Класове думи
и
поставяне на тагове на части
от речта

Съдържание

1. Класове думи
2. Поставяне на тагове на частите на речта
   a. Алгоритъм за поставяне на тагове HMM
   b. Поставяне на тагове, базирано на трансформационни правила
   c. Множествени тагове и съставни думи
   d. Непознати думи
3. Клас- базирани N-грами

Класове думи

Класове думи

- Дума - звуков комплекс, свързан с определен смисъл, притежаващ определени граматически характеристики.
- Морфологията - граматическото значение на думите .
- Части на речта - думите в морфологията.

Класове думи
Части на речта

- Съществително име - назовава лица и предмети;
- Прилагателно име - назовава качество на лица и предмети;
- Числително име - показва брой или поредност на лица и предмети;
- Местоимение - замества някакво име;
- Глагол - означава действие или състояние;

Класове думи

Части на речта

- Наречие- пояснява някакво действие;
- Предлог- свързва думите;
- Съюз- свързва думите и изреченията;
- Частица- подсилва израза или с която се образуват думи и форми;
- Междуметие- се изразяват непосредствени чувства или се имитират природни, животински и други звукови шумове.

Класове думи

Супер категории на частите на речта

- Затворен тип - имат относително фиксиран брой членове(предлозите са затворен клас, защото има фиксиран брой от тях в българския език).
- Отворени класове - съществителни, глаголи, прилагателни и наречия.

Класове думи

Съществителни имена. Видове

- нарицателни - назовават всички предмети или същества от един и същи вид (*стол*, *маса*, *ученик*
- собствени -назовават един предмет или едно същество (*Пламен*, *София*);
- конкретни -назовават материални, веществени неща (*маса*, *стол*, *легло*).
- абстрактни-назовават нематериални, духовни неща (*любов*, *омраза*, *идея*) ;

Класове думи
Съществителни имена. Видове

- Обикновени - назовават единични предмети или лица, които може да се броят (*стол*, *маса*, *ученик*).
- Събирателни - назовават множество от предмети, лица и други които не можем или не искаме да броим, а ги възприемаме като нещо цяло (*клас*, *група*, *мравуняк*).

Класове думи

Съществителни имена. Характеристики

- Род : мъжки, женски и среден.
- Форми за определеност- морфологически показател за определеност е наличието на специална морфема, наричана определителен член.
- Членуване - образуването на форми с определителен член. Съществуват пълен (*градът*, *кралят*, *жените*) и кратък член (*народа*, *краля*, *жена*).

Класове думи
Съществителни имена. Характеристики

- Звателна форма – специална форма за обръщение на същ. имена от мъжки и женски род, единствено число (*Иван – Иване, господин – господине, учителка – учителко, Елена - Елено*).

Съществително име (N)

№	Х-ка	Стойност	Пример	Таг
1	Вид	нарицателно	книга	c
		собствено	Иван	p
2	Род	мъжки	стол	m
		женски	маса	f
		среден	вретено	n
3	Число	единствено	момче	s
		множествено	столове	p
		бройно	(два) стола	t
4	Падеж	именителен	народ	n
		звателен	народе	v

Класове думи
Глагол. Видове

- Преходни (транзитивни) – има пряко и непряко допълнение (*Ученикът пише писмо.*; *Ученикът пише с писалка.*).
- Непреходни (нетранзитивни) – има само непряко допълнение (*Ученикът отговаря на учителя.*).
- Лични глаголи – има лице, вършител на действието (*Иван чете книга.*).
- Безлични глаголи – липсва лице, вършител на действието (*Вали.*).

Класове думи
Глагол. Видове

- Обикновени (невъзвратни глаголи) – означават действие, което не се извършва върху вършителя, а е насочено към други лица и предмети (*Ученикът мие чашата.*).
- Възвратни глаголи – означават действие, което се извършва върху самия вършител (*Ученикът се мие.*).
- Спомагателни глаголи – служат за образуване на глаголни форми (*съм*).

