Neural Cross-Lingual Transfer and Limited Annotated Data for Named Entity Recognition in Danish

Barbara Plank
Department of Computer Science
ITU, IT University of Copenhagen
Denmark
bplank@itu.dk

Abstract

Named Entity Recognition (NER) has greatly advanced by the introduction of deep neural architectures. However, the success of these methods depends on large amounts of training data. The scarcity of publicly-available human-labeled datasets has resulted in limited evaluation of existing NER systems, as is the case for Danish. This paper studies the effectiveness of cross-lingual transfer for Danish, evaluates its complementarity to limited gold data, and sheds light on performance of Danish NER.

1 Introduction

Named entity recognition is a key step for natural language understanding (NLU), and important for information extraction, relation extraction, question answering and even privacy protection. However, the scarcity of publicly-available human annotated datasets has resulted in a lack of evaluation for languages beyond a selected set (e.g., those covered in early shared tasks like Dutch, German, English, Spanish), despite the fact that NER tools exists or recently emerged for other languages. One such case is Danish, for which NER dates back as early as (Bick, 2004) and tools exist (Bick, 2004; Derczynski et al., 2014; Johannessen et al., 2005; Al-Rfou et al., 2013) but lack empirical evaluation.

Contemporarily, there exists a surge of interest in porting NLU components quickly and cheaply to new languages. This includes cross-lingual transfer methods that exploit resources from existing high-resource languages for zero-shot or few-shot learning. This line of research is blooming, particularly since the advent of neural NER, which holds the state of the art (Yadav and Bethard, 2018). However, neither neural tagging nor cross-lingual transfer has been explored for Danish NER, a gap we seek to fill in this paper.

Contributions We present a) publicly-available evaluation data to encourage research on Danish NER; b) an empirical comparison of two existing NER systems for Danish to a neural model; c) an empirical evaluation of learning an effective NER tagger for Danish via cross-lingual transfer paired with very little labeled data.

2 Approach

We investigate the following questions: **RQ1:** To what extent can we transfer a NER tagger to Danish from existing English resources? **RQ2:** How does cross-lingual transfer compare to annotating a very small amount of in-language data (zero-shot vs few-shot learning)? **RQ3:** How accurate are existing NER systems for Danish?

2.1 NER annotation

To answer these questions, we need gold annotated data. Access to existing resources is limited as they are not available online or behind a paywall. Therefore, we annotate NERs on top of publicly available data.[1]

In line with limited budget for annotation (Garrette and Baldridge, 2013), we add an annotation layer for Named Entities to the development and test sets of the Danish section of the Universal Dependencies (UD) treebank (Nivre et al., 2016; Johannsen et al., 2015). To answer RQ2, we further annotate a very small portion of the training data, i.e., the first 5,000 and 10,000 tokens. Examples are shown in Figure 1. Dataset statistics are provided in Table 2.

The Danish UD treebank (`Danish-DDT`) is a conversion of the Copenhagen Dependency Treebank (CDT). CDT (Kromann et al., 2003) consists of 5,512 sentences and 100k tokens, originating from the PAROLE-DK project (Bilgram and Keson, 1998). In contrast to original CDT and the PAROLE tokenization scheme, starting from the Danish UD has the advantage that it is closer to everyday language, as it splits tokens which were originally joined (such as 'i_alt').

We follow the CoNLL 2003 annotation guidelines (Tjong Kim Sang and De Meulder, 2003) and annotate proper names of four types: person (PER), location (LOC), organization (ORG) and miscellaneous (MISC). MISC contains for example names of products, drinks or film titles.

2.2 Cross-lingual transfer

We train a model on English (a medium and high resource setup, see details in Section 3) and transfer it to Danish, examining the following setups.

[1] `https://github.com/UniversalDependencies/UD_Danish-DDT`

B-LOC	O	O	O	O	O	O	O
Rom	blev	ikke	bygget	på	èn	dag	.
O	O	O	B-PER	O	O	B-MISC	I-MISC
vinyl	,	som	Elvis	indspillede	i	Sun	Records

Table 1: Example annotations.

