Statisztikai koreferenciafeloldó rendszer magyar nyelvre – első eredmények

Munkácsy Gergely, Farkas Richárd

Szegedi Tudományegyetem, Informatikai tanszékcsoport
Szeged, Árpád tér 2.
rfarkas@inf.u-szeged.hu

Kivonat: Cikkünkben bemutatjuk az első statisztikai (gépi tanulás alapú) módszert koreferenciafeloldásra magyar nyelvű szövegekben. Ehhez a SzegedKoref korpuszon [1] tanítottuk a HOTCoref rendszert [2], majd a rendszer egyes moduljait alakítottuk át a magyar korpusznak megfelelően.

1. Bevezetés

A természetes nyelvű szövegekre jellemző, hogy kevés bennük az ismétlés, a szerzők törekednek a változatosságra, többféle kifejezést használnak ugyanarra az entitásra hivatkozásnál, pl. *kutya*, *eb*. Ezeknek a szövegeknek a megértéséhez szükség van arra, hogy tudjuk, hogy a kifejezések melyik másik kifejezésre utalnak, vagy épp melyek azok a szövegrészek, melyek azonos egyedre utalnak. Ez a koreferencafeloldás feladata.

Cikkünkben bemutatjuk az első statisztikai (gépitanulás-alapú) módszert koreferenciafeloldásra magyar nyelvű szövegekben. Ehhez a SzegedKoref korpuszon [1] tanítottuk a HOTCoref rendszert [2], majd a rendszer egyes moduljait alakítottuk át a magyar korpusznak megfelelően.

2. Korpusz

A SzegedKoref korpusz [1] egy magyar nyelvű, teljes mértékben kézzel annotált koreferenciakorpusz, mely azzal a céllal készült, hogy alapjául szolgáljon különböző statisztikai (adatvezérelt) algoritmusok tanításának és kiértékelésének. Az annotálás alapja a Szeged Korpusz volt, melyet egy újabb réteggel bővítettek. Ezek közül is azokat a szövegeken választották, amik viszonylag hosszabbak, az egy-két mondatos dokumentumokon kevésbé érdekes a koreferenciafeloldási feladat.

A SzegedKoref folyamatosan bővül, munkánkhoz a 2015 eleji változatot használtuk fel, mely két részből állt össze:

- Iskolai fogalmazások, elbeszélések
- Újsághírek

Az adatbázisból az egyszerűség kedvéért kísérleteinkhez töröltük a zéró névmásokat. Az így létrejövő adatbázisban lévő 400 dokumentum összesen 9 565 mondatot és 123 971 tokent tartalmaz. Ezekből 18 854 szerepel koreferencialáncban.

3. Koreferenciafeloldó rendszer

Munkánk alapjául a stuttgarti egyetemen fejlesztett HOTCoref (Higher Order Tree Coreference) program [2] szolgált, mely a CoNLL (Conference on Computational Natural Language Learning) Shared Task [3] adatain a legjobb eredményeket adja jelenleg. A HOTCoref első lépésben szabályok alapján kiválasztja a lehetséges anaforajelölteket, majd felügyelt gépi tanulási módszertant követve alakítja az említési láncokat, azaz azokat az említéscsoportokat amelyek egy entitásra vonatkoznak. A gépi tanulási megközelítés az egyes csoportokat látens faszerkezettel reprezentálja.

A HOTCoref magyar nyelvhez igazításának első lépése a jelöltek azonosítására szolgáló algoritmus honosítása volt. Ehhez statisztikákat gyűjtöttünk a tanító adatbázisban előforduló anaforákról (szófajok és konstituens-elemzésbeli nem terminális címkék).

A gépi tanulási rész itt is a jól megválasztott jellemzőkészleten áll vagy bukik. Ennek magyarra átalakításához a Szeged Treebank morfológiai és szintaktikai leírói alapján átírtuk a tulajdonnév, határozottság, számosságra utaló jegyeket. Egy másik fontos átalakítás a magyar ún. headFinder szabályok implementálása, mely a kifejezések fejét keresi meg. Ez egy olyan szabályrendszer, amely megadja, hogy ha egy utalás több szóból áll, akkor abból melyik a legfontosabb. Ezt a mondat konstituenselemzése alapján döntöttük el. Például a *nagy piros labda* kifejezés esetén a *labda* token a fej. Végül töröltük a magyarban nem értelmezhető jellemzőket (pl. gender).

Megvizsgáltuk továbbá, hogy a ragozott alakok hordoznak-e hasznos információt a koreferenciafeloldási feladatban. Azt tapasztaltuk, hogy az eredmények drasztikusan (átlagos 15 százalékponttal) emelkedtek, ha szótöveket használunk a lexikai jellemzők kinyerésekor. Ennek magyarázata kettős. Egyrészt a tanító adatbázis viszonylag kicsi, nagyon alacsony az egyes ragozott alakok gyakorisága, ami a jellemzők extrém ritkaságát vonja maga után. Másrészt az említések összerendelése elsősorban szemantikai, és nem morfoszintaktikai alapon dönthető el.

4. Eredmények

Ellentétben más problémáknál használt leszámolós módszerekkel, egy koreferenciajelölés pontosságának értékelése vitatott feladat. Sokfajta metrika létezik, melyek különbözően jellemzik az egyes mintákat. Viszont az, hogy melyiket érdemes használni, az nem nyilvánvaló. A nemzetközi trendet követve négy különböző metrika (MUC, BCUC, CEAFM, CEAFE) mellett is kiértékeltük[1] a koreferenciafeloldót [3].

[1] a kiértékelő szkript elérhető: http://conll.cemantix.org/2011/software.html

Az átalakításokkal az alábbi eredményeket éri el a rendszer tökéletes (gold standard) morfoszintaktikai jelölések mellett:

	Fedés	Pontosság	F1
MUC	35,69	49	41,3
BCUB	34,21	49,09	40,32
CEAFM	40,47	54,45	46,43
CEAFE	39,11	50,72	44,16
Átlag	37,37	50,815	**43,0525**

Összehasonlításként egy ugyanekkora tanító adatbázist használó angol rendszer eredményei:

	Fedés	Pontosság	F1
MUC	64,85	64,43	64,64
BCUB	50,44	52,39	51,4
CEAFM	54,98	54,94	54,96
CEAFE	47,38	48,21	47,79
Átlag	54,4125	54,9925	**54,6975**

5. Összegzés

Ezek az eredmények első, de megismételhető empirikus eredmények. Az átalakított HOTCoref rendszert kérésre bárkinek odaadjuk. Számos ponton javítható még a rendszer a jövőben, például szemantikai információk (WordNet, tulajdonnév-kategorizáció stb.) beépítésével.

Köszönetnyilvánítás

Farkas Richárd kutatásait az MTA Bolyai János ösztöndíja támogatta.

Bibliográfia

1. Farkas R., Vincze V., Hegedűs K.: SzegedKoref: kézzel annotált magyar nyelvű koreferenciakorpusz. In: XI. Magyar Számítógépes Nyelvészeti Konferencia, Szeged (2015) 312–319
2. Björkelund, A., Kuhn, J. Learning structured perceptrons for coreference resolution with latent antecedents and non-local features. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics (2014) 47–57