Szótáralapú kémiai NE-felismerő rendszer

Nyilas Sándor, Németh Gábor, Almási Attila

Szegedi Tudományegyetem, Informatikai Tanszékcsoport
Szeged, Árpád tér 2.
{nyilasster, nemeth.gabor3, vizipal}@gmail.com

Kivonat: A MASZEKER projekt szabadalmakon futó szemantikus keresőrendszer kifejlesztését célozta meg, melynek az orvostudományi és kémiai szabadalmak esetében egyik lényegi lépése a kémiai névelemek felismerése. Ehhez szükség volt egy szótárfájl létrehozására, mivel a névelemeket jelölő program nem elemzi szemantikusan a mondatok szavait, és nem az alapján dönt, hogy melyik szó kémiai névelem és melyik nem. A szótárfájl, amely soronként elkülönített szavakból áll, tartalmazza a kémiai névelemeket. Ennek a szótárfájlnak a rendszeres frissítése és karbantartása szükséges ahhoz, hogy a program minden kémiai névelemet fel tudjon ismerni.

1 A szótárfájl előállításáról

A kémiai vegyületneveket tartalmazó szótárfájlt az Environmental Chemistry oldalról[1] gyűjtöttük ki, mely soronként egy vegyületnevet és egy tabulátorral elválasztott egy- vagy kétjegyű előfordulási számot tartalmaz, amelyet eddig figyelmen kívül hagytunk.

1.1 Az eredeti szótárfájl

A szótárfájlból több névelem (NE) is hiányzott (pl. *sodium*), ezért szükség volt a szólista bővítésére. Mivel más kémiai névelem-adatbázis nem állt rendelkezésünkre, de feltételeztük, hogy az összetett kémiai NE-k tartalmazzák a hiányzó, elemibb NE-ket (pl. a (*9-Octadecenoic acid (Z)-, iron(3+) salt*) tartalmazza az *iron*-t), ezért az összetett NE-k felbontása mellett döntöttünk. Ehhez szükség volt egy bővítmény létrehozására, melyet az alábbiak szerint hajtottunk végre: az eredeti szótárfájlról két másolatot készítettünk, melyeken a következő változtatásokat hajtottuk végre:

- kis- és nagybetűket nem megkülönböztetve ábécérendbe rendeztük az NE-ket,
- minden sor végéről a tabulátort és az utána szereplő számot töröltük,
- az 1. másolatban minden nem szám- és betűkaraktert kicseréltünk szóköz karakterre, az összes szóközt sortörésre cseréltük, s az így kapott szavakat a következő algoritmussal szűrtük:
  - a tördelésnél keletkezett felesleges szavak közül kivettük a duplikátumokat;
  - a könnyebb kezelhetőség kedvéért a dupla sortöréseket is töröltük;

[1] http://environmentalchemistry.com/yogi/chemicals/

• az 1. másolatot hozzáadtuk a 2.-hoz, majd újból szűrtük a dupla sortöréseket és a duplikátumokat;

• a 2. másolatból a felesleges szavakat kézileg eltávolítottuk. E 2. másolatot a beillesztést követően az első verziójú szótárfájlnak tekintjük.

Mivel a program nem tett különbséget a rövidítéseknél a kis- és nagybetűk között, helytelenül került feljelölésre az *at* mint prepozíció és helyesen az *At* mint *Astatine*. A hiba kiküszöbölése érdekében a szótárfájlt kettévágtuk a legfeljebb három, illetve az annál több karakterből álló szavakra. Ezentúl két lista létezik, melyeket együttesen nevezünk szótárfájlnak. Az első listafájlt (három és annál kevesebb karakterből álló szavak), kis- és nagybetűt megkülönböztetve, a másikat (háromnál több karakterekből álló NE-k) továbbra is kis- és nagybetűt nem megkülönböztetve vizsgáljuk. A két listafájl jelenleg 81959 vegyületnevet tartalmaz.

