: Formal semantics for recognizing textual entailment and determining semantic similarity. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 642–646, Dublin, Ireland

8. Bjerva, J. and Praet, R. (2015). Word embeddings pointing the way for late antiquity. In *9th SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences and Humanities (LaTeCH 2015)*, pages 53–57

9. Bjerva, J. and Börstell, C. (2016). Morphological Complexity Influences Verb-Object Order in Swedish Sign Language. In *Proceedings of the Workshop on Computational Linguistics for Linguistic Complexity (CL4LC)*, pages 137–141

10. Bjerva, J. (2016). Byte-based language identification with deep convolutional networks. In *Proceedings of the Third Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial3)*, pages 119–125

11. Busger op Vollenbroek, M., Carlotto, T., Kreutz, T., Medvedeva, M., Pool, C., Bjerva, J., Haagsma, H., and Nissim, M. (2016). GronUP: Groningen user profiling. In *Proceedings of CLEF 2016*

12. Haagsma, H. and Bjerva, J. (2016). Detecting novel metaphor using selectional preference information. In *Proceedings of the Fourth Workshop on Metaphor in NLP*, pages 10–17

13. Bjerva, J., Bos, J., and Haagsma, H. (2016a). The Meaning Factory at SemEval-2016 Task 8: Producing AMRs with Boxer. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 1179–1184

14. Bjerva, J. and Praet, R. (2016). Rethinking intertextuality through a word-space and social network approach – the case of Cassiodorus. *Journal of Data Mining and Digital Humanities, accepted, in revision.*

15. Bos, J., Basile, V., Evang, K., Venhuizen, N. J., and Bjerva, J. (2017). *The Groningen Meaning Bank*, pages 463–496. Springer Netherlands, Dordrecht

16. Abzianidze, L., Bjerva, J., Evang, K., Haagsma, H., van Noord, R., Ludmann, P., Nguyen, D.-D., and Bos, J. (2017). The Parallel Meaning Bank: Towards a Multilingual Corpus of Translations Annotated with Compositional Meaning Representations. In *EACL*, pages 242–247

17. Östling, R. and Bjerva, J. (2017). SU-RUG at the CoNLL-SIGMORPHON 2017 shared task: Morphological Inflection with Attentional sequence-to-sequence models. In *Proceedings of the 2017 Meeting of SIGMORPHON*, Vancouver, Canada. Association for Computational Linguistics

18. Sjons, J., Hörberg, T., Östling, R., and Bjerva, J. (2017). Articulation rate in swedish child-directed speech increases as a function of the age of the child even when surprisal is controlled for. In *Proceedings of Interspeech 2017*, Stockholm, Sweden

19. Kulmizev, A., Blankers, B., Bjerva, J., Nissim, M., van Noord, G., Plank, B., and Wieling, M. (2017). The power of character n-grams in native language identification. In *Proceedings of Shared Task on NLI at BEA17*

20. Bjerva, J., Grigonytė, G., Östling, R., and Plank, B. (2017). Neural networks and spelling features for native language identification. In *Proceedings of Shared Task on NLI at BEA17*