HunNer korpusz – projektjavaslat

2006. augusztus

A projekt célja létrehozni egy félmillió szövegszavas manuálisan NE annotált korpuszt, ami

- jól használható gépi tanuláson alapuló NE címkézők tanítására,
- szabványos kiértékelő korpuszként sokan tudják használni
- miközben elő- és utófeldolgozó eszközöktől független.

A projektet a MOKK, Nytud és az SZTE (ábécé sorrend) közösen indítja, a költségeket egyenlően felosztva.

1. Költségbecslés (a Szeged Korpusz taggelése alapján

A Szeged korpusz tapasztalatai alapján, ahol mintegy 200.000 szó került taggelésre (a korpusz felcimkézése mintegy 200 munkaórát igényelt a későbbi javítási fázisokkal együtt) a következő becslést tehetjük a várható költségekre: Ha két független annotátort alkalmazunk, akkor a betanítással, és a taggelési különbségek kiértékelésével, javításával 100.000 szövegszó taggelése mintegy 300.000 forint költséggel az esetleges bonyolultabb cimkézési előírásokat is beszámítva, biztosan megoldható (1.000 Ft/munkaóra költséget számolva).

A korpusz jellegét úgy lenne célszerű meghatározni, hogy a korpuszt alkotó szövegek heterogénak legyenek (témájukat tekintve). A korpusz méretének lehetővé kellene tenni, hogy az azon tanított modellek

- általános szövegen megállják a helyüket
- specifikus szövegen nagyon jól tudjanak működni, azaz a korpusz egyes részei (különböző domain-ek) kellően nagyok legyenek önálló train adatbazisként (erre egy minimalis mérethatár 150-250 ezer token)

Ez alapján minimum 500K - 1 millió szövegszó cimkézése tehetné lehetővé, hogy a korpuszban sporttal, turizmussal, gazdasággal, politikával, stb. foglalkozó, megfelelő méretű (4-5 különböző domain) részkorpuszok legyenek. Ez a méret lehetővé tenné, hogy a korpuszt sokféle célra használni lehessen, azaz széles körben használt referencia-korpusz legyen. A mostani projekt célja lehet ennek egy kisebb részének elkészítése is, későbbi esetleges pályázati/szponzori források reményében.

Eredeti terveink szerint a MOKK, Nytud, SZTE hármas 300-300e forinttal száll be a projekttel, ami 300e szövegszó címkézését teszi lehetővé. Kérdés, hogy már most, első körben is célunk-e egy nagyobb, drágább korpusz létrehozása.

2. Annotációs séma

Első körben a szokásos `PERS`, `LOC`, `ORG` típusokat különböztetjük meg, de a projekt egyik feladata kialakítani egy *egységes annotációs útmutatót is.* A MOKK által használt LDC útmutató[1] több ponton eltér a Szeged Korpusznál használt szabályoktól (amelyek nagyrészt a CoNLL konferenciák annotálási szabályain alapulnak /de csak tulajdonnévi szófajú kifejezések lettek jelölve/). Csak néhányat említve:

- Az LDC útmutató szerint a neveket megelőző névelőket a név részeként kell jelölni, mivel együtt alkotnak egy frázist; míg a szegedinél ez nem tartozik bele.
- Az LDC alapján a szövegben NE-ként kell jelölni az *-i* képzős, helyre referáló, jelzői pozícióban álló mellékneveket, illetve a nemzetiségneveket is, míg a szegediben ezek egyáltalán nincsenek jelölve. Az LDC szerint a szervezetre referáló *-i* képzős alakokat is jelölni kell. Tehát taggelendő az *Az amerikai elnök*, *az amerikai étterem*, stb. is.
- A Szeged Korpusz annotálásakor egy további, úgynevezett `MISC` tag-et is használtunk, ami a különféle termékeket, márkákat, művek címeit, stb. fedte le. Azaz ebbe a kategóriába került minden tulajdonnév (ami éppúgy entitás, mint a korábban említett 3 osztály) ami nem tartozott az ORG, LOC, PERSON kategóriák egyikébe sem.
- Szegeden nem használatos a TITLE/ROLE tag, ami a személyneveket megelőző címekre, beosztásokra utal.

