Szemantikai gráf alapú mondatelemző modul kidolgozása IS-NLI értelmezőhöz

Kovács László

[1] Miskolci Egyetem, Általános Informatikai Tanszék,
3515 Miskolc-Egyetemváros
kovacs@iit.uni-miskolc.hu

Kivonat: Az NLI lekérdező modulok egyik alapfeladata a természetes nyelven beérkező parancsok átkonvertálása a feldolgozó modul saját parancsnyelvére. Napjainkban az NLP transzformációs feladat megoldási módszerek között dominálnak a generatív vagy a statisztikai algoritmuson alapuló eljárások. A cikk egy fogalmi hálót mint közvetítő elemet tartalmazó NLP modul modelljét ismerteti. A kidolgozott rendszer a Dependency Grammar modellen alapul.

1 Bevezetés

Az információs rendszereket tekintve a természetes nyelvi feldolgozás (NLP, natural language processing) egyik legfontosabb alkalmazási területét az emberközeli lekérdező felületek jelentik. A lekérdező modulok egyik alapfeladata a természetes nyelven beérkező parancsok átkonvertálása a feldolgozó modul saját parancsnyelvére. A természetes nyelvi interfésszel rendelkező információs rendszerek gyökerei az 1970-re nyúlnak vissza. Az úttörő LUNAR projekt a holdkőzetek adatbázisában való lekérdezésekhez dolgozott ki NLI (természetes nyelvű interfész) felületet. A RENDEZVOUS (Codd, 1977) rendszer volt az első általános célú adatbázis NLI modul. Az NLP transzformációs feladat megoldási módszerek között dominálnak a generatív vagy a statisztikai algoritmuson alapuló eljárások. A generatív esetben, mely alatt most azt értjük, hogy a kódba beépítjük a két nyelv általunk fontosnak tartott szabályrendszerét és a meghatározzuk a két szabályrendszer közötti közvetlen konverziót. E módszer előnye az egzakt működés, a feltárt szabályrendszernek való pontos megfeleltetés. Hátránya viszont, hogy ismertnek és algoritmizálhatónak kell lennie az alkalmazott nyelvek nyelvtanának. Ekkor a transzformáció jósága a nyelvtan leképezés jóságától függ. A statisztikai módszereknél, melyek egyik leggyakrabban használt formája a Markov-modelleken alapuló módszerek, a tanítómintából kinyert valószínűségi szabályok alkotják a konverzió magját A legtöbb statisztika alapú módszernél vagy a szabad szöveges forrásokra építenek, vagy nyelvi annotációt alkalmaznak a tanulás hatékonyság javítására. A tapasztalatok azt mutatják, hogy az alapszöveg, a nyers szintaktika önmagában nem elegendő a nyelvtan hatékony feltárására. A nyelvtani annotáció más részről jelentős többletterhet jelent, és igen nagy tanító mintahalmazt igényel. A cikkben bemutatott módszer alapvonása, hogy a nyelvi konverziót egy köztes szemantikát leíró formalizmuson keresztül hajtjuk végre. A

szemantikai hálóval történő köztes tartalom annotáció egyik fő előnye a nagyobb rugalmasság a nyelvfüggetlenségben és így a különböző nyelvek közötti konverzióban. Az irodalomban viszonylag szerény az ilyen szemantikai mediátoron keresztüli parancsértelmezőre vonatkozó vizsgálatok száma, mivel ezen megoldás csak nemrég került a vizsgálatok központjába [10]. A dolgozat bemutatja a javasolt transzformációs modult, a nyelvi és a szemantikai reprezentációs alak közötti konverzió lépéseit. A modell egyik fontos eleme a megfelelő szemantika ábrázolásai mechanizmus kiválasztása. Elemzésünk azt mutatta, hogy a tradicionális szemantikai reprezentációk nem adnak kellő hatékonyságot a transzformációban, ezért egy HECG-nek elnevezett szemantikai háló került kidolgozásra a modulhoz. A HECG ontológia leírás és az interfész nyelvek közötti transzformáció több lépcsőben megy végbe. A konverzióhoz szükséges nyelvtani elemek deklaratívan adhatók meg mint működési paraméterek. A konverzió első fázisában a szemantikai gráfból egy szógráf képezhető. A szógráf jellegében igen sok közös vonással rendelkezik a word dependency graph nyelvi reprezentációval, mely mintegy átmenetet képez a szintaxis és a szemantika között. A leképzés második fázisában a szógráfból szavak szekvenciája generálódik. A kidolgozott séma mintarendszer keretekben működik és a hatékony implementáció kidolgozása esetén fontos alkalmazási területeket kiszolgálására válhat alkalmassá.