	Evaluation		Training	
	DEV	TEST	TINY	SMALL
Sentences	564	565	272	604
Tokens	10,332	10,023	4,669	10,069
Types	3,640	3,424	1,918	3,525
TTR	0.35	0.34	0.41	0.35
Sent.w/ NE	220	226	96	206
Sent.w/ NE%	39%	34%	35%	34%
Entities	348	393	153	341

Table 2: Overview of the annotated Danish NER data. Around 35%-39% of the sentences contain NEs. TTR: type-token ratio.

- **Zero-shot**: Direct transfer of the English model via aligned bilingual embeddings.
- **In-Language**: Training the neural model on very small amounts of in-language Danish training data only. We test two setups, training on the tiny data alone; or with unsupervised transfer via word embedding initialization (+Poly).
- **Few-shot direct transfer**: Training the neural model on English and Danish jointly, including bilingual embeddings.
- **Few-shot fine-tuning**: Training the neural model first on English, and fine-tuning it on Danish. This examines whether fine-tuning is better than training the model from scratch on both.

3 Experiments

As source data, we use the English CoNLL 2003 NER dataset (Tjong Kim Sang and De Meulder, 2003) with BIO tagging.

We study two setups for the source side: a MEDIUM and LARGE source data setup. For LARGE we use the entire CoNLL 2003

		neural in-lang.		neural transfer		
	TnT	plain	+Poly	+MEDIUM src	+LARGE src	FINETUNE
zero-shot	—	—	—	58.29	61.18	—
TINY	37.48	36.17	56.05	67.14	67.49	62.07
SMALL	44.30	51.90	67.18	**70.82**	70.01	65.63

Table 3: F_1 score on the development set for low-resource training setups (none, tiny 5k or small 10k labeled Danish sentences). Transfer via multilingual embeddings from MEDIUM (3.2k sentences, 51k tokens) or LARGE English source data (14k sentences/203k tokens).

training data as starting point, which contains around 14,000 sentences and 200,000 tokens. To emulate a lower-resource setup, we consider a MEDIUM setup, for which we employ the development data from CoNLL 2003 as training data (3,250 sentences and 51,000 tokens). The CoNLL data contains a high density of entities (79-80% of the sentences) but is lexically less rich (TTR of 0.11-0.19), compared to our Danish annotated data (Table 2), which is orders of magnitudes smaller, lexical richer but less dense on entities.

Model and Evaluation We train a bilstm-CRF similar to (Xie et al., 2018; Johnson et al., 2019). As pre-trained word embeddings we use the Polyglot embeddings (Al-Rfou et al., 2013). The word embeddings dimensionality is 64. The remaining hyperparameters were determined on the English CoNLL data. The word LSTM size was set to 50. Character embeddings are 50-dimensional. The character LSTM is 50 dimensions. Dropout was set to 0.25. We use Stochastic Gradient Descent with a learning rate of 0.1 and early stopping. We use the evaluation script from the CoNLL shared task and report mean F1 score over three runs.

Cross-lingual mapping We map the existing Danish Polyglot embeddings to the English embedding space by using an unsupervised alignment method which does not require parallel data. In particular, we use character-identical words as seeds for the Procrustes rotation method introduced in `MUSE` (Conneau et al., 2017).

4 Results

Table 3 presents the main results. There are several take-aways.

Cross-lingual transfer is powerful (RQ1). Zero-shot learning reaches an F1 score of 58% in the MEDIUM setup, which outperforms training the neural tagger on very limited gold data (plain). Neural NER is better than traditional HMM-based tagging (TnT) (Brants, 2000) and greatly improves by unsupervised word embedding initialization (+Poly). It is noteworthy that zero-shot transfer benefits only to a limiting degree from more source data (F1 increases by 3% when training on all English CoNLL data).