[1] `http://projects.ldc.upenn.edu/LCTL/Specifications/SimpleNamedEntityGuidelinesV6.5.pdf`

További példák az annotáció során felmerült problémás esetekből:

- A jelölendő nemzetiségnevek meglehetősen problémásak: sok esetben nem könnyű eldönteni, hogy hely- vagy szervezetnévként szerepelnek. Például *Az amerikai elnök támadást indított Afganisztán ellen* mondatban az LDC útmutató szerint az *amerikai* szervezetet, míg az *Afganisztán* helyet jelöl.
- A hely- és szervezetnevek elkülönítése más esetben is nehéz, pl. a következő mondatban: *A Nemzeti Múzeumban történt balesetben ketten megsérültek* a *Nemzeti Múzeum* helyet jelöl, és nem szervezetet.

SZTE álláspontja a taggelésről

- A lokatív értelemben használt szervezetneveknél maradhatnánk a *Tag for Meaning* elvnél, azaz az ORG/LOC cimkéknél a mondatbeli szerep szerint taggelnénk őket.
- A neveket megelőző névelők jelölése teljesen feleslegesnek tűnik.
- A MISC osztály használatát célszerűnek tartanánk (így teljesülhetne az, hogy minden tulajdonnév cimkézésre kerülne), míg a TITLE/ROLE jelölését - az LDC útmutatóban is elég speciálisan kezelik, csak személynevek előtt jelölik be - elhagyhatónak érezzük.
- A melléknévi pozícióba került entity-k jelölését nem érezzük fontosnak, tekintve, hogy sokszor problémás eldönteni a megfelelő cimkét (lásd fentebb), valamint az ilyen bejelölések sokszor a szöveg szempontjából nem jelentős információt hordoznának (pl. *a magyar étel* esetén a *magyar*mint hely. Mindenképp megfontolandó, hogy ezek jelölése helyett inkább a többi kategóriára építsünk nagyobb (vagy pontosabb) korpuszt.
- Esetleg meg lehet fontolni az LDC és a Szeged-féle annotáció közti konvertálhatóság biztosítását. Ehhez azonban a két taggelési stílus metszetén kívül eső (fentebb részletezett) entitásoknak új osztálycimke kellene, ami növelné a költségeket.

3. Az annotáció menete

Páros számú annotátorokkal dolgozunk, akik minden szöveget egymástól függetlenül, külön helyen annotálnak. Az annotátorok közötti egyetértést mérjük:

The numerical comparision between two alignments was calculated by the following standard method: Two named entities are considered identically tagged, if the start positions, the end positions and the tags are identical. The similarity score for two annotations is:

$$\frac{2 * |identicallytaggedentities|}{|entitiestaggedbyannotatorA| + |entitiestaggedbyannotatorB|}$$

Az annotátorok mindig csak a rögzített útmutató alapján dolgozhatnak, amit a menet közben felmerülő problémás kérdések megvitatásával folyamatosan fejlesztünk. Első lépésként ki kell dolgozni a kezdeti útmutatót. Ez alapján betanítjuk az annotátorokat, akik 10k szöveget önállóan feldolgoznak. A kezdeti egyetértés remélhetőleg eléri a 80%-ot. Az első körben megvizsgáljuk az eltéréseket, ha kell, módosítjuk az útmutatót. Ezen a szövegen addig dolgoznak újra és újra az annotátorok, amíg el nem érik a 95%-ot. Ezután indulhat csak a fennmaradó szöveg annotálása.

Az annotálás minden fázisának eredményét meg kell tartanunk.

