

A szemantikai reprezentáció generálása

vázlat, v0.1

Simonyi András, ALL

2011. március

1. Előfeltevések

1.1. A dependenciaelemzés menete

Feltételezzük, hogy a szintaktikus dependenciaelemzés az algoritmus részleteitől függetlenül egy olyan ciklusnak tekinthető, amely végighalad az elemzendő mondat szavain,¹ és ezeket egyenként megpróbálja hozzákötni a szöveg más szavaihoz.

1.2. Szemantikai reprezentációk

Az általánosság kedvéért feltesszük, hogy a szövegekhez rendelendő szemantikai reprezentációk olyan irányított gráfok, amelyeknek mind a csúcsai, mind az élei címkékkel lehetnek ellátva.

1.3. A reprezentációk összekapcsolása a szintaktikai elemzés dependenciagráfjával

A szemantikus elemzés célja nem pusztán a szöveg tartalmának megfelelő gráf létrehozása, hanem a szöveg dependenciagráfjának a reprezentáció megfelelő részeivel való összekötése is: a szemantikai elemzésnek generálnia kell egy *Val* relációt, amely a dependenciagráf elemeit, tehát a szavakat és a közöttük lévő szintaktikai kapcsolatokat a nekik megfelelő reprezentációelemekhez rendeli.

¹Pontosabban atomi kifejezésein. A továbbiakban szó alatt az elemzés szempontjából atominak tekinthető kifejezést fogunk érteni.

1.4. A MagNP-k szemantikai reprezentációi

Mivel a MagNP-k elemzése a dependenciaelemzéstől függetlenül történik, ezért ezeket olyan kifejezéseknek tekintjük, amelyek a dependenciaelemzés szempontjából atomiak, és amelyeknek a reprezentációja a *Val* reláció megfelelő részével együtt már rendelkezésre áll.

Azt is feltesszük, hogy az előfeldolgozás minden egyes MagNP esetében kijelöli a hozzátartozó szemantikai reprezentáció azon csúcsát, amely defaultként szolgál más szemantikai reprezentációkkal való összekötéskor. Ezek tipikusan a MagNP-k fejének megfelelő csúcsok, amelyeket a *Con* függvény rendel a MagNP-khez.

2. A szemantikai reprezentáció generálása a dependenciaelemzéssel párhuzamosan

A szemantikai reprezentáció létrehozásának kulcslépéseit a szintaktikai elemzés menetéről tett feltevésünknek megfelelően az elemzés következő két lépéstípusához kötjük:

1. Az elemző ciklusa egy szóhoz ér, és megkezdí annak feldolgozását
2. Az elemző összeköt két szót egy szintaktikai relációval

A szintaktikai elemzés 1. típusú lépéseinél az esetek jelentős részében már generálható az éppen vizsgált szóhoz tartozó szemantikai reprezentáció, illetve a *Val* reláció megfelelő értéke(i). Ebből adódóan az elemzésnek a következő feladatot kell ekkor végrehajtania:

- **Ha** az aktuális szónak még nincs reprezentációja, és a szó reprezentációja már generálható, **akkor** le kell generálni a reprezentációt.
- **Ha** az aktuális szónak van reprezentációja **és** még nincs összekötve a szóval, **akkor** a megfelelő rendezett párral/párokkal ki kell bővíteni a *Val* relációt.

Az elemzés 2. típusú lépései *után* azokat a reprezentációelemeket kell létrehozni, amelyek az összekötött két szó szintaktikai kapcsolatának alapján generálhatóak, tehát meg kell tenni a következő lépéseket:

- **Ha** az összekötött szavak között megállapított szintaktikai kapcsolat alapján létrehozható új reprezentáció, **akkor** az le kell generálni.

- **Ha** a szintaktikai kapcsolatnak van reprezentációja, akkor a megfelelő rendezett párral/párokkal ki kell bővíteni a *Val* relációt.

A következő szakaszokban részletezzük az itt általánosan leírt lépéseket.