Класове думи
Глагол. Граматически категории

- Лице (1-во, 2-ро, 3-то)
- Число (ед.ч. и мн.ч.)
- Време:
  - сегашно – *чета*;
  - минало свършено – *писах*;
  - бъдеще – *ще пиша*;
  - минало несвършено – *пишех*;

Класове думи
Глагол. Граматически категории

- Наклонение
  - изявително (индикатив) – *Аз работех в градината.*
  - повелително (императив) – *Пишете!*
  - условно – *Бих си купил тази книга.*
  - преизказно – *Той пишел писма всяка седмица.*

Класове думи
Глагол. Граматически категории

- Залог
  - деятелен – *Учителят изпитва ученика.*
  - страдателен – *Книгата е прочетена от ученика.*

Глагол (V)

№	Х-ка	Стойност	Пример	Таг
1	Вид	основен	говоря	m
		спомагателен	съм	a
2	Наклонение	изявително	говоря	i
		повелително	говорете	m
		условно	говорил	p
		преизказно	говорейки	g
3	Време	сегашно	говоря	p
		минало несв.	говорех	i
		минало неопр.	говорено	s
		минало св.	говорих	a
4	Лице	първо	говоря	1

Класове думи
Прилагателно име. Видове

- качествени - означават цвят, вкус, размер, форма, физически качества, духовни качества.
- относителни - означават свойство свързано с веществен произход, местен произход, принадлежност, предназначение.
- Прилагателните имена се изменят по род и число.

Прилагателно (А)

№	Х-ка	Стойност	Пример	Таг
1	Вид		-	-
2	Степен		-	-
3	Род	мъжки	бедният	m
		женски	бедна	f
		среден	бедното	n
4	Число	единствено	бедния	s
		множествено	бедните	p
5	Падеж			-
6	Членуване	нечленувано	бедна	n

Класове думи
Наречие

- пълнозначна дума (има собствено лексикално значение).
- самостойна дума (може да има самостоятелна служба в изречението).
- неизменяема дума(не се изменя с помощта на окончания, няма различни форми).

Наречие (R)

№	Х-ка	Стойност	Пример	Таг
1	Вид	основни местоименни	тук умно	g a
2	Степен			-
3	Число			-
4	Лице			-

Класове думи
Предлози

- Могат да свързват:
  - глагол със съществително име (*Чета за изпит.*);
  - съществително със съществително (*Сок от малини*);
  - съществително с прилагателно (*Дъх на хубав парфюм*);
  - глагол с прилагателно име (*Говоря на малкото дете*).

Класове думи
Предлози

- Примери- *без, в, вместо, въпреки, върху, до, за, зад, заради, като, край, към, между, на, над, освен, от, по, поради, под, преди, през, при, с, след, според, спрямо, сред, срещу, у, и чрез.*
- Предложни изрази – *предвид на, въз основа на, независимо от* и други.

Предлози (S)

№	Х-ка	Стойност	Пример	Таг
1	Вид		на, в	р
2	Форма			-
3	Падеж			-

Класове думи
Съюзи. Видове

- Свързват думи или еднакви части в простото изречението и прости изречения в сложното.

Различаваме:

- истински съюз (*а, но, че* и други);
- местоимения съюзи(въпросителни и относителни местоимения – *кой, който* и други);
- наречия съюзи (въпросителни и относителни наречия – *как, както, къде* и *където*).

Съюзи (С)

№	Х-ка	Стойност	Пример	Таг
1	Вид	съчинителни	а, ала, или	c
		подчинителни	че, да	s
2	Форма	проста	а, и, че	s
		сложна	и да, за да	c
3	Подтипове			-
4	Число			-
5	Лице			-

Класове думи
Частици. Видове

- показателни – *ето*, *ей*, *хей*;
- призивни – *бе*, *брей*, *бре*;
- въпросителни – *ли*, *дали*, *нали*, *нима*;
- потвърдителни – *да*, *аха*, *зер*;
- отрицателни – *не*, *ни*, *нито*;
- пожелателни – *да* , *нека*, *дано*;
- усилващи – *даже*, *дори*;
- емотивни – *леле*, *ха*.

