Název práce: Rozpoznávání pojmenovaných entit pomocí neuronových sítí

Autor: Jana Straková

Ústav: Ústav formální a aplikované lingvistiky

Vedoucí doktorské práce: prof. RNDr. Jan Hajič, Dr., Ústav formální a aplikované lingvistiky

Abstrakt: Obor rozpoznávání pojmenovaných entit v češtině (tj. úkol automaticky identifikovat a klasifikovat významné části textu, jako například jména lidí, míst a organizací) se významně rozvinul po vydání českého korpusu pojmenovaných entit, Czech Named Entity Corpus (CNEC). Tato doktorská práce předkládá autorské výsledky v oblasti rozpoznávání pojmenovaných entit, zejména v češtině. Publikuje práci a výzkum provedený v průběhu přípravy CNEC a později během jeho evaluace. Dále shrnuje autorské výsledky, které představují nejlepší známé výsledky v rozpoznávání českých pojmenovaných entit. Na základě jednoduché neuronové sítě s výstupní funkcí softmax a standardní sadou klasifikačních rysů je popsána metodologie a výsledky, ze kterých později vznikl otevřený software pro rozpoznávání pojmenovaných entit, NameTag. Doktorská práce je zakončena popisem rozpoznávače založeném na rekurentních neuronových sítích s embeddingy slov a embeddingy založenými na znacích, které představují výsledky současného výzkumu v oblasti neuronových sítí. Rozpoznávač nevyžaduje tvorbu klasifikačních rysů a dosahuje v současné době nejlepších známých výsledků v oblasti rozpoznávání pojmenovaných entit v češtině.

Klíčová slova: rozpoznávání pojmenovaných entit, Czech Named Entity Corpus, neuronové sítě, rekurentní neuronové sítě, softmax, embeddingy slov, embeddingy založené na znacích