DaLUKE: Strengthening Danish NLP Using Weak Knowledge-Enhancement

Søren Winkel Holm, Asger Laurits Schulz

December 16, 2021

Abstract

- Fokusér på problemet med NLP i middelressourcesprog generelt
- Dansk er et case study, hvor vi undersøger LUKE-arkitekturen og videnshancement generelt som mulig løsning
- Inkludér reklame for kode

1 Introduction

- Kort gennemgang af status i dansk NLP
- Kort introduktion af LUKE + nævn andre videnshancement-metoder
- Det mest generelle, motiverende spørgsmål: Hvad skal der til for at få det elegant LUKE-ideal til at virke med meget mere begrænset data?
- Mindre ressourcestærkt sprog =¿ mindre ressourcestærke anvendere, så fokus på lille model

2 Methods

- Data: Augmentering og forskellige kilder
- Hovedmodel: Hvad var forskellig fra DaLUKE
- Fremgangsmåde for træning
- Lille model: Fremgangsmåde for destillering/pruning
- Præsentér vores nye arkitektur til lavere entitetsdimension
- Præsentér NER-opgaven

Name	Data	Base Model
Control	Da. Wiki. w. entity links, Da. Gigaword	RoBERTa Base
Auto-annotated	Da. Wiki. w. entity links, Da. Gigaword with auto. entities	RoBERTa Base
Big model	Da. Wiki. w. entity links, Da. Gigaword	RoBERTa Large
Low entity dim.	Da. Wiki. w. entity links, Da. Gigaword	252-dimensional ent., RoBERTa Base

Table 1: Learning rate: $1.2 \cdot 10^{-4}$, batch size: 8160 examples

3 Results

- Prætræningsperformance med forskellige modeller
- NER-performance sammenlignet med dansk niveau

4 Discussion

- Prætrænings-eksperimenter: Data, arkitektur og transferlæring kontrolleret
- En smule undersøgelse af maskeret opgave/downstream fejl
- Konklusioner: AI-succes bliver bestemt af nogle andre underliggende parametre i højressourcedomænet end i den dataknappe situation