Частици (Q)

№	Х-ка	Стойност	Пример	Таг
1	Вид	отрицателни	не, ни	z
		призивни	ма, бе	g
		сравнителни	по, най	c
		словообразуващи	да, ще	v
		въпросителни	ли, дали	q
		пожелателни	да, дано	o
2	Форма	прости	а, не	s
		съставни	хайде де	c

Класове думи
Числителни имена. Видове

- числителни бройни (показват броя на съществата и предметите- *един*, *два*, *три*, *пет*)
- числителни редни(показват реда на съществата и предметите – *първи*, *втори*, *трети*)

Числителни имена

(М)==========

№	Х-ка	Стойност	Пример	Таг
1	Вид	бройни	един	c
		редни	втори	o
2	Род	мъжки	един	m
		женски	една	f
		среден	едно	n
3	Число	единствено	един	s
		множествено	едни	p
4	Падеж		-	-
5	Форма	арабска	1984	d
		римска	IX	r

Класове думи

Местоимения. Видове

- лични (*аз, ти, той, тя, то, ние, вие, те*);
- притежателни (*мой, твой, негов, ми, ти, му*);
- възвратни (*себе, себе си, свой, своя, свои*);
- показателни (*този, онзи, такъв, онакъв, инакъв, толкав*);
- въпросителни (*кой, какъв*);
- относителни (*който, какъвто, чийто*);
- неопределителни (*някой, някакъв, нечий*);
- отрицателни (*никой, никакъв*);
- обобщителни (*всеки, всякакъв, всичкия*).

Местоимение (P)

№	Х-ка	Стойност	Пример	Таг
1	Вид	лични	аз	p
		показателни	този	d
		неопределени	някой	i
		притежателни	мой	s
		въпросителни	кои	q
		относителни	който	r
		възвратни	себе	x
		отрицателни	никой	z
		обобщителни	всеки	g
2	Лице	първо	аз	1
		второ	ти	2
		трето	той	3

Класове думи
Междуметия

Представляват:

- имитация на естествени природни явления (*бъл-бук*);
- наподобяване на изкуствено сътворени от човека изделия (*скръц*);
- имитация на звукове издадени от животни (*меее*)

Междуметие (I)

№	Х-ка	Стойност	Пример	Таг
1	Вид			-
2	Форма	проста	ах	s
		съставна	боже мой!	c

Поставяне на тагове на частите на речта

- Процес, при който се определят частите на речта или други лексикални класове чрез тагове за всяка дума от определен сборника с тагове.
- Поставянето на тагове играе все по-важна роля в разпознаването на речта, в синтактичния разбор на естествения език и във възстановяването на информацията.

Поставяне на тагове

Алгоритми за поставяне на тагове

- Начални данни за алгоритъма - низ от думи и определен набор от тагове.
- Изходни данни - единственият, най-добър таг за всяка дума.

Например:

Те <лично местоимение> *имат* <глагол> *лакиран* <прилагателно> *под* <съществително>.

//многозначност при определянето на таг за конкретна дума

Поставяне на тагове на частите на речта

Поставяне на тагове
Алгоритми за поставяне на тагове

- Рещение на проблема с многозначността- повечето тагове, свързани с думите, не са с еднакви вероятности.
- Можем да използваме факта, свързан с честотата на таговете като прост алгоритъм, избирайки най-вероятния таг за всяка многозначнаслена дума. Gale, Church и Yarowsky (1992) предлагат този алгоритъм да се използва като основа за сравнение на всички останали алгоритми.

Поставяне на тагове
Алгоритми за поставяне на тагове

- Charniak, Hendrickson, Jacobson и Perkowitz (1993) показват, че версиите на този основен алгоритъм постига точност от 90-91% (точността е процента от думи, които получават верния таг, където “верния” означава “определен от човек”).

