Főnevek a Magyar WordNetben

Hatvani Csaba[1], Kocsor András[1], Miháltz Márton[2],
Szarvas György[1], Szécsi Katalin[2]

[1] Szegedi Tudományegyetem, Informatikai Tanszékcsoport
6720 Szeged, Árpád tér 2.
{hacso, kocsor, szarvas}@inf.u-szeged.hu
[2] MorphoLogic Kft., 1126 Budapest, Orbánhegyi út 5.
{mihaltz, szecsi}@morphologic.hu

Kivonat: A Magyar WordNet, a többnyelvű BalkaNet/EuroWordNet rendszerekhez kapcsolódó magyar nyelvű wordnet-ontológia fejlesztése három intézmény részvételével 2005-ben indult egy hároméves pályázati projekt keretében (GVOP-AKF-2004-3.1.1). A tanulmány a Magyar WordNet ontológia teljes főnévi részének felépítését mutatja be. Részletesen leírjuk azokat a módszertani elveket és bővítési lépéseket, melyeknek segítségével kialakítottuk a jelenleg mintegy 20.000 főnévi csomópontból (synsetből) álló lexikális fogalmi hálózatot. Ismertetjük továbbá az általunk követett bővítési módszertan minőségi vizsgálatának módszereit és eredményeit is.

1. Bevezetés

A Magyar WordNet (HuWN) ontológia ([1], [4]) létrehozásával a magyar nyelv bekapcsolódott a Princeton WordNet (PWN) ([3]) architektúrájára épülő, EuroWordNet (EWN) ([6]) és a BalkaNet (BN) ([2], [5]) többnyelvű ontológiarendszerekbe. A napjainkra a legtöbb nagy európai nyelvet tömörítő kezdeményezéshez való csatlakozással számos nyelvtechnológiai probléma, mint például a gépi fordítás előtt is új távlatok nyílhatnak meg. Az alábbiakban bemutatjuk a Magyar WordNet főnévi állományának főbb jellemzőit, az ontológia kialakítása során alkalmazott módszertani elveket, megoldásokat, valamint egy vizsgálat eredményeit, mellyel bővítési módszertanunk minőségét vizsgáltuk.

2. Módszertani elvek

A legfőbb célunk a Magyar Wordnet ontológia kialakítása során olyan felső szintű, általános nyelvi tudást megjelenítő fogalmak felvétele volt, melyekhez a későbbiek során könnyen kapcsolhatók kisebb, domain-specifikus fogalomi hálók (mint pl. a későbbiekben létrehozandó gazdasági szókincset leíró ontológia).

A *fogalmi sűrűség elvének* nevezett, gyakorlati szempontból lényegesnek tartott elv alatt azt a törekvést értjük, hogy a létrehozott Magyar WordNet ontológiában az összes olyan fogalom szerepeljen, ami egy másik, az ontológiába felvett csomópont által szimbolizált jelentést magában foglal, annál általánosabb. A fogalmi sűrűség kritérium teljesíthető, ha minden bővítési szakasz után képezzük a főnévi hálózatnak az angol wordnet hipernima-relációi szerinti lezártját, és az esetlegesen hiányzó synsetekkel bővítjük azt.

3. A főnévi hálózat bővítése

A munka korábbi szakaszában elkészítettük a BalkaNet közös fogalmi készletének, a BalkaNet Concept Set (BCS) synsetjeinek magyar reprezentációját. A BCS 8516 synsetje (köztük 5896 főnévi) tartalmazza a EWN projekt 8 nyelvében, valamint a BN további 5 nyelvében legfontosabbnak tartott, az ontológiai hierarchia szempontjából alapvetőnek számító fogalmakat, melyeket minden nyelven implementáltak, biztosítva ezzel a nyelvek közötti minimális átjárhatóságot A munkáról bővebben lásd [4].

Ezt a főnévi magontológiát a bővítettük ki 19.500 tételesre, az alábbiakban részletesen ismertetjük ennek menetét.

