

Miből lesz a robot MÁV-pénztáros?

Nemeskey Dávid, Recski Gábor, Zséder Attila

MTA SZTAKI

Nyelvtechnológiai Kutatócsoport

e-mail: ndavid,recski,zseder@sztaki.hu

A MÁV-pénztáros demonstrációban a felhasználó különféle vonatjegyeket vagy menetrendi információkat kérhet a programtól természetes nyelven. A rendszer két fő összetevőből áll: az egyik egy keretrendszer, amely lehetővé teszi, hogy a rendszer különböző komponensei akár más-más gépeken, aszinkron módon fus- sanak, míg a másik a tényleges szemantikai kód. A teljes program pythonban íródott, viszont az egyes komponensek bármilyen nyelvűek lehetnek.

A keretrendszer egy egyszerű eseményvezérelt architektúrát valósít meg, amelybe tetszőleges funkciójú összetevőket (*plugin*-eket) kapcsolhatunk be. Ezekről a komponensektől csak annyit követelünk meg, hogy a más komponen- sektől a keretrendszeren át érkező üzenetekre – melyek bármilyen python objek- tumot tartalmazhatnak – valamely más komponens számára értelmes választ ad- janak. A keretrendszer segítségével rugalmasan működhetnek együtt a különböző elemzők, következtetőrendszerek vagy bármilyen külső erőforrás.

A tényleges nyelvi megértést végző szemantikai modul is egy a keretrend- szerbe kapcsolódó *plugin*-ek közül. A *MachineCore* nevű komponens gondosko- dik a felhasználói üzenetek belső reprezentálásáról, a háttértudások tudásbázisba építéséről, a mondatok szintaktikai és szemantikai elemzéséről és a következ- tetésről. A felhasználtól érkező szöveges (nem hangalapú) üzenetek feldol- gozásához a szemantikai magnak morfológiailag elemzett és fő mondattani össze- tevőkre bontott adatokra van szüksége. Ezeket az adatokat a *hunmorph* [1] és a *hunchunk* [2] eszközökkel nyerjük ki. A jelen alkalmazás céljaira ezután egy egy- szerű, mindössze néhány reguláris kifejezésből álló dátum- és időpontfelismerő komponens is lefuttatunk.

Az így elkészült adat ezután készen áll a szemantikai feldolgozásra. Az elemzés és következtetés motorja a *Spreading Activation (SA)*, melynek alap- egységei az ún. *gépek* (machine, definícióját l. [3], 10. fejezet), majd a külvilág, pontosabban bizonyos, külvilággal érintkező *plugin*-ek felé az eredményeket att- ribútum-érték mátrixokon (AVM) keresztül kommunikálja.

Az egyes szavak jelentésének leírására a kutatócsoport kidolgozott egy de- finíciós szintaxist [4] és fejlesztett egy parszert, mely ezt a tudást ugyanabban a gépalapú reprezentációban ábrázolja, amellyel a rendszerünk dolgozik. Így egy- szerű szöveges fájlokkal írhatjuk le a fogalmak jelentését. Ez a modul lehetőséget teremt arra is, hogy a szavak fogalmi szintű leírásához különböző nyelvű szavak, kifejezések is kapcsolódhassanak. A demóhoz létrehoztunk egy MÁV-jegyekkel és vonatokkal kapcsolatos rövid leíró fájlt, amely csak a témakörrel kapcsolatos szavakat definiálja.

Az SA a konstrukciós nyelvtanból ismert konstrukciókat futtatja le az előfeldolgozott szövegen – a szavak között található kapcsolatokat ily módon ismerjük fel. Ilyen konstrukció például a főnévi csoportokon belüli jelzők főnévhez való kapcsolása vagy az igei vonzatkeretek kitöltése. Ez utóbbihoz még arra is szükség volt, hogy az igék definiálásakor helyet hagyjunk az egyes vonzatoknak, és egy külön fájlban írjuk le, hogy mely vonzatok általában milyen szerepet töltenek be a mondatban.

Az SA másik fontos funkciója, hogy teljes futása során számontartja egy ún. *aktív tömbben*, mely szavakkal, illetve gépekkel dolgozunk. Így ha olyan szavakat lát az aktív gépek között, melyeknek ismerjük a jelentését (mivel definiáltuk őket), akkor a jelentésüket is beírjuk ebbe az aktív tömbbe, vagy fordítva: ha úgy látja, hogy már létrejött egy gépekből álló struktúra, amely épp egy szó jelentésének felel meg, akkor a szót is felvesszük. Pl. a *menetrend* és a *vonat* szavak a mi rendszerünkben életre hívják az *elvira* gépet, vagy amikor a rendszer meglátja a *megy* igét, akkor helyettesíti azzal a struktúrával, amely leírja, hogy ez az ige mit jelent valójában.

Végül az SA gondoskodik arról is, hogy az AVM-ek a futás során kitöltődjenek azokkal az értékekkel, amelyeket a külső erőforrások megkövetelnek. Egyszerre több AVM is kitöltődhet, mert a mondat egyes szavaiból még nem tudhatjuk, hogy a kér(d)és mire vonatkozott, így arra is fel kell készülni, hogy a felhasználó menetjegyet vagy helyjegyet kért, esetleg mindkettőt, de arra is, hogy valójában a menetrendről érdeklődött. Az SA során létrejövő aktivációk sorozatából, a gépek és AVM-ek versengéséből valamelyik (esetleg több) győztesen kerül ki, és az kerül eredményként a felhasználó elé. Ez jelen esetben egy vagy két nyomtatott jegyet, esetleg egy külön böngészőablakban megjelenített menetrendi információt jelent.

Hivatkozások

1. Trón, V., Gyepesi, G., Halácsy, P., Kornai, A., Németh, L., Varga, D.: Hunmorph: open source word analysis. In Jansche, M., ed.: Proc. ACL 2005 Software Workshop. ACL, Ann Arbor (2005) 77–85
2. Recski, G., Varga, D., Zséder, A., Kornai, A.: Főnévi csoportok azonosítása magyar-angol párhuzamos korpuszban. In: VI. Magyar Számítógépes Nyelvészeti Konferencia. (2009)
3. Eilenberg, S.: Automata, Languages, and Machines. Volume A. Academic Press (1974)
4. Kornai, A., Makrai, M.: A 4lang fogalmi szótár. In: IX. Magyar Számítógépes Nyelvészeti Konferencia. (2013)