2 A program működése

A szótárfájlt és a vizsgált szabadalmat betöltjük a programba, majd megvizsgáljuk, hogy a szótárfájlból soronként betöltött NE-k megtalálhatók-e a vizsgált szabadalomban. Az eredmények javításához feszítő-szűrő szabályrendszert alkalmazunk.

2.1 Feszítés

Mikor a szó átadódik a feszítő algoritmusnak, már biztosak lehetünk abban, hogy vegyületnevet találtunk. Először a kezdő-, majd a záróindex értékét változtatja a kód, értelemszerűen a kezdőindexet csökkentve, a záróindexet pedig növelve azért, hogy a vegyületnév-töredék indexeit ráfeszítse az egész vegyületnévre. A folyamat akkor áll meg, ha a program megfelelő karakterpárt talál. Ezek a karakterpárok a vegyületnév elejét vagy végét jelzik. Ha egy szó után szünet van, és a következő karakter valamilyen betű vagy szám, akkor ott a vége a vegyületnévnek. Ugyanez igaz a szó elejével kapcsolatban is. Ezen kívül a következő kritériumok esetében áll meg a feszítés, és dönt úgy, hogy a vegyületnévnek az adott helyen eleje vagy vége van:

eleje:

- <bármilyen karakter és/vagy szám> és szóköz
- <írásjelek (, . ! ? (stb...)> és szóköz
- <sorvége, kocsivissza, \0 jelek> és szóköz
- pontosvessző

vége:

- szóköz és <bármilyen karakter és/vagy szám>
- szóköz és <írásjelek (, . ! ?) stb...)>
- <szóköz, írásjelek> és <sorvége, kocsivissza, \0 jelek>

2.2 Szűrés

Mikor a feszítés befejeződik, átadja a találatok kezdő- és végindexeinek listáját a szűrő algoritmusnak, mely azért felel, hogy a találatok közül mindig csak a „legbővebb" legyen feljelölve (pl. csak a (*Threitol 1,4-bis (methanesulfonate*)) és ne a *Threitol* és *methanesulfonate* külön-külön), ezért egy algoritmussal a kezdő és az ahhoz tartozó végindexeket rendezi. Ezután a program addig hasonlítja össze a találatokat, ameddig a szűrő már nem változtat a találati listán.

A szűrő egyik funkciója az, hogy megvizsgálja, hogy az n-edik találatnak a k kezdőpontjához és v végpontjához képest hol helyezkedik el az n+1-edik találatnak a k2 kezdőpontja és a v2 végpontja. A sorba rendezés miatt alapfeltevés, hogy k<k2:

- ha v<k2, akkor nem változik
- ha v>=k2, de v<=v2, akkor v-t egyenlővé tesszük v2-vel és az n+1 találatot töröljük
- ha v>=k2, de v>v2, akkor v megtartja az értékét, és az n+1 találatot töröljük.

Ha két találat közt csak egy karakter távolság van, akkor a kettőt egynek veszi, és megkapja a két találat legkisebb kezdőértéket és a legnagyobb végértéket.

3 A névelemek annotációja során felmerült problémákról

A főigénypontokban szereplő NE-ket három csoportba rendeztük: 1) kémiai elemek (nitrogén), elemcsoportok (halogének), vegyületek (Na_2O) stb.; 2) általános anyagnevek (só), vegyületfajták (szénhidrát) stb.; 3) konkrét betegségek (Alzheimer-kór), betegségcsoportok (immunhiányos betegségek) és tünetek (másnaposság).

