Tulajdonnév-felismerés német nyelven gépi tanulással

Ragács Attila

A szakdolgozatomban a természetesnyelv-feldolgozás területén belüli információkinyerés egy részfeladatával, a tulajdonnév-felismeréssel foglalkozom német nyelven. A dolgozat első fejezetében egy általános bevezető olvasható erről a feladatról az, illetve ismertetem a használt modellt a fontosabb paraméterek bemutatásával. Ez egy Conditional Random Field alapú modell, amely elérhető német nyelven is és ezek közül az egyik legújabb, bár már így is 10 éves. Ennek előnye a manapság gyakrabban használt neurális hálókkal szemben, hogy lehetséges a feature engineering, azaz meg tudjuk szabni, milyen paramétereket használjon a modell, míg egy neurális háló maga generálná ezeket. Így a fejlesztés során felmerülő konkrétabb problémák is megoldhatóvá válnak.

Mivel ennek a modellnek egyes összetevői elavultak, nagy hangsúlyt fektettem ezek helyettesítésére újabb módszerekkel a második fejezetben. Az ehhez felhasznált korpusz a 2014es GermEval Shared Task-hoz szolgáltatott adat, ami egy internetes hírportálokról és a wikipédiáról származó német szöveg több, mint 590000 szóval. Az érintett módszerek a Gazetteer listák használata, amelyek ismert tulajdonneveket tartalmaznak és újabb szóbeágyazások felhasználása hasonlósági osztályok létrehozásához.

A különböző kísérletek kiértékelései a harmadik fejezetben olvashatók. A legpontosabb modellt a fejlesztés után kipróbáltam a GermEval Shared Task résztvevőinek is biztosított teszthalmazon, a kapott eredmény így összevethető a versenyzőkkel.