2 A szófüggőség alapú nyelvtan modellek

A nyelvtanok egyik elterjedt osztályozása szempontja, hogy mit tekintünk mondat egységnek: a szavakat, szóláncokat vagy a szavak közötti függőségi rendszert. A függőség alapú rendszerek fő jellemzője a fejfüggő asszimmetria és az a törekvés, hogy a fej és tagelemek közötti kapcsolatot szemantikai alapokon nyugvó függőségi relációkkal írjuk le. A függőségi nyelvtan (Dependency Grammar) modelljét a francia Lucien Tesnière [1] dolgozta ki. A modell alapegysége a stemma, amely a szavak között fennálló szintaktikai függőségi viszony grafikus reprezentációjának tekinthető. A modell értelmezésében az ige tekinthető a legmagasabb helyen álló szónak, amely felügyeli, vezérli az alatta elhelyezkedő kiegészítőket, csatolmányokat. A csatolmányok maguk is lehetnek összetettek, rendelkezhetnek saját csatolmányokkal.

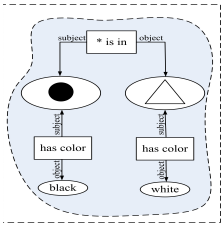
Tesnière elmélete jelentős hatással volt a nyelvészek széles táborára, azokra, akik a szemantika fontosságát a szintaktika elé helyezték. A [2] mű részletes áttekintést ad erről a területről.

Klein és Simmons [3] ezen függőségi nyelvtant alkalmazta gépi fordítást végző rendszerükre. A Valency elmélet [4] és Meaning-Text elmélet [5] néhány példái a mai is folyó függőségi nyelvtant támogató kutatásoknak. Schank is ezen irányból kiindulva alkotta meg a Conceptual Dependency Graph [7] modelljét, melynek sajátossága, hogy a háló elemei a szavak helyett a fogalmakat reprezentálják. A függőségi nyelvtan szerepét széles körben elemző mű a [8]. A függőségi nyelvtanok egy további lényeges bázisa a Case Grammar modell [9] is, melyben a függőségi élek címkézettek. Az Extensible Dependency Grammar [10] és a Word Grammar [6] olyan újszerű nyelvi modelleket képviselnek, melynek célja egy egységes modellbe egyesíteni a szemantikai és szintaktikai elemeket.

A függőségi nyelvtanok egyik előnyös tulajdonsága a magyar nyelv vonatkozásában, hogy a rendszer tudja kezelni a szabadsorrendű szerkezeteket is. Emellett lehetőség van nem folytonos szóláncot alkotó struktúrák kezelésére is. A DG alapú reprezentációban az élek így nemcsak szintaktikai szereppel bírnak, alkalmasak a szemantikai szerep jelölésére is.

3 A HECG szemantikai háló

A szemantikát, a jelentést megadó hatékony leírások közé tartozik a szemantikai háló, amely egy ontológiai modellt valósít meg. A szemantikai háló (Sloman 2003) egy olyan irányított gráf, melynek csomópontjai a fogalmakat reprezentálják és a köztük lévő élek a különböző relációkat jelölik. Az ontológia területe a problémák, a vizsgált világ fogalmi szinten történő leírásával foglalkozik. Az ontológiai rendszerek egyik lényeges vonása, amely megkülönbözteti őket a hagyományos szemantikai modellektől, hogy szabályalapú logikai kezelő nyelvvel is rendelkeznek. A mögötte álló következtető motor segítségével ellenőrizni lehet a modell konzisztenciáját, illetve új tények levezetését is biztosítani tudja a rendszer. Az ontológialeíró nyelvek között a két leginkább elterjedt nyelv az RDF és az OWL. Az RDF nyelvben az ábrázolás alapelemei körébe az erőforrások, a literálok és az állítások tartoznak. Az erőforrásoknak két fő típusa van: egyed és tulajdonság. Az állítás egy (p,s,o) hármassal adható meg, ahol a p egy tulajdonság, s egy erőforrás és o egy erőforrás vagy literál. Jelentését tekintve a p egy predikátumot, egy állítmányt takar. Az s szimbólum a szubjektum, az alany, míg az o az objektum, az érték. A RDF modellben az állítások vonatkozhatnak nemcsak elemi egyedekre, hanem más állításokra is. Az OWL nyelv az RDF nyelv kiterjesztésének tekinthető. A hozzátett új funkcióelemek köre magába foglalja az adattípus kezelést, a tulajdonság minősítését, a számosság ellenőrzést és egyéb új megszorítási elemeket.