To compare cross-lingual transfer to limited gold data (RQ2), we observe that training the neural system on the small amount of data together with Polyglot embeddings is close to the tiny-shot transfer setup. Few-shot learning greatly improves over zero-shot learning. The most beneficial way is to *add* the target data to the source, in comparison to fine-tuning. This shows that access to a tiny or small amount of training data is effective. Adding gold data with cross-lingual transfer is the best setup.

DEV	All	PER	LOC	ORG	MISC
Majority	44.8	61.8	0.0	0.0	—
DKIE	55.4	65.7	58.5	20.3	—
DKIE July 23	58.9	68.9	63.6	23.3	
Polyglot	64.5	73.7	**73.4**	36.8	—
Ours	**70.8**	**83.3**	71.8	**60.0**	23.9
TEST	All	PER	LOC	ORG	MISC
Polyglot	61.6	78.4	**69.7**	24.7	—
Ours	**66.0**	**86.6**	63.6	**42.5**	24.8

Table 4: F_1 score on the Danish dev set.

In both MEDIUM and LARGE setups are further gains obtained by adding TINY or SMALL amounts of Danish gold data. Interestingly, a) fine-tuning is less effective; b) it is better to transfer from a medium-sized setup than from the entire CoNLL source data.

Existing systems (RQ3) perform poorly (Table 4). Polyglot (Al-Rfou et al., 2013) is better than DKIE (Derczynski et al., 2014). Our best system is a cross-lingual transfer NER from MEDIUM source data paired with SMALL amounts of gold data. Per-Entity evaluation shows that ours outperforms Polyglot except for Location, which is consistent across evaluation data (Table 4). Overall we find that very little data paired with dense representations yields an effective NER quickly.

5 Related Work

Named Entity Recognition has a long history in NLP research. While interest in NER originally arose mostly from a question answering perspective, it developed into an independent task through the pioneering shared task organized by the Message Understanding Conference (MUC) (Grishman and Sundheim, 1996; Grishman, 1998). Since then, many shared task for NER have been organized, including CoNLL (Tjong Kim Sang and De Meulder, 2003) for newswire and WNUT for social media data (Baldwin et al., 2015). While Danish NER tools and data exists (Bick, 2004; Derczynski et al., 2014; Johannessen et al., 2005; Al-Rfou et al., 2013), there was a lack of reporting F1 scores. Supersense tagging, a task close to NER has received attention (Martínez Alonso et al., 2015).

The range of methods that have been proposed for NER is broad. Early methods focused on hand-crated rule-based methods with lexicons and orthographic features. They were followed by feature-engineering rich statistical approaches (Nadeau and Sekine, 2007). Since the advent of deep learning and the seminal work by (Collobert et al., 2011), state-of-the-art NER systems typically rely on feature-inferring encoder-decoder models that extract dense embeddings from word and subword embeddings, including affixes (Yadav and Bethard, 2018), often outperforming neural architectures that include lexicon information such as gazetteers.

Recently, there has been a surge of interest in cross-lingual transfer of NER models (Xie et al., 2018). This includes work on transfer between distant languages (Rahimi et al., 2019) and work on projecting from multiple source languages (Johnson et al., 2019).

6 Conclusions

We contribute to the transfer learning literature by providing a first study on the effectiveness of exploiting English NER data to boost Danish NER performance.[2] We presented a publicly-available evaluation dataset and compare our neural cross-lingual Danish NER tagger to existing systems. Our experiments show that a very small amount of in-language NER data pushes cross-lingual transfer, resulting in an effective Danish NER system.

[2]Available at: `https://github.com/ITUnlp/transfer_ner`

Acknowledgements

We kindly acknowledge the support of NVIDIA Corporation for the donation of the GPUs and Amazon for an Amazon Research Award.

References

Rami Al-Rfou, Bryan Perozzi, and Steven Skiena. 2013. Polyglot: Distributed word representations for multilingual NLP. In *ACL*.

