Angol címek felismerése

Pohl Gábor[1], Ugray Gábor[2]

[1] Pázmány Péter Katolikus Egyetem, Információs Technológiai Kar
1083 Budapest, Práter utca 50/A
pohl@itk.ppke.hu

[2] MorphoLogic Kft.
1126 Budapest, Orbánhegyi út 5.
ugray@morphologic.hu

Kivonat: Angol szövegek szintaktikai elemzése során nehézséget jelentenek a címekre jellemző sajátos nyelvtani szerkezetek: hiányos névelőhasználat, egyes segédigék elhagyása stb. A probléma a szintaktikai szabályok paraméterezésével küzdhető le anélkül, hogy elfogadhatatlan kompromisszumokat kelljen hozni akár a címek lefedettségét, akár a jólformált mondatok pontosságát illetően. A szöveget alkotó szegmensek besorolására aszerint, hogy címről van-e szó vagy sem, egy döntési fát tanítottunk be weboldalakról gyűjtött, kézzel osztályozott két korpusz segítségével. Az egyik korpuszon tanítva és a másikon tesztelve a módszerrel 95% körüli pontosságot sikerült elérni. Az eredmény megfelelőnek bizonyult ahhoz, hogy beépítsük a MorphoLogicnál fejlesztett angol-magyar gépi fordítórendszerbe. A cikkben ismertetjük a nyelvtani probléma természetét, a tanításhoz használt korpuszokat, a relevánsnak bizonyult tulajdonságokat és a tanítási módszert, valamint annak értékelését.

1 Angol címek felismerése és elemzése

1.1 Szintaxis

Az angol címek szintaxisa jelentős mértékben különbözik a közönséges angol mondatokétól. Tekintsük az alábbi két, címben illetve folyó szövegben található előfordulást:

1. Driver killed in accident

2. A driver has been killed in an accident.

Jól látható, hogy a címben használatos formából hiányoznak a határozatlan névelők, a *has* segédige ragozott alakja, a passzív szerkezethez tartozó *be*, valamint a mondat végén nincs írásjel. A *driver* mint determinálatlan, egyetlen megszámlálható főnévből álló főnévi csoport a szabályok megengedőbbé tételével még lefedhető lenne, ez az irány azonban önmagában véve végzetes túlgenerálást eredményez a bonyolult, ámde jólformált mondatok esetében. A sikeres elemzéshez tehát meg kell különböztetni a címeket a folyó szövegtől.

1.2 Paraméterezhető nyelvtan

A MetaMorpho rendszer [2][4] nyelvtani elemzőjének [3] fejlesztésekor a problémára új, a fordítóprogramokban ismereteink szerint eddig nem alkalmazott megoldást dolgoztunk ki. Az ötlet azon a felismerésen alapul, hogy ha egy mondatról még a fordítás előtt el lehet dönteni, hogy cím-e, akkor utána a nyelvtan paraméterezhető ennek megfelelően.

A nyelvtan paraméterezése a gyakorlatban annyit jelent, hogy az elemzendő szövegről a nyelvtanon kívül kiértékelt információk feltételként befolyásolják egyes szabályok alkalmazását. Amennyiben például a felismerő a bemenetként kapott szövegről megállapította, hogy az cím, a fent leírt determinálatlan NP-t létrehozó szabály működése engedélyezett, ellenkező esetben nem.

1.2 *Dokumentumformátumok*

Modern dokumentumformátumok esetében azt várhatnánk, hogy a címeket valamilyen speciális módon megjelölik a szerzők. A gyakorlatban azonban még ha létezik is ilyen formátumfüggő jelölési lehetőség, annak használata nem konzisztens. A különböző dokumentumformátumokban fellelhető eltérő jelölések felismerése gyakorlatilag megvalósíthatatlan feladat a dokumentumtípusok sokfélesége és zártsága miatt, ráadásul elterjedt dokumentumtípusokban (pl. PDF, PostScript) nincsenek is ilyenek. Ezért amellett döntöttünk, hogy a formázatlan szöveg (plain text) alapján próbáljuk egy osztályozó segítségével megkülönböztetni a címeket és a nem címeket.