Néhány gyakoribb hibatípus [1]:

- A program nem különíti el a NE-k főnévi és jelzői használatát: pl. *antibiotic* – az angolban főnév és melléknév is lehet; főnévi használatban jelöltük csak.
- Az előforduló helyesírási hibák miatt a program nem megfelelően szegmentál bizonyos elemeket: pl. *…alkarylamino, fluoro, chloro, bromo iodo and trifluoromethyl…* – két, egyébként külön jelölendő NE-t egynek vett; az annotáció a szándékolt tartalomnak megfelelően történt.
- Szófaji problémák:
    1. *water-soluble*, *wax-like* kifejezések a magyarban nem NE-k – nem jelöltük;
    2. *carboxylic*, *enantiomeric* jelzők – nem jelöltük;
    3. *O-glycosidically* határozószó – nem jelöltük.

4 Eredmények

A névelem-felismerő program találati pontossága a fejlesztésekkel rohamosan nőtt. Ezt egy segédrendszerrel teszteltük, amely összehasonlította a kézzel annotált 313 dokumentumot az automatikusan feljelöltekkel. A két feljelölés között jelentős számbeli különbség mutatkozott: pl. a *salt* a kézi feljelölés alapján 17, a gépi feljelölés szerint pedig 43 alkalommal fordul elő NE-ként.

Eddig négy különböző programverzió készült, a negyedik még tesztfázisban van.

- ***verzió 0.1***: szűrés és feszítés nélkül a program a tesztelés során csak ~ 6,5%-ot ért el. A programban nem volt szűrési rendszer: a gépi megjelölés jóval több, mint a kézi, mert összetettebb vegyületnevek esetén az elemibb NE is feljelölésre kerültek, és az is annotációnak számított.
- ***verzió 0.2***: a szűrőrendszer beiktatása után már 70%-os teljesítményt ért el a program. A gépi annotációk száma számottevően kisebb volt, mint a kézzel feljelölteké. Szükség volt a szótárfájl bővítésére és egyben szűrésére is, mert még ~ 2000 feljelölés fölösleges, illetve helytelen volt.
- ***verzió 0.3***: a szótárfájl bővítése és szűrése után az F-mérték 90.13%-ra javult.
- ***verzió 0.4***: finomítottunk az annotálási elveken és három kategóriát vettünk fel: ***speciális NE-k***, ***általános NE-k***, ***betegségek*** (l. 3. fejezet). A tesztanyagban a kézzel jelölt annotációkat a fentieknek megfelelően módosítottuk. A szabályrendszer átalakítására nem volt szükség, csupán a program beolvasási és osztályba sorolási rendszerén kellett változtatni. Az NE-k és betegségek megkülönböztetése érdekében a szótárfájlt – a korábbi kettő helyett – négyfelé vágtuk, majd beillesztettük a programba a négy fájl beolvasását. Amikor a program egy NE-t felismer, feljelöli, és besorolja a megfelelő osztályba. Az, hogy melyik osztályba kerül egy NE, kizárólag attól függ, hogy a kifejezés melyik szótárfájlból származik. A betegségek szótárfájlt bővíteni kell egy másik adatbázis[2] segítségével. A kézi jelölés módosítása után az F-mérték 95.25%-ra javult.

1. táblázat

	verzió 0.1	verzió 0.2	verzió 0.3	verzió 0.4
Gépi NE-k száma	17 779	9 799	11 407	10 373
Kézi NE-k száma	10 874	10 874	10 874	11 355
Helyes NE-k száma:	932	7 306	10 041	10 348
Precision / Recall:	8.57 / 5.24	67.18 / 74.56	92.33 / 88.02	91.13 / 99.75
F:	6.50	70.68	90.13	95.25

[2] http://www.who.int/classifications/icd/en/

Köszönetnyilvánítás

A kutatást – részben – a MASZEKER kódnevű projekt keretében az NKTH támogatta.

Bibliográfia

1. Vincze V., Nagy Á., Klausz Á., Almási A., Kiss M.: Nyelvészeti problémák a szabadalmak feldolgozásában. In: Tanács A., Vincze V. (szerk.): VII. Magyar Számítógépes Nyelvészeti Konferencia. Szeged, Szegedi Tudományegyetem (2010) 168–179