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, ami lehetővé teszi, hogy a rendszer különböző komponensei akár más-más gépeken, aszinkron módon fussanak, 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. Ezektől a komponensektől csak annyit követelünk meg, hogy a más komponensektől a keretrendszeren át érkező üzenetekre – melyek bármilyen python objemktumot tartalmazhatnak – valamely más komponens számára értelmes választ adjanak. A keretrendszer segítségével rugalmasan működhetnek együtt a különböző elemzők, következtetőrendszerek, vagy bármilyen külső erőforrás.

A tényleges nyelvi megértést végző szemantikai modul is egy a keretrendszerbe kapocsolódó plugin-ek közül. A `MachineCore` nevű komponens gondoskodik a felhasználói üzenetek belső reprezentálásáról, a háttértudások tudásbázisba építéséről, a mondatok szintaktikai és szemantikai elemzéséről és a következtetésről. A felhasználótól érkező szöveges (nem hangalapú) üzenetek feldolgozásához a szemantikai magnak morfológiailag elemzett és fő mondattani összetevőkre bontott adatokra van szüksége. Ezeket az adatokat a `hunmorph` (Trón 2005) és a `hunchunk` (Recski et al. 2009) eszközökkel nyerjük ki. A jelen alkalmazás céljaira ezután egy egyszerű, mindössze néhány reguláris kifejezésből álló dátum- és időpontfelismerő komponenst is lefuttatunk.

Az így elkészült adat ezután készen áll a szemantikai feldolgozásra. Az elemzés és következtetés motorja a *Spreading Activation (SA)*, melynek alapegységei az ún. *gépek* (machine, definícióját ld. Eilenberg 1974 ch. 10), majd a külvilág, pontosabbab bizonyos, külvilággal érintkező plugin-ek felé az eredményeket attribútum-érték mátrixokon (AVM) keresztül kommunikálja.

Az egyes szavak jelentésének leírására a kutatócsoport kidolgozott egy definíciós szintaxist (Kornai és Makrai 2012) és fejlesztett egy parszert, mely ezt a tudást ugyanabban a gép-alapú reprezentációban ábrázolja, amivel a rendszerünk dolgozik. Így egyszerű 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, ami 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, ami épp egy szó jelentésének felel meg, akkor a szót is felvesszük. Pl. a *menetrend* és a *vonat* szavak a mi rendszerünkben életrehívják az *elvira* gépet, vagy amikor a rendszer meglátja a *megy* igét, akkor helyettesíti azzal a struktúrával, ami leírja, hogy ez az ige mit jelent valójaban.

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.