2 Korpuszalapú címosztályozás

2.1 Döntési fa

Szabályalapú osztályozó létrehozásához nem állt rendelkezésünkre elég ismeret ahhoz, hogy a címként fordítandó szövegrészeket a nem címként fordítandóktól meg tudjuk különböztetni, ezért egy gépi tanuló algoritmus korpusz alapú tanítása mellett döntöttünk. A lehetséges gépi tanuló módszerek közötti választáskor fontos szempont volt, hogy az offline tanított osztályozó működése átlátható, illetve a későbbiekben könnyen – és lehetőleg szabályalapú változatban – beépíthető legyen a fordítórendszerbe, azaz ne feketedobozként működjön. A döntési fák ismert rossz tulajdonságaik (pl. instabilitás [1]) mellett is a legjobb választásnak tűntek, mivel egyszerre képesek nominális és numerikus tulajdonságok (feature-ök) kezelésére, valamint az eredményül kapott döntési fa ember által könnyen olvasható, procedurális programozási nyelveken könnyen implementálható. Egy lecsupaszított (pruned) döntési fa alkalmazása mellett szólt, hogy nem láthattuk előre, a szöveg szegmenseiből esetlegesen kinyerhető tulajdonságértékek közül melyek fogják majd befolyásolni döntésünket, azaz a releváns tulajdonságok meghatározását is a döntési fa tanító algoritmusra bíztuk.

A kísérlethez a WEKA[3] gépi tanulórendszer [6] *J48* döntési fa osztályozóját választottuk, amely Quinlan C4.5 döntési fa tanuló algoritmusán [5] alapul.

2.2 A címek megkülönböztető tulajdonságai

A döntési fa tanításához meg kellett határoznunk, hogy a tanítóhalmazba az egyes szövegegységek milyen tulajdonságait vegyük fel, illetve hogy mit válasszunk szövegegységnek. Az utóbbi döntés esetében a mondat és a bekezdés közül a bekezdést választottunk vizsgálatunk alapjául, mivel megfigyeléseink azt mutatták, hogy a címek mindig külön bekezdésbe kerülnek, azaz nincsenek, vagy nagyon ritkák a cím és nem cím szövegrészeket vegyesen tartalmazó bekezdések. A bekezdésszint választása azzal az előnnyel is együtt járt, hogy nem kellett számolnunk a címek (azaz rendhagyó tulajdonságokat mutató mondatok) esetében ismeretlen pontossággal működő mondatszegmentáló modul hibáival. Vizsgálatunk alapjául így a bekezdések következő tulajdonságait választottuk:

- szavak száma;
- záró írásjel (illetve ennek hiánya);
- névelők aránya;
- *be* és *have* különböző alakjainak aránya;
- nagybetűvel kezdődő szavak aránya.

Első kísérleteinknél a fentieken kívül a mondatok gépi mondatszegmentáló segítségével meghatározott számát is rögzítettük.

2.3 Tanító és kiértékelő minták előállítása

Mivel a MetaMorpho fordítórendszer fő célkitűzései közé tartozik weboldalak angolról magyarra fordítása, egy weblapokból készített kisméretű korpuszt hoztunk létre a kísérleteinkhez. Két hírportálról (cnn.com és newyorker.com) gyűjtöttünk össze oldalakat, melyekből kinyertük a bekezdésként értelmezhető szegmenseket és minden bekezdés esetében kézzel meghatároztuk, hogy cím-e. A CNN korpusz 776, a NEWYORKER korpusz 1739 bekezdést tartalmazott. A korpusz bekezdéseihez az előző pontban részletezett tulajdonságokat meghatározva készítettünk a WEKA rendszerben alkalmazható ARFF formátumú mintahalmazt. Mindkét mintahalmaz több címet tartalmazott, mint nem címet. Ez az ilyen jellegű weboldalak tulajdonsága, hiszen minden oldalon több másikra hívják fel címek segítségével a figyelmet.

2.4 Tanítás és kiértékelés

Az egyes adathalmazokkal külön-külön illetve az adathalmazokat kombinálva is betanítottunk döntési fákat. Először a CNN korpuszon tanítottuk és teszteltük keresztkiér-

[3] Waikato Environment for Knowledge Analysis

tékeléssel (tenfold cross-validation) a döntési fát. Az eredmény meglepően jó volt, a CNN korpusz bekezdéseinek 96,7%-át helyesen osztályozta a keresztkiértékeléssel tanított döntési fa. A fát megvizsgálva meglepve tapasztaltuk, hogy csupán egyetlen, a bekezdések szószámban mért hosszára vonatkozó szabállyal 95%-os pontosság volt elérhető. A NEWYORKER korpuszon tesztelve a fát (most nem keresztkiértékeléssel tanítva) 90,6% volt a jó döntések aránya, csak egy szabályt alkalmazva viszont 91%-volt, ami az mutatja, hogy a döntési fa a CNN korpusz tulajdonságait túlzottan pontosan megtanulta, a NEWYORKER korpuszon így a kevésbé speciális egyetlen szabály jobbnak bizonyult.

A NEWYORKER korpusz már egy kicsit nehezebben klasszifikálhatónak tűnt. Keresztkiértékeléssel 93,7%-os osztályozási pontosságot ért el a fa, egyetlen, a bekezdés hosszára vonatkozó szabályt alkalmazva pedig 92,3%-ot. A CNN korpuszon tesztelve a fát 96,1% illetve 93% (csak egy szabályt alkalmazva) volt a helyesen osztályozott bekezdések aránya, ami igen jónak tekinthető.

