Намиране на следите в изречения на италиански език чрез статистически методи за машинно обучение и банка от дървета

Тодор Арнаудов

ФМИ на ПУ
Пловдив 4003, бул. България 236
todprog@yahoo.com

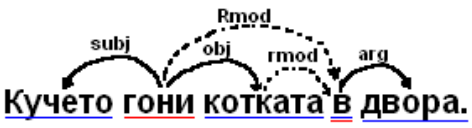
Резюме

В настоящата работа разказваме за опит за намиране на следите в изречения на италиански език чрез обучение на статистическия PoS-тагер TnT на Thorsten Brants (Brants, 1999) с обърната банка от дървета на Университета в Торино (Bosco, 2003). Работата започва със запознаване на читателя с понятията нелокална зависимост, граматики на зависимостите, празни елементи. Даваме пример за важността на разпознаването на нелокални зависимости за семантично правилния машинен превод. Продължаваме с дефиниция на "следа" и с обяснение на понятието "проективност" на синтактично-семантична структура от зависимости, във връзка с преобразуването ѝ в конституентно синтактично дърво. Споменаваме статистически методи за разпознаване и възстановяване на нелокални зависимости и препращаме читателя към докторската дисертация на Pèter Dienes (Dienes, 2003) от Университета на Заарланд в Заарбрюкен (Universität des Saarlandes, Saarland University), която предлага методи за разрешаване на нелокални зависимости за английски език. Завършваме с доклада за нашия експеримент и с резултатите получени по метода 10-fole cross-validation.

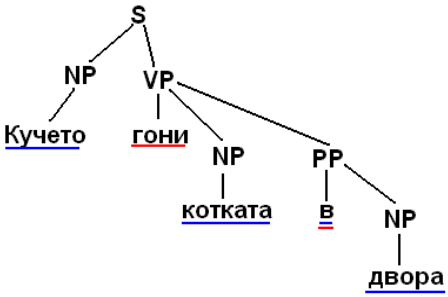
1. Нелокални зависимости, граматики на зависимостите

Намирането на следите в изречения е подзадача на по-общата задача за разрешаване на нелокални зависимости. Нелокалните зависимости (НЛЗ) са семантични и граматични релации между думи, когато зависимите думи, наречени "главна" и "зависима" (подчинена), са отделени една от друга от други думи, които не участват във въпросната релация. За изразяване и разпознаване на такива отношения се използват т.нар. граматики на зависимостите.

Представянето на структурата на изречение чрез зависимостите между думите в него, съдържа информация за всички отношения предикат-аргумент, включително нелокалните. Това е много полезна особеност, която не е заложена явно в конституентните представяния, и която прави граматиките на зависимостите много по-удобни за моделиране на езици със свободен словоред, какъвто е италианският. При езиците със свободен словоред отношенията предикат-аргумент не следват толкова просто от наредбата на думите, колкото при езиците със строг словоред, и нелокалните зависимости са с по-висока честота.



Фиг.1 – пример за представяне на изречение с граматика на зависимостите. С червено са подчертани главните думи, а със синьо - зависимите. Вижда се, че е възможно една дума да бъде едновременно и главна, и зависима.



Фиг.2 – конституентно представяне на същото изречение:

1.1. Примери за нелокални зависимости в изречения на английски език по (Dienes, 2003). "Празни елементи" .

Английският език като цяло се смята за език със строг и подреден словоред, при който аргументите на предикатите се появяват в добре определена последователност на съседство помежду си. Въпреки това, за него са характерни нелокални зависимости, които не следват това общо правило.

*It is difficult to understand what **I** want to **do**. (Dienes, 2003)*

Тук фразата от една дума "I", която е непосредствено подчинена на "do", е отделена от своята главна дума от последователността "want to"; при това, нито една дума от последните две не е зависима от "do".

Разпознаването на този вид нелокални зависимости е много важно в практиката, първо защото НЛЗ често са семантични аргументи в изречението и като следствие от това биха могли да покажат структурата предикат-аргумент; от друга страна, познаването на структурата от локални и нелокални зависимости в изречението е полезно в почти всички области от обработката на естествен език. Бихме изброили отговаряне на въпроси, извличане на информация, машинен превод, моделиране на езика, извличане на речници и пр.

Пряката практическа полза от разпознаването на НЛЗ ще илюстрираме като цитираме експеримента на Dienes (Dienes, 2003), направен с изречения, в които има НЛЗ от горния вид. По-точно:

I promised John **[EE = I]** to remember you.
I find it difficult **[EE = I]** to remember you.

В превод на български и с явно изразяване на имплицитно заложената в думите семантика (в квадратни скоби), горните изречения биха звучали приблизително така:

[Аз] Обещах на Джон [аз] да си спомням за теб.
Трудно ми е [на мене] да си спомня за теб.

Нито една от тестваните системи за машинен превод към немски и френски в работата на (Dienes, 2003) не успява да разпознае нелокалната зависимост едновременно и в двата случая.