3.2 Lokális alapfogalmak

A EWN és BN projektekben alkalmazott metodológiát követve először elkészítettük a lokális főnévi alapfogalmakat (Local Base Concepts, LBC), vagyis a magontológiába tartozó, de a közös halmazban (BCS) már nem szereplő fogalmakat reprezentáló synseteket. Ehhez korpuszstatisztikai módszereket alkalmaztunk: a Magyar Nemzeti Szövegtár főnévi gyakorisági listáját, illetve az Értelmező Kéziszótár (EKSZ) egy elektronikus változatában a főnévi definíciók szemantikai elemzéseit. A leggyakoribb MNSZ-ben szereplő, valamint az EKSZ definíciókban leggyakrabban genus proximum-ként szereplő főneveknek heurisztikusan megállapítottuk a leggyakoribb jelentéseit. A magyarra lefordított BCS-ben (BCSHu) felvett EKSZ azonosítók segítségével meghatároztuk ezek közül azokat a fogalmakat, amelyekhez még nem létezett synset a BCSHu-ban. Ezek alapján 250 szójelentéshez vettünk fel új synseteket, illetve EKSZ hivatkozásokat létező, megegyező jelentésű synsetekhez. A magyar főnévi mag-ontológia ezek után nagy valószínűséggel tartalmazza a BalkaNet/EuroWordnet alapfogalmain túl a magyar nyelvben legfontosabb kiinduló jelentéseket is.

3.2 Koncentrikus bővítés

A BCS és a magyar nyelvre fontosnak ítélt LBC-k elkészítése után a főnévi fogalmi háló bővítése során az angol nyelvre meglévő fogalmi hálót tekintettük kiindulási alapnak. Célszerű választás volt minden bővítési szakaszban a már elkészült magyar hálózat angol nyelvű képéből közvetlenül elérhető csomópontok közül válogatni. Így egyrészt az angol oldalról automatikusan teljesült a fogalmi sűrűség elve, másrészt –

lévén a magontológiából indultunk ki – többnyire általánosabb, a hipernima-hierarchiában magasabb szinten levő fogalmak kerültek a jelöltek közé.

Mivel a felsőbb szintű, absztraktabb fogalmaknak tipikusan egynél több hiponimájuk van, a fejlesztés során végig 30-40 ezer közvetlenül elérhető angol fogalom alkotta a jelöltek halmazát. Egy munkafázisban általában néhány ezer csomóponttal bővítettük az ontológiát, így a jelöltek közül szükségessé vált a céljainknak legmegfelelőbbek kiválasztása. A rangsoroláshoz négy, egymással nem feltétlenül összhangban lévő szempontot használtunk:

Fordíthatóság: A fogalomjelölt előkészíthető volt a korábban kidolgozott automatikus fordítási heurisztikákkal ([1], [4]). Ebben az esetben a synset létrehozása magyar nyelven egyszerűbben és gyorsabban elvégezhető volt, hiszen egy vagy több literál azonnal az annotátor rendelkezésére állt magyarul is.
Gyakoriság: A fogalomjelölt literáljai angol nyelvű korpuszokban (British National Corpus, American National Corpus First Release, SemCor) gyakran fordultak elő. Ez legtöbbször azt jelzi, hogy az adott szó a kommunikációban gyakran előkerülő fogalmat takar, azaz a felvétele a Magyar WordNetbe indokolt.
Nyelvek közötti átfedés: A megfelelő synset az angolon kívül sok más nyelvű wordnetben is előfordul. Ilyen synsetek felvételével egyrészt maximalizálhatjuk a Magyar Wordnet más nyelvekkel való átfedését, ami pl. fordítási feladatokhoz előnyös lehet, másrészt olyan fogalmakat veszünk fel, melyeket több más kutatócsoport is fontosnak ítélt, hiszen felvette az adott nyelv ontológiájába.
Relációk száma: A bővítés első szakaszában figyelembe vettük, hogy az adott synset hány új fogalmat tesz elérhetővé. A sok hiponimával rendelkező gyűjtőfogalmak felvétele célszerű volt, mert növelte a későbbi bővítési fázisokhoz a beválasztható synsetek számát.

Minden lépésben a gyakoriság és a nyelvek közötti átfedés (továbbá első körben a relációk száma) alapján rangsorolt fogalmakat választottunk ki a magyar ontológia bővítésére úgy, hogy az automatikus fordítással rendelkező synsetekből 3-4-szer többet vettünk fel, mint a fordítással nem rendelkező jelöltekből. Az úgynevezett koncentrikus bővítés során az első fázisban 2705, a második szakaszban 4385, majd végül további 800 fogalmat dolgoztunk ki.

3.3 Teljes részfák felvétele

A főnévi synset-állomány iteratív koncentrikusan kifelé terjeszkedő bővítése mellett kiválasztottunk néhány speciális területet, ahol minden PWN-ben ismert fogalmat lefordítottunk, vagyis az adott fogalmi körhöz tartozó teljes hipernima-részfákat átemeltünk. Ezzel az ontológia általános enciklopédikus tudását igyekeztünk az adott területeken teljessé tenni.