4. Eszköz

A projekt előkészítő fázisában annotációs eszközt kell választani. Eddig felmerült ötletek:

- http://callisto.mitre.org/
- LDC által használt SimpleNet[2]
- CLaRK

Elvárás, hogy küslő annotációs sémát kezelni képes eszközt válasszunk.

5. Külső annotáció

A korpusz minden feldolgozását, így a mondatrabontást, a tokenizációt és a tulajdonnevek feljelölését is külso (standoff) annotációval jelöljük. Ennek lényege, hogy az eredeti dokumentumokat sima szövegként (UTF-8 v. Latin2) rögzítjük. Az annotációkat nem beágyazott XML-ként, hanem egy külső fájlban jelöljük úgy, hogy megadjuk, hogy az eredeti szöveg melyik karaktertartományára vonatkozik a címkézés, és hogy milyen címkét kap a szövegrészlet.

[2] `http://projects.ldc.upenn.edu/LowDensity/NE/tools/LDC_LoDLTools.exe`

Ennek elonye, hogy az annotálást teljesen különválasztjuk a használt feldolgozó eszközöktől, és közben minden információt megtartunk. Például a

```
<s><w>Ez</w> <w>egy</w> <w>mondat</w> <c> .</c>
<s><w>Meg</w> <w>a</w> <w>második</w> <c> .</c>
```

címkézett szövegből nem derül ki, hogy az első mondat utáni pont után volt-e szóköz, vagy nem, ami pedig fontos egy mondatrabontó tanításához. Magasabb szinten, ha a beágyazott XML-t választanánk, akkor egyetlen egy tokenizálóhoz kötnénk magunkat.

A standoff annotáció formátumáról szintén a projekt során kell dönteni.

6. A korpusz forrása

A korpusz elsődleges forrása magyar nyelven írt hírek: üzleti, belföldi és külföldi. Fontos, hogy valódi hírek teljes szövegét dolgozzuk fel, és ne vágjunk ki dokumentumokból darabokat.

Az Origo 2005-ös, a Magyar Hírlap 2002-es és az MTI hírei talán felhasználhatóak. Elsőért a MOKK, a másodikért a Nyelvtudományi Intézet, a harmadikért a Szegedi Egyetem tud felhasználási jogokat kérni.

A projekt során kell még eldöntenünk, hogy milyen további szöveg stílusokat veszünk be: turisztika, sport, művészeti stb. A MOKK javaslata:

műfaj	szövegszó (ezer)
gazdaság	100
sport	50
hazai politika	50
nemzetközi politika	50
törvények/rendeletek	50
tudomány/technika	50
bboard, blog	50
szoftver-kézikönyvek	50
filmszövegek, irodalom	50
összesen	500

A szövegeket úgy érdemes választani, hogy minden műfajban 1/5 rész angolból magyarra fordított szöveg legyen, mert ez fontos a gépi fordítási alkalmazásokhoz.

7. Jogok

Alapvető cél, hogy a létrejött korpuszt bárki teljesen szabadon használhassa, de ne publikálhassa módosított változatát (ugyanakkor új standoff annotációt hozzáadhat). Ennek egyetlen akadálya lehet, ha a hírforrások nem akarják letölthetővé tenni a híreiket.

8. Eldöntendő kérdések - feladatok

1. kell-e a projektkonzorciumot bővíteni?
2. mennyi erőforrás szánunk a korpuszra?
3. milyen jogosultsággal akarjuk a korpuszt később szabadon engedni (sok pénz esetén a konzorcium megtarthatná magának a korpuszt egy időre)
4. annotációs útmutató elkészítése (sok szakmai vita alapja)
5. annotáló szoftver kiválasztása
6. annotátorok keresése, szerződéskötés
7. korpusz forrásának és leendő méretének meghatározása
8. a nyers dokumentumok jogainak letisztázása (egységes szerződést köthetnénk)
9. végső korpusz séma rögzítése
10. a nyers szövegek előfeldolgozása (2plain text, tokenizálás, mondatra-bontás)
11. annotálás, utómunkálatok