3. Szó eléréskor generálható szemantikai reprezentációk

3.1. A koreferencia kezelése

Ha egy elért k kifejezésről az előfeldolgozás megállapította, hogy koreferens egy olyan l kifejezéssel, amelynek már van szemantikai reprezentációja (a *Val* reláció már rendel hozzá értéket), akkor k reprezentációi meg kell hogy egyezzenek a koreferens kifejezés reprezentációival, tehát tetszőleges m gráf-elemre, ha *Val* tartalmazza $\langle l, m \rangle$ -et, akkor *Val*-t ki kell bővíteni a $\langle k, m \rangle$ rendezett párral.

3.2. Perdurant jelentésű (predikatív) szavak

A perdurant jelentésű szavak tipikusan igék vagy igenevek, de vannak közöttük nem igéből képzett főnevek is (pl. *method*). Ha egy szó perdurant jelentésű, akkor szemantikus reprezentációja generálható szintaktikus kapcsolatainak ismerete nélkül:

- A szemantikai gráfba be kell illeszteni egy új csúcspontot.
- A szemantikus lexikon által a szóhoz rendelt synset-azonosítóval meg kell címkézni a létrehozott csúcspontot.²

A szó összekötése a reprezentációval triviális: a *Val* reláció egyszerűen az \langle aktuális szó, újonnan létrehozott csúcs \rangle rendezett párral bővül.

3.2.1. A reprezentáció további részeinek generálása

Bizonyos esetekben ez az alapreprezentáció kiegészítésre szorulhat. Pl. ha a kérdéses perdurant jelentésű szó az *előállítás kontextusában* fordul elő, és a szemantikus lexikon szerint olyan folyamatot jelöl, amely valamilyen eredménnyel jár, akkor a folyamat eredménye már az elemzésnek ezen a pontján

²Az egyszerűség kedvéért feltesszük, hogy a szemantikus lexikon minden szóra visszaad valamilyen értéket, mivel a lexikonba explicite nem szereplő szavakra a lexikon default synset-azonosítókat adhat vissza.

felvehető a reprezentációba, függetlenül attól, hogy van-e rá explicit hivatkozás a szövegben.

3.3. Határozószavak

A határozószavak elérésekor a perdurant jelentésű szavakhoz hasonlóan kell eljárni:

- A szemantikai gráfba be kell illeszteni egy új csúcspontot, és
- a szemantikus lexikon által a szóhoz rendelt synset-azonosítóval meg kell címkézni a létrehozott csúcspontot.

3.4. Relációt jelölő szavak

A relációt jelölő szavak (*contain*, *comprise* stb.) többségénél a dependenciaviszonyok ismerete nélkül nem generálható a szó teljes jelentése, csak egy él, amely azt reprezentálja, hogy a kifejezés relációt jelöl. Ennek megfelelően ezeknek a szavaknak az elérésekor tipikusan csak a következő lépések hajthatók végre:

- A reprezentációba be kell illeszteni egy új élt.
- Az élt a szemantikus lexikon lekérdezése alapján meg kell címkézni a szó által jelölhető összes relációt tartalmazó legszűkebb relációval (ez lehet akár a legtágabb, bármely két individuum között fennálló reláció is).

A szó összekötése a reprezentációval itt is egyszerűen a *Val* reláció <aktuális szó, újonnan létrehozott él> rendezett párral való bővítésével történik.

3.5. Boole-konnektívumok (negáció, konjunkció, diszjunkció)

TODO – a szintaktikai elemző egyelőre nem kezeli őket.

4. Szintaktikai kapcsolat bejelölésekor generálható szemantikai reprezentációk

4.1. Perdurant jelentésű szó összekötése egy, a vonzatkere-téhez tartozó argumentummal

Ebben az esetben feltételezhetjük, hogy a perduranthoz és az argumentumhoz tartozó szemantikai reprezentációk már adottak. Ekkor a szintaktikai kapcsolat szemantikai reprezentációjának előállításához

- meg kell határozni az argumentum reprezentációjának azt a csúcspontját, amelyhez majd a perdurantot reprezentáló csúcsot kötjük (MagNP-*knél* ezt alapesetben a *Con* függvény adja meg),
- össze kell kötni a két csúcsot egy, a perdurantból az argumentumba mutató éllel,
- és a létrehozott élt meg kell címkézni annak a relációnak (tipikusan tematikusszerep-relációnak) az azonosítójával,³ amelyet a perdurant jelentésű szó adott vonzatához rendel a szemantikus lexikon.