A következő fogalmi körökre alkalmaztuk ezt az eljárást:

- földrajzi nevek (országok, fővárosok, nagyvárosok, országon belüli (tag)államok (pl. USA államok), földrajzi területek (geopolitikai régiók), egyéb régiók, földrészek, "víznevek" (tavak, folyók, tengerek, öblök, óceánok, vízesések), hegycsúcsok, szigetek);
- emberi nyelvek (és nyelvcsaládok);

- embercsoportok (népek, ill. egy-egy régió lakosai);
- a világ országainak pénzegységei.

Összesen 3,200 synsetet vettünk fel ezen kritériumok alapján.

A gazdasági szakontológia számára ezzel a módszerrel felvettünk további 940 fogalmat a gazdaság, vállalkozás és a kereskedelem szakterületéről is.

3.4 Domain synsetek

A PWN 2.0-s verziójában bevezetett domain-relációk segítségével reprezentálni lehet a korábbi szemantikai relációkkal (főneveknél: hipernímia, holonímia, antonímia) ki nem fejezhető kapcsolatokat, illetve szerepük lefedi a hagyományos (értelmező) szótárak tárgyterületi, nyelvhasználati minősítő kódjainak funkcióját is. A reláció egy *domain synset* mint összefogó kategória és egy vagy több *domain term synset* mint elem között ábrázol tematikus/nyelvhasználati kapcsolatot. A domain relációnak három fajtája van: tartalmi/tematikus/szemantikai kapcsolatot kifejező (category), térbeli kapcsolatot kifejező (region), valamint nyelvhasználati kategóriát kifejező (usage).

Annak érdekében, hogy a HuWN PWN-re támaszkodó későbbi bővítése során akadálymentesen lehessen a domain-relációkat az angolból átvenni, felvettük az összes, PWN-ben szereplő category és region domain synsetet. A region domain fogalmak körét kiegészítettük speciálisan magyar régiók gyűjtésével is. A usage domain relációk használatát elvetettük, a PWN-ben tapasztalt inkonzisztenciák miatt: néhol egy synset-re vonatkozó usage minősítés nem minden szinonimára (literálra) érvényes, és ez a kódolás nem teszi egyértelművé, hogy melyek azok. (A PWN-ben ugyanakkor előfordul az is, hogy pont a usage minősítés alapján választanak szét két synsetet.) Az általunk bevezetett, literál-szintű nyelvhasználati kódolás, melynek segítségével a synset minden egyes elemére külön-külön megadhatunk minősítő kódokat, rugalmasabb megoldást biztosít.

3.5 Tulajdonnevek

A nemzeti wordnetek kisebb-nagyobb számban tartalmaznak named entity (NE, „névvel rendelkező entitás”, itt most tulajdonnév) jellegű synseteket is. Ezek között vannak „univerzálisak”, pl. a világ országai, fővárosai; világirodalmi jelentőségű alkotók, híres képzőművészek, tudósok, politikusok, és vannak az adott nemzethez, országhoz köthető, helyi fontosságú NE-k, pl. az adott ország megyéi, régiói, települései; a nemzeti irodalom, képzőművészet, tudomány, politika nagyságai. A Magyar WordNet ilyen irányú bővítéséhez tematikus NE-listákat gyűjtöttünk. Ezek az alábbi fő kategóriákba sorolhatók:

- földrajzi nevek (ország, megye, település, egyéb (hegy, víz stb.))
- intézmény jellegű nevek (cégek, nevezetességek, kórházak, színházak, mozik, légitársaságok stb.)
- személynevek (keresztnevek, családnevek, híres emberek nevei (művészek, történelmi alakok stb.))
- címek (újságok)

- márkanevek (termékek, árucikkek)

A listákat áttekintve az alábbi feladatokat határoztuk meg:

1. Egységesítés (formátum- és kódkonverziók)
2. Szelekció (mely kategóriákat, ill. az adott kategóriából mely NE-ket integráljuk)
3. Kollekció (a kiválasztott NE-knek milyen alakváltozatait, szinonimáit, körülírásait vegyük fel)

3.5.1 Synset-szintű beépítésre kijelölt NE-k

Bizonyos tematikus listák kiválasztott elemei közvetlenül bekerülnek a Magyar WordNetbe. A preferálandó kategóriák:

- földrajzi nevek:
  - országnevek
  - magyar megyék
  - magyar települések
  - világvárosok
- intézménynevek:
  - magyar nevezetességek
- személynevek:
  - magyar keresztnevek
  - híres emberek

Ha egy tulajdonnévnek van a magyarban meghonosodott írásmódja, literálváltozata, akkor alapelv, hogy minden esetben a magyar írásmódú alakokat tüntessük fel.