1. ábra. ECG modell minta.

A kidolgozott HECG modell egy olyan fogalmi modellt jelöl, melyben a szerkezet építő elemei az egymásba foglalható elemi állításatomok. Egy elemi állítás magja az ige vagy predikátum. A predikátumhoz csatolható elemeket argumentumoknak nevezzük. Mind az élek, mind az elemek címkézettek, ahol a címke több elemi információt hordoz, mint a kapcsolat szemantikai tartalma, a kapcsolat megvalósulási megszorításai. A kapott fogalmi hálóból egy fókusz-állítás megadásával egy kapcsolati fa feszíthető ki, amely az elemek függőségi rendszerét is kifejezi. Az 1. ábra egy mintahálót mutat be.

4 A mintarendszer architektúrája, működése

A kidolgozott rendszer kétirányú konverziót valósít meg a HECG modell és egy szimbolikus nyelv között. A konverzió menete az alábbi alaplépésekre bontható fel:

- a hálóhoz a kijelölt predikátum alapján egy kifeszítő, függőségi fa generálása
- a fához egy szó-fa generálása, ahol a fogalmak mögé a hozzá csatolható szavak kerülnek be egy megadott tezauruszból
- a szavak módosítása az élekhez rendelt nyelvtani ragok alapján
- a szavakból a mondat generálása a sorrendiségi megszorításokat figyelembe véve.

A fordított irányú konverziónál elsőként a mondat elemeit határozzuk meg szavakra és morfémákra bontással. Az elemzés főbb lépései:

- Morfémaelemző segítségével a szavak szerkezetének feltárása
- A szavak morfémaelemzésével a ragok meghatározása
- A szótövek alapján a szó fogalmi kategóriáinak kijelölése
- A ragok alapján a kapcsolható argumentum élek kijelölése
- A szógráfhoz rendelt sorrendiség előírás összevetése a beérkező mondat sorrendiségével
- A vizsgált fogalomháló és a mintamondat távolság mértékének meghatározása
- A legközelebbi háló kiválasztása, mint a mondat jelentését reprezentáló háló.

A megadott algoritmus segítségével a mintarendszerben a magyar nyelv adott témakörhöz tartozó mondatait egy predikátum kalkulusbeli formára alakította, mely a későbbi lépésekben SQL vagy más nyelvre konvertálható tovább.

Hivatkozások

1. Tesnière, L.: Éléments de syntaxe structurale. Paris: Klincksieck (1959)
2. Sowa, J. F.: Semantic networks. In: Shapiro, S. C. (ed.): Encyclopedia of Artificial Intelligence. 2nd ed., Wiley. (1992)
3. Klein, S., Simmons, R. F. Syntactic dependence and the computer generation of coherent discourse. Mechanical Translation 7 (1963)
4. Hudson, D R.: Language Networks: The new Word Grammar. Oxford University Press (2007)
5. Mel'cuk, I. A.: Towards a linguistic ''Meaning ⇔ Text' model. In: Kiefer, F. (ed.): Trends in Soviet Theoretical Linguistics. Dordrecht, Reidel (1973) 35–57
6. Steele, J. (ed.): Meaning-Text Theory. Ottawa, University of Ottawa Press (1990)
7. McEnery, A., Xiao, R., Tono, Y.: Corpus-Based Language Studies: An Advanced Resource Book. In: Ser. Routledge Applied Linguistics. Routledge (2005)
8. Hudson, R.: Recent developments in dependency theory. In Jacobs, J., v. Stechow, A., Sternefeld, W., Vennemann, T. (eds.): Syntax. Ein internationales Handbuch zeitgenöössischer Forschung. Berlin, Walter de Gruyter (1993) 329–338
9. Fillmore, C. J.: The case for case. In: Bach, E., Harms, R. T. (eds.): Universals in Linguistic Theory. New York, Holt, Rinehart and Winston (1968) 1–88
10 Debusmann, R.: Extensible Dependency Grammar: A modular grammar formalism based on multigraph description. PhD thesis (2006)