Общата семантична грешка, която допускат всички системи, е че породеното от тях изречение явно изразява смисъла "I promised John **he to remember** you", вместо "**I to remember** you".

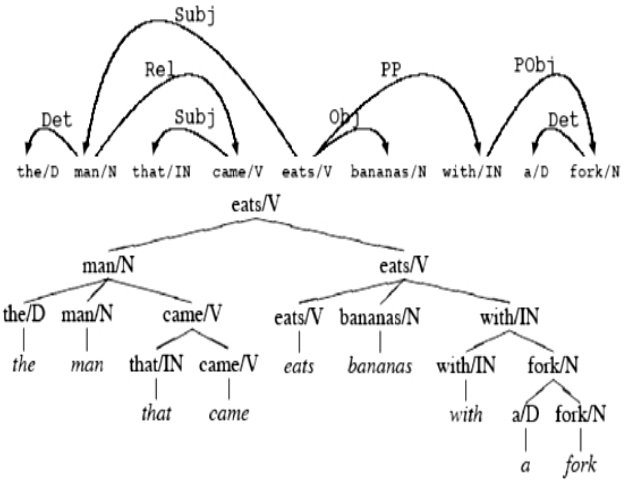
Тествахме една от системите – Systran - сега, през 2006 г., и установихме че допуска същата грешка.

Вмъкнахме този пример не само за да покажем важността на разпознаването на НЛЗ в практиката, а и за да въведем категорията "празен елемент" (empty element, EE), означена по-горе с **EE.** Празните елементи са виртуални заместители на липсващи, но семантично подразбиращи се елементи от дървото на синтактичния разбор, които са въведени по-рано в изречението. В примерите на английски, празният елемент и в двата случая е заместител на "I". "I", която неявно въвежда празния елемент, се нарича **предшественик (antecedent)** на този празен елемент.

2. Следи, намиране на следи , проективност и линеаризиране

"Следа" (trace) наричаме релацията между празен елемент и предшественик**.** Намирането на следа е откриване на местата, където има празен елемент, и възстановяване на правилния предшественик (глава) в дървото на представянето с граматика на зависимостите.

За да може предсавянето с граф на зависимости да бъде еднозначно преобразувано в дърво на синтактичния разбор в класическо конституентно представяне, се налагат някои ограничения върху структурите, представящи изречения чрез зависимости (Schneider, 2003), (Dienes, 2003). Едно от основните ограничения е да няма пресичащи се връзки на зависимости. Когато структурното представяне спазва тези ограничения, то се нарича "проективно" и от него по детерминиран начин може да се изведе конституентното представяне на изречението.



Фиг.3 - Проекция - пример на (Schneider, 2003).

Неспазването на условията не би позволило пораждане на детерминирано конституентно представяне, защото биха се появили връзки между листата, т.е. би се нарушила дървовидната структура в разгърнатото конституентно представяне.

Трябва да се има предвид, че структурите при граматиките на зависимостите не са проективни по начало, но могат да бъдат направени такива чрез подходящи преобразования. (Dienes, 2003)

3. Статистически методи за разпознаване и възстановяване на нелокални зависимости чрез машинно обучение

Методите за статистически синтактичен анализ се разработват и усъвършенстват от различни изследователи още от началото на 80-те години. Само ще спомена, че освен граматики на зависимостите, се използват още вероятностни безконтекстни граматики (Probabilistic Context-Free Grammars), Вероятностна граматика на фрази с прекъсната структурата (Probabilistic Discontinuous Phrase Structure Grammar); граматики със съединяване, размяна и вмъкване на дървета (Tree-Adjoining, Tree-Substitution, Tree-insertion); унификационни граматики и др.

Препоръчаме на любознателния читател изчерпателното описание в докторската дисертация на Dienes (Dienes, 2003), където той е представил и своя разрешител на нелокални зависимости за английски език. Механизмите на действието на разрешитля на

Dienes се основават на съчетание между прости, но добре настроени крайни автомати и безконтекстни граматики.

4. Намиране на следите в изречения на италиански чрез PoS-тагера TnT (Brants, 1999)

Не са ни известни опити за автоматично търсене на следи в изречения на италиански.

В нашата работа използвахме по-ограничена дефиниция за "следа" и поставихме по-леки цели по намирането на следи, отколкото в дефиницията от т. 2. В нашия случай целта беше да открием мястото на празните елементи, без да търсим предшествениците им.

Вместо да имплементираме самообучаващи се статистически алгоритми от самото начало, опитахме да използваме вездесъщия адаптивен PoS-tagger TnT на Торстен Брантс (Brants, 1999).

Обучихме TnT да разпознава следи в италиански изречения чрез готова банка от дървета. За обучение използвахме банката от дървета на университета в Торино (Turin University Treebank, TUT), която е построена върху граматика на зависимостите. Изреченията в банката са проективни и в нея се съдържа и информация за местата на следите (Bosco, 2003).