A synset-szintű beépítéssel kapcsolatos részfeladatok:

1. A szelektálandó állományok kézi leválogatása (a beépítendő NE-k megjelölése), a kiválasztott NE-k írásképének ellenőrzése, javítása.
2. Szelektálás közben az „ömlesztett” anyag finomítása, pl. a híres ember kategórián belül megadni az albesorolás(oka)t (festő, író, költő, hadvezér, politikus, fizikus stb.).
3. Annak ellenőrzése, hogy a kiválogatott literálok szerepelnek-e az angol nyelvű wordnetben (automatikus ellenőrzésnél probléma lehet, ha eltér a magyar és az angol írásmód, pl. Róma – Rome – Roma, Olaszország – Italy). Itt az is járható út, ha az angol wordnetes ellenőrzés előtt automatikusan generáljuk ezeket a synseteket, és a kézi ellenőrzéskor történik meg az angol wordnetbe való bekötés, ha ott már létezik ez a synset.
4. A bekötési synset (hipernima) meglétének ellenőrzése, szükség esetén felvétele és magyarítása.

3.5.2 Csak táblázatszintű, pointeres beépítésre javasolt NE-k

A szelektáláskor „kihulló" elemek is vannak annyira értékesek, hogy elérhetővé tegyük őket a wordnetből. Ezek nem önálló synsetként jelennek meg az ontológiában, hanem listaszerűen. Ilyen módon kerülnek be az alábbi témakörökből NE-k:

- intézménynevek:
  – az összes fennmaradó
- személynevek:
  – magyar családnevek
- címek
- márkanevek

Az ezzel kapcsolatos feladatok:

1. A formátum egységesítése (egységes ékezethasználat, kiegészítő információk törlése stb.)
2. A pointeres megoldás technikai részleteinek meghatározása (új reláció bevezetése: synsetekről tematikus listák felé)

4. A bővítési módszerek kiértékelése

4.1 Vizsgálati módszer

Egyrészt arra voltunk kíváncsiak, hogy a Magyar WordNet központi részét alkotó synsetek (BCSHu) mennyire relevánsak a magyar beszélők számára, azaz valóban a mag-ontológiában van-e a helyük. Másrészt az egyes bővítési módszerek hatékonyságát is számszerűsíteni szerettük volna. Ehhez az egyes bővítési eljárások során felvett synsetekből választott véletlen mintákat értékeltettünk két magyar beszélővel. A mérés a következőképpen folyt le:

a) A vizsgálandó synset-halmazokból 200-as véletlen mintákat vettünk.
b) A minták synsetjeit egyenként értékelte a két magyar anyanyelvű személy egymástól függetlenül. Minden synsetet 1-től 10-ig kellett pontozniuk. A nagyobb szám a fogalom nagyobb relevanciáját jelenti az értékelő személy számára. A két annotátor közötti egyetértés az összes értékelésre átlagolva 78,67%-os volt (egy adott synset értékelésében az egyetértést 100%-osnak vettük, ha mind a két annotátor ugyanazt a pontot adta, 0%-osnak, ha a különbség maximális (9 pontnyi) volt. Az értékeket átlagoltuk a synsetek és az összes minta felett.)
c) A két értékelő pontszámait synsetenként átlagoltuk, majd kiszámoltuk a 200 synset pontátlagainak átlagát és szórását.

4.2 Eredmények, értékelés

Az 1. és a 2. táblázat oszlopai jelentik azokat a főnévi wordnet-szegmenseket, melyek mindegyikéből 200 synsetes véletlen mintát generáltunk. Az egyes szegmensek:

NONBCS: az angol wordnet BCS-en kívüli synset-állománya
BCS1: a mag-ontológia 1-es szintje
BCS2: a mag-ontológia 2-es szintje
BCS3: a mag-ontológia 3-as szintje
CONC_1: az 1. koncentrikus bővítési körben felvett synsetek
TREE: a teljes részfák felvételekor bekerült synsetek
CONC_2_CAND: a 2. koncentrikus bővítési kör jelöltjei: a hipernima-reláció mentén elérhető synsetek
LIT_FREQ: korpuszokból készült szóalak-gyakorisági listák alapján felvett synsetek a 2. koncentrikus bővítési kör jelöltjei közül
ILI_OVL: a nyelvek közötti átfedések száma alapján felvett synsetek a 2. koncentrikus bővítési kör jelöltjei közül

	NONBCS	BCS1	BCS2	BCS3	CONC_1	TREE
átlag	4,51	6,56	6,21	5,03	5,71	4,21
szórás	2,48	2,78	2,20	2,45	1,71	2,61

1. táblázat

Az 1. táblázat NONBCS eredménye igazolja várakozásainkat. Megállapítható, hogy ha a bővítés során a bekerülő fogalmakat véletlenszerűen választottuk volna ki az angol wordnetből, akkor átlagosan kisebb relevanciájú fogalmak kerültek volna a Magyar WordNetbe. Ez tulajdonképpen a CONC_1 halmazának kontrollcsoportja.