A két korpusz unióján tanítva a fát az eredmény túl bonyolultnak tűnt. A paramétereket változtatva a fát sikerült redukálni, ugyanakkor azt tapasztaltuk, hogy a NEWYORKER korpuszon tanított jóval kisebb méretű fa is azonosan jó, 94,8%-os pontosságot ért el. A fa egyszerűsége miatt a túltanulás veszélye is jóval kisebb volt ebben az esetben, így a kísérlet végén ezt, az 1. ábrán látható fát találtuk legjobbnak.

Fontos megjegyezni, hogy minden döntési fánál a bekezdések szavakban mért hossza volt a legfontosabb döntési tényező; a többi tényező meglehetősen inkonzisztens módon változott a különböző fákat tekintve. Ez utóbb betudható a tanítóhalmazok különbözőségének, a döntési fa tanulás instabilitásának, illetve annak, hogy a tényezők súlya meglehetősen kicsi volt. A NEWYORKER korpuszon tanított fa esetében szerencsés, hogy a többi döntési tényezőhöz képest nehezebben meghatározhatók (a *be* és *have* különböző alakjainak aránya, a névelők, illetve a nagybetűs szavak aránya) estek ki a fa visszavágása (pruning) során.

```
wordCount <= 8: 1 (1133.0/70.0)
wordCount > 8
|   wordCount <= 18
|   |   endingMark = NONE
|   |   |   wordCount <= 11: 1 (30.0/3.0)
|   |   |   wordCount > 11: 0 (15.0)
|   |   endingMark = quest_m: 0 (19.0/4.0)
|   |   endingMark = excl_m: 0 (0.0)
|   |   endingMark = punct
|   |   |   wordCount <= 16: 0 (111.0/18.0)
|   |   |   wordCount > 16
|   |   |   |   wordCount <= 17: 0 (4.0/1.0)
|   |   |   |   wordCount > 17: 1 (3.0)
|   |   endingMark = colon: 0 (1.0)
|   wordCount > 18: 0 (423.0/4.0)
```

1. ábra: a NEWYORKER korpuszon tanított döntési fa

2.4.1 Következtetések

A kísérlet végén levonhattuk a következtetést, hogy az angol címek általában rövid bekezdések, rövidebbek 9 szónál, a nem cím bekezdések pedig hosszabbak. Utólag nem is tűnik különösnek ez az állítás, az viszont igen, hogy csupán ezzel az egy szabállyal 90% fölötti osztályozási pontosságot lehet elérni. Köszönhető ez annak, hogy ritkák az egyetlen rövid mondatot tartalmazó nem cím bekezdések, talán csak dialógusokban szerepelnek ilyenek, ezek elemzésére pedig – a címekhez hasonlóan hiányos szerkezetük miatt – lehet, hogy amúgy is szerencsésebb a címnyelvtant választani. A kísérlet eredménye arra is rámutat, hogy szerencsés választás volt a döntést bekezdésszinten meghozni.

2.5 A döntési fa alapú osztályozó modul alkalmazása

A döntési fa alapú címosztályozó C++ implementációját beépítettük a MorphoLogic MetaMorpho fordítórendszerébe, amelyhez a cím, illetve közönséges mondatok elemzésére és fordítására alkalmas alternatív nyelvtani szabályrendszereket Ugray Gábor és Merényi Csaba (MorphoLogic) dolgozták ki.

Referenciák

1. Li, Ruey-Hsia; Belford, Geneva G.: Instability of decision tree classification algorithms. Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining, 2002.
2. Prószéky Gábor; Tihanyi László: MetaMorpho: A Pattern-based Machine Translation Project. In: Proceedings of the 24th 'Translating and the Computer' Conference. London, United Kingdom, 19–24 (2002)
3. Prószéky Gábor; Tihanyi László; Ugray Gábor: Moose: A Robust High-Performance Parser and Generator. EAMT Workshop, Malta, 2004
4. Tihanyi László: A MetaMorpho projekt története. Magyar Számítógépes Nyelvészeti Konferencia 2003, Szeged.
5. Quinlan, J. Ross: C4.5: Programs for Machine Learning. Morgan Kaufmann, San Francisco, 1993.
6. WEKA (Waikato Environment for Knowledge Analysis)
   Web page: http://www.cs.waikato.ac.nz/~ml/weka/index.html)
   Book: Witten, Ian H.; Frank, Eibe: Data Mining: Practical machine learning tools with Java implementations. Morgan Kaufmann, San Francisco, 2000.