Word Embeddings vs Word Types for Sequence Labeling: the Curious Case of CV Parsing

Melanie Tosik†, Carsten L. Hansen‡, Gerard Goossen‡, and Mihai Rotaru‡

†University of Potsdam, ‡Textkernel B.V.

Objective

We explore new methods of improving upon Curriculum Vitæ (CV or resume) parsing for German. Our approach integrates word embeddings as features for a probabilistic sequence labeling model that relies on the Conditional Random Field (CRF) framework.

Introduction

Information extraction from CVs is one of the success stories of applying NLP in industry.

- Traditional approach: word types as features for CRF or HMM
- Challenge: high variance in data, many unknown words, poor generalization to CVs from new sectors
- Possible solution: annotate more CVs from these sectors – expensive
- Our approach: replace word types with continuous vector representations of words that can be induced from large, unlabeled data sets

CV Extraction Task

The task is to extract entities from sections in the CV (e.g. personal, experience, education, or skills).

- Personal section entities: *name, address, birthday, phone number, nationality*, and *email address.*
- Experience section entities: *job title, job duration* and *company and location.*

Model setup

- Conditional Random Fields (L-BFGS) for phrase extraction implemented in *CRFsuite*
- Word embeddings learned from 200k German CVs containing ∼145.5M tokens
- Generating word embeddings: *word2vec* (Skip-gram; default parameters; 150 dimensions)

Features

- Hand-crafted features: beginning / end of line, unknown words, digits, single chars, multi-spaces, capitalization, first / last token of line, most frequent words
- Word types: one-hot representation of all words occurring at least twice
- Word embeddings: generated w/ *word2vec*

Results

Model [%]	Personal Prec	Personal Rec	Personal F1	Experience Prec	Experience Rec	Experience F1
Hand-crafted features	94.5	94.0	94.3	84.7	69.8	76.4
Word types	94.7	91.2	92.3	85.3	67.7	75.3
Word embeddings	94.9	93.1	93.9	87.0	74.6	80.3
Word types + features	95.2	95.0	95.1	88.4	74.3	80.6
Word embeddings + features	96.3	95.7	**96.0**	89.6	79.2	**84.0**

Table 1: Macro-averaged precision, recall, and F1 on main test partition for Personal section and Experience Section.

Model [%]	Test set Prec	Test set Rec	Test set F1	OOS set Prec	OOS set Rec	OOS set F1
Word types + features	88.4	74.3	80.6	82.3	57.0	65.6
Word embeddings + features	89.6	79.2	84.0	83.3	**67.1**	73.8

Table 2: Experience phrase extraction on test partition and out-of-sample dataset.

Data

- Main set standard train, dev and test split
- Out-of-sample set to evaluate on intrinsically different data

	Main set Train	Main set Dev	Main set Test	OOS Test
#Docs	1010	233	214	25
#Pers	6736	1634	1388	n/a
#Exp	20687	4569	4410	356

Table 3: Distribution of documents and entities across the two data sets.

Conclusions

Word embeddings can be successfully applied to CV Parsing.

- Best results on both extraction tasks are obtained by the model which combines word embeddings and hand-crafted features, outperforming word types.
- Results on personal sections show that hand-crafted features outperform word types and word embeddings alone.
- Improvements are consistent throughout different sections of target documents.
- Effect of the word embeddings is strongest on semi-structured, out-of-sample data.
- Best-performing word embeddings are generated from a large set of German CVs.

textkernel