A BCS1, BCS2 és BCS3 mérések kilógnak a sorból, abban az értelemben, hogy ezeket a synset-halmazokat nem mi válogattuk össze, hanem a EuroWordNet és a BalkaNet ontológiákból „örököltük”. Itt azt mérhettük meg, hogy az említett projektek által kulcsfontosságúnak tekintett fogalmak mennyire relevánsak a két magyar értékelő számára. A kapott eredmények átlagai és szórásai is elmaradtak várakozásainktól. Magasabb relevanciára és kisebb szórásértékekre számítottunk.

A CONC_1-ben, azaz az első koncentrikus bővítési körben olyan synseteket építettünk be a Magyar WordNetbe, melyek a mag-ontológiából egy lépésben elérhetők a hiponima-reláció mentén. A közepesnek mondható átlag a BCS viszonylag gyengébb eredményéből származtatható. Az alacsony szórásérték a synsetek értékességének homogenitását mutatja.

A TREE gyenge eredménye könnyen megmagyarázható. Egy teljes részfa beépítése azt jelenti, hogy felveszünk minden synsetet, amely egy bizonyos kijelölt csomópont alatt található. Pl. népek, nyelvek, pénznemek teljes részfái. Ezek között – a kiemelt jelentőségű fogalmak mellett – nagyon sok kevéssé ismert, illetve ritkán használt fogalom is szerepel. A TREE magas szórásértéke ezt a feltételezést alátámasztja.

	CONC_2_CAND	LIT_FREQ	ILI_OVL
átlag	4,25	5,26	8,32
szórás	2,27	1,74	1,25

2. táblázat

A CONC_2_CAND halmaz kb. 35 ezer főnévi synset-jelöltet tartalmazott a 2. koncentrikus bővítési körre. Az e halmazon elvégzett mérés azt mutatja, hogy ha véletlenszerűen választottuk volna ebből a 2. kör synsetjeit, akkor milyen értékes fogalmak kerültek volna be a Magyar WordNetbe. Ez tulajdonképpen a másik két oszlop halmazának kontrollcsoportja.

A LIT_FREQ eredményei a gyakorisági listák hasznosságát igazolják.

Az ILI_OVL-en mért átlag és szórás egyaránt kiemelkedő volt. Ilyen jó eredményeket nem vártunk ezen a szegmensen. A véletlen mintához viszonyított majdnem kétszeres átlag és alig több mint félszeres szórásérték e módszert mutatja messze a leghatékonyabbnak.

Összegzésként elmondható, hogy érdemes munkát fektetni a bővítés szisztematizálásába. A Magyar WordNet egyik értékét a synsetek magas relevanciája jelenti.

Bibliográfia

1. Alexin, Z., Csirik, J., Kocsor, A., Miháltz, M.: Construction of the Hungarian EuroWordNet Ontology and its Application to Information Extraction. In: Proceedings of the Third International WordNet Conference, Seogwipo, Jeju Island, Korea (2006) 291–292.
2. Christodoulakis, D. N. (ed.) (2004): Design and Development of a Multilingual Balkan Wordnet. BalkaNet Final Report. http://www.ceid.upatras.gr/Balkanet/deliverables/finalreport_sub.pdf
3. Fellbaum, C. (ed.): WordNet: An Electronic Lexical Database. Cambridge, MA: MIT Press (1998)
4. Miháltz, M.: Magyar EuroWordNet projekt: bemutatás és helyzetjelentés. III. Magyar Számítógépes Nyelvészeti Konferencia, Szeged (2005) 68–78
5. Tufiş, D., Cristea, D., Stamou, S.: BalkaNet: Aims, Methods, Results and Perspectives. A General Overview. In Romanian Journal of Information Science and Technology Special Issue, vol. 7, no. 1-2 (2004)
6. Vossen, P. (ed.): EuroWordNet General Document, Version 3. University of Amsterdam (1999)