Timothy Baldwin, Marie-Catherine de Marneffe, Bo Han, Young-Bum Kim, Alan Ritter, and Wei Xu. 2015. Shared tasks of the 2015 workshop on noisy user-generated text: Twitter lexical normalization and named entity recognition. In *Proceedings of the Workshop on Noisy User-generated Text*, pages 126–135.

Eckhard Bick. 2004. A named entity recognizer for danish. In *LREC*.

Thomas Bilgram and Britt Keson. 1998. The construction of a tagged danish corpus. In *Proceedings of the 11th Nordic Conference of Computational Linguistics (NODALIDA 1998)*, pages 129–139.

Thorsten Brants. 2000. Tnt: a statistical part-of-speech tagger. In *Proceedings of the sixth conference on Applied natural language processing*, pages 224–231. Association for Computational Linguistics.

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.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2017. Word translation without parallel data. *arXiv preprint arXiv:1710.04087*.

Leon Derczynski, Camilla Vilhelmsen Field, and Kenneth S Bøgh. 2014. Dkie: Open source information extraction for danish. In *Proceedings of the Demonstrations at the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 61–64.

Dan Garrette and Jason Baldridge. 2013. Learning a part-of-speech tagger from two hours of annotation. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 138–147.

Ralph Grishman. 1998. Research in information extraction: 1996-98. In *Proceedings of the TIPSTER Text Program: Phase III*, pages 57–60, Baltimore, Maryland, USA. Association for Computational Linguistics.

Ralph Grishman and Beth Sundheim. 1996. Message understanding conference- 6: A brief history. In *COLING 1996 Volume 1: The 16th International Conference on Computational Linguistics*.

Janne Bondi Johannessen, Kristin Hagen, Åsne Haaland, Andra Björk Jónsdottir, Anders Nøklestad, Dimitris Kokkinakis, Paul Meurer, Eckhard Bick, and Dorte Haltrup. 2005. Named entity recognition for the mainland scandinavian languages. *Literary and Linguistic Computing*, 20(1):91–102.

Anders Johannsen, Héctor Martínez Alonso, and Barbara Plank. 2015. Universal dependencies for danish. In *International Workshop on Treebanks and Linguistic Theories (TLT14)*, page 157.

Andrew Johnson, Penny Karanasou, Judith Gaspers, and Dietrich Klakow. 2019. Cross-lingual transfer learning for Japanese named entity recognition. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Industry Papers)*, Minneapolis - Minnesota. Association for Computational Linguistics.

Matthias T Kromann, Line Mikkelsen, and Stine Kern Lynge. 2003. Danish dependency treebank. In *Proc. TLT*, pages 217–220. Citeseer.

Héctor Martínez Alonso, Anders Johannsen, Sussi Olsen, Sanni Nimb, Nicolai Hartvig Sørensen, Anna Braasch, Anders Søgaard, and Bolette Sandford Pedersen. 2015. Supersense tagging for Danish. In *Proceedings of the 20th Nordic Conference of Computational Linguistics (NODALIDA 2015)*, Vilnius, Lithuania.

Linköping University Electronic Press, Sweden.

David Nadeau and Satoshi Sekine. 2007. A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 30(1):3–26.

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 Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 1659–1666.

Afshin Rahimi, Yuan Li, and Trevor Cohn. 2019. Multilingual ner transfer for low-resource languages. In *Proceedings of the 2019 Conference of the Chapter of the Association for Computational Linguistics*.

Erik F Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the conll-2003 shared task: Language-independent named entity recognition. *arXiv preprint cs/0306050*.

Jiateng Xie, Zhilin Yang, Graham Neubig, Noah A Smith, and Jaime Carbonell. 2018. Neural cross-lingual named entity recognition with minimal resources. *arXiv preprint arXiv:1808.09861*.

Vikas Yadav and Steven Bethard. 2018. A survey on recent advances in named entity recognition from deep learning models. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2145–2158.