Алгоритъм за поставяне на тагове HMM

- Въведен през 80-те години на миналия век (Church, 1988; DeRose, 1988; Garside, 1987).
- За основа на този алгоритъм се използва поставянето на най-подходящия таг за отделна дума, изведена от контекста.
- Нека за начало видим как чрез този алгоритъм се поставя таг на единична дума. Първо задаваме основния таг, след това задаваме примери и най-накрая се обосноваваме.

Биграмен алгоритъм HMM

- Избираме тага t_i за дума w_i, което най-често се задава чрез предходния таг и текущата дума:

  $\mathbf{t_i = argmax\ P\ (t_j \mid t_{i-1}, w_i)}$

- Според някои предположения на Марков, ние преобразуваме тази формула до:

  $\mathbf{t_i = argmax\ P\ (t_j \mid t_{i-1})\ P(w_i \mid t_j)}$

Биграмен алгоритъм HMM

- Пример за обосновка:

  Котката е под масата.

  Те имат лакиран под.

- Биграмната версия на алгоритъма HMM прави предположение, че проблема за подходящия таг може да се реши чрез разглеждането на околните думи и тагове т.е. *лакиран под* и *под масата*

Алгоритъм HMM, приложен върху изречение

- За всяко изречение изчисляваме най-вероятната последователност от тагове T=t1,t2,…tn, която се задава от последователността от думи в изречение (W):

$$T=\underset{T \in \tau}{\operatorname{argmax}} P(T|W)$$

- Чрез закона на Bayes вероятността P(T|W) може да се изрази като:

$$P(T|W)=[P(T)P(W|T)]/P(W)$$

Алгоритъм HMM, приложен върху изречение

- Избераме последователността от тагове, които усъвършенстват вероятността:

$$P(T|W)=[P(T)P(W|T)]/P(W)$$

до:

$$P(T|W)=\underset{T \in \tau}{\operatorname{argmax}}\ P(T)P(W|T)\ /\ P(W)$$

Алгоритъм HMM, приложен върху изречение

- Пропускаме вероятността за определена дума P(W), понеже тя е същата като вероятността за всеки таг от изречението
- $T = \underset{T \in \tau}{\operatorname{argmax}}\ P(T)P(W|T) / P(W)$
- От основното правило за възможността имаме:
- $P(T)P(W|T) = \prod_{i-1}^{\pi} P(wi|w1t1 \ldots wi\text{-}1ti\text{-}1ti)P(ti|w1t1 \ldots wi\text{-}1ti\text{-}1)$

Триграмен модел за поставяне на тагове

- Правим предположения, че вероятността на дадена дума оказва влияние върху таг чрез формулата:

  $$P(w_i | w_1 t1 \ldots wi\text{-}1, ti\text{-}1 ti) = p(wi | ti)$$

- Правим предположението, че историята на тага може да бъде приближена от най-използваните два тага:

  $$P(wi | w1 t1 \ldots wi\text{-}1, ti\text{-}1 ti) = P(ti | ti\text{-}2 ti\text{-}1)$$

Триграмен модел за поставяне на тагове

- Избираме последователността от тагове, които са по-вероятни:

$$P(t1)P(t2|t1) = \prod_{i=3}^{n} P(ti|ti\text{-}2ti\text{-}1) / \prod_{i=1}^{n} P(wi|ti)$$

- Използваме максималната вероятност, за да изчислим тези вероятности:

$$P(ti|ti\text{-}2ti\text{-}1) = c(ti\text{-}2ti\text{-}1ti) / c(ti\text{-}2ti\text{-}1)$$

$$P(wi|ti) = c(wi,ti) / c(ti)$$

Поставяне на тагове, базирано на трансформационни правила

Brill(1995) предложил нов метод на поставяне на тагове - поставяне на тагове, базирано на трансформационни правила (TBL), като се ръководел от езиковите правила и от алгоритъма HMM.

TBL е също базиран на правила, свързани с това кой таг трябва да се зададе на дадена дума.

TBL е механична техника, в която съществуват правила, които се индуцират от фактите.

Поставяне на тагове, базирано на трансформационни правила