Val a <behúzott szintaktikus él, behúzott szemantikus él> párral bővül.

4.1.1. A reprezentáció további részeinek generálása

Bizonyos esetekben a fenti lépések által előállított reprezentáció még kiegészítésre szorulhat, pl. az ECM/Raising/Control igék esetében az argumentumként szereplő perdurant jelentésű szó reprezentációját nem csak a fej reprezentációjával, hanem más argumentumokkal (pl. az alannyal) is össze kell kötni. Az ilyen típusú igékre vonatkozó szabályokat a szemantikus lexikon fogja leírni.

4.2. Perdurant jelentésű szó összekötése szabad határozóval

A szabad határozó funkciójú, az elemzés szempontjából atominak tekintett kifejezések között megkülönböztethetünk határozószókat és prepozíciós frázisokat – ezt a két kategóriát a szemantikai reprezentáció előállításakor különböző módon kell kezelni.

³A következőkben a tömörség kedvéért egy ontológiai reláció vagy osztály azonosítójával való megcímkézés helyett gyakran a relációval vagy osztállyal való megcímkézésről fogunk beszélni. Ha nem jelezzük külön ennek az ellenkezőjét, akkor ezekben az esetekben mindig a megfelelő azonosítóval való címkézésről van szó.

4.2.1. Határozószók

Határozószó és egy predikatív szó összekötésekor a szemantikai reprezentáció létrehozásához

- fel kell venni egy, a *Val* által a predikatív szóhoz rendelt csúcsból a határozószóhoz rendelt csúcsba mutató élt,
- a létrehozott élt meg kell címkézni a HASQUALITY relációval,
- és *Val*-t ki kell bővíteni a <szintaktikai él, létrehozott szemantikai él> rendezett párral.

4.2.2. Prepozíciós frázisok

A prepozíciós frázisokat a szintaktikai elemző „<prepozíció>-complement” címkéjű élekkel köti a hozzájuk tartozó fejhez. A szemantikus reprezentáció létrehozásához

- a szemantikus lexikon segítségével meg kell állapítani a kérdéses prepozíciónak megfelelő ontológiai relációt, és azt, hogy ennek az ontológiai relációnak a prepozíció után álló NP reprezentációjának melyik csúcsához kell kötni.
- a szemantikus reprezentációban létre kell hozni egy új, a perdurant jelentésű szót reprezentáló csúcsból az NP reprezentációjának megfelelő csúcsába mutató, a szemantikus lexikon által megadott relációval címkézett élt, és
- *Val*-t ki kell bővíteni a <prepozíció,létrehozott él> és a <szintaktikai él,létrehozott él> rendezett párokkal.

4.3. Relációt jelölő szó összekötése argumentumaival

Mint ezt már említettük, a relációt jelentő szavak többségének pontos reprezentációja csak bizonyos szintaktikai viszonyok ismeretében hozható létre, mivel az ilyen kifejezéseknek megfelelő ontológiai reláció függ az argumentumok szemantikai típusától.

Az egyes relációt jelentő szavakra vonatkozó pontos szabályokat a szemantikus lexikon fogja tartalmazni. Példaként itt bemutatjuk az egyik leggyakoribb ilyen kifejezés, a *contain* kezelését.

4.3.1. Példa: *contain*

A *contain* előfordulásainak teljes reprezentációja csak akkor hozható létre, ha mind az alany, mind a tárgy ismert, és szemantikai reprezentációjuk is rendelkezésre áll. Ennek megfelelően *mindkét* argumentum szintaktikai összekötése után a következő lépések hajtandók végre:

- ha a *Val* által a *contain* előforduláshoz rendelt élnek van címkéje, akkor azt el kell távolítani.
- **ha** az alany szemantikai reprezentációja az OBJECT típusba tartozik (a *Con* függvény a kifejezéshez egy OBJECT címkéjű csúcsot rendel), **akkor**
 - **ha** a tárgy is OBJECT típusú, **akkor** a *Val* által a *contain*-előforduláshoz rendelt élt meg kell címkézni a PARTOF relációval, és az él kiinduló csúcsának a szintaktikai alany reprezentációjának OBJECT címkéjű csúcsát, míg beérkező csúcsának a szintaktikai tárgy reprezentációjának OBJECT címkéjű csúcsát kell megtenni.
 - **ha** a tárgy AMOUNTOFMATTER típusú, **akkor** a *Val* által a *contain*-előforduláshoz rendelt élt meg kell címkézni a CONSTITUTEDBY relációval, és az él kiinduló csúcsának a szintaktikai alany reprezentációjának OBJECT címkéjű csúcsát, míg beérkező csúcsának a szintaktikai tárgy reprezentációjának AMOUNTOFMATTER címkéjű csúcsát kell megtenni.
 - **ha** a tárgy SUBSTANCE típusú, **akkor** a *Val* által a *contain*-előforduláshoz rendelt élt meg kell címkézni a CONSTITUTEDBY relációval, és fel kell venni egy újabb, az említett élhez egy újonnan létrehozott AMOUNTOFMATTER címkéjű csúcson keresztül csatlakozó élt a HASSUBSTANCE címkével. Az előbb él kiinduló csúcsának a szintaktikai alany reprezentációjának OBJECT címkéjű csúcsát, míg az utóbbi beérkező csúcsának a szintaktikai tárgy reprezentációjának SUBSTANCE címkéjű csúcsát kell megtenni.
- **ha** az alany szemantikai reprezentációja az AMOUNTOFMATTER típusba tartozik, **akkor**
 - **ha** a tárgy is AMOUNTOFMATTER típusú, **akkor** a *Val* által a *contain*-előforduláshoz rendelt élt meg kell címkézni a PARTOF relációval, és az él kiinduló csúcsának a szintaktikai alany reprezentációjának AMOUNTOFMATTER címkéjű csúcsát, míg beérkező

csúcsának a szintaktikai tárgy reprezentációjának AMOUNTOFMATTER címkéjű csúcsát kell megtenni.

- **ha** a tárgy SUBSTANCE típusú, **akkor** a *Val* által a *contain*-előforduláshoz rendelt élt meg kell címkézni a PARTOF relációval, és fel kell venni egy újabb, az említett élhez egy újonnan létrehozott AMOUNTOFMATTER címkéjű csúcson keresztül csatlakozó élt a HASSUBSTANCE címkével. Az előbbi él kiinduló csúcsának a szintaktikai alany reprezentációjának AMOUNTOFMATTER címkéjű csúcsát, míg az utóbbi beérkező csúcsának a szintaktikai tárgy reprezentációjának SUBSTANCE címkéjű csúcsát kell megtenni.

4.4. Boole-konnektívumok

TODO – a szintaktikai elemző egyelőre nem kezeli őket.

5. Visszacsatolás a szemantikai és a szintaktikai elemzés között

A talán legegyszerűbb visszacsatolási mechanizmus a következő módon működik: Az elemző egy-egy szintaktikus él behúzása után nem hoz létre automatikusan egy megfelelő szemantikai reprezentációt, hanem jelezheti, hogy a kérdéses szintaktikai élhez tartozó szemantikus reprezentáció létrehozása megsértene bizonyos szemantikai megkötéseket. Ekkor a szintaktikus elemzés visszalép, és megpróbálja a problémás él által összekötött szavakat máshová kötni. A visszacsatoláskor többek között a következő kikötések vehetők figyelembe:

- **Argumentumrestrikciók:** a szemantikus lexikon tartalmaz olyan megszorításokat, amelyek kizárják bizonyos típusú csomópontok bizonyos típusú éllel való összekötését, pl. sok igei vonzatkeret esetében megadja, hogy az argumentumhelyeket milyen típusú entitások tölthetik ki.
- **Inkompatibilis/inkonzisztens jellemzők:** olyan információ is szerepelhet a szemantikus lexikonban, hogy két synset szigorúan inkompatibilis, tehát egy reprezentációelem nem címkézhető meg egyidejűleg mindkét, a synsetekhez tartozó azonosítóval.