Форматът на TUT е четириколонен, а PoS-тагерът TnT работи с двуколен формат. За да обучим TnT, трябваше да извършим преобразуване и да опростим представянето на банката.

Чрез скриптове на Perl генерирахме "равен" корпус без тагове, и опростен двуколонен корпус с тагове, в който думите, които са следи според TUT, са отбелязани с "Yes"; а думите, които не са следи – с "No".

Чрез TnT създадохме 10 модела, всеки от които бе обучен върху 90% от обема на банката от дървета с тагове Yes/No. Използвахме останалите съответни 10% от банката за проверка на валидността на резултатите от работата на TnT по метода 10-fole cross-validation.

Десет пъти прилагахме тагера върху съответните 10%-ови извадки от корпуса без тагове за следи с моделите, обучени върху 90-те % от корпуса.

4.1. Резултати в цифри

93.78	94.03	93.07	93.05	93.86
93.17	93.62	93.35	92.83	93.52

Средно: 93.428%

Табл.1: Общ сбор на съвпаденията на двата вида тагове "Yes" и "No" между оригиналния корпус и файла, анотиран от TnT.

За съжаление височината на процентите в Табл.1 са заблуждаващи, поради значителното преобладаване на тагове "No" над таговете "Yes" в банката от дървета - над 15:1. В същото време броят на таговете за следа "Yes" в анотираните от TnT файлове в някои от 10-те случаи пада до под 1/10 от броя на таговете "Yes" в оригиналния корпус. От поставените от TnT малко тагове "Yes", верни са едва около 1/10.

В крайна сметка се стигна до много нисък резултат, като в един от случаите TnT дори не успя да познае нито една следа.

1/247	1/238	2/266	0/272	1/238
2/266	1/245	1/261	1/276	1/262

Средно: 0.43%

Табл.2: Съвпадения на тагове "Yes" между оригиналния корпус и файла, анотиран от TnT.

79.5	79.5	80.7	80.9	79.1
79.9	79.6	78.7	80.4	81.3

Средно: 79.96%

Табл.3: Съвпадения на тагове "No" между оригиналния корпус и файла, анотиран от TnT.

4.2. Защо?

Първото което можем да предположим като вероятна причина за неуспешния експеримент, е недостатъчен обем на използваната банка от дървета, тъй като размерът ѝ е по-малко от 45000 реда.

Имаме известни съмнения за особеност на формàта на банката от дървета, която може би отчасти е объркvала построяването на правилни вероятностни модели за следите.

Става въпрос за наличието на следи към приложения (appositions), анотирани на 3 реда с две скоби и цифра. Напр.:

```
t     (...)
(     PUNCT 17      OPEN+PARENTHETICAL
1     NUM.1  17     APPOSITION
)     PUNCT 17      CLOSE+PARENTHETICAL
```

Така при "преподаването" на банката на TnT, се получава таг за следа на знака "(", защото не е определено как би трябвало да се разглежда въпросната ситуация.

В оригиналния корпус се срещат и отделни анотации за следа преди знак за край на изречение, което в корпуса за обучение изглежда така, сякаш знакът точка е следа.

5. Заключение

Експериментът за намиране на следи в изречения на итлиански език чрез адаптивния PoS-тагер TnT и банката от дървета на Университета в Торино беше неуспешен. Постигнотото съвпадение на таговете по метода 10-fole validation е едва 0.43%.

Не можем да дадем категоричен отговор на въпроса каква е причината за резултатите да бъдат толкова лоши, но имаме предвид два предполагаеми фактора:

1. **Малкия размер на банката от дървета.**
2. **Вероятна неприложимост на този метод за разпознаване на следи в изречения на италиански език**

6. Забележки

Изследването е извършено като курсов проект по избираемата дисциплина *"Perl за лингвисти*" през зимния триместър на учебната 2005/2006 във ФМИ на Пловдивски университет "Паисий Хилендарски" при преподавателя Атанс Чанев, докторант по Когнитивни науки в Университета в Тренто, Италия. Благодаря му за напътствията и помощта.

Тази работа не би съществувала и без съдействието на декана на ФМИ доц.д-р. Димитър Мекеров, който осигури факса, с който се свързах с Торстен Брантс, така че да получа разрешение да ползвам TnT.

Тодор Арнаудов Пловдив, 17.3.2006

7. Литература

1. Brants, Thorsten (1999). TnT – A Statistical Part-of-Speech Tagger.
2. Dienes, Pèter (2003). Statistical parsing with non-local dependencies.
3. Bosco, Cristina (2003). A grammatical relation system for treebank annotation.
4. Schneider, Gerold (2003). Learning to Disambiguate Syntactic Relations. *Linguistik online 17/5/03,* http://www.linguistik-online.de/17_03/schneider.html
5. *TUT (Turin University Treebank) - http://www.di.unito.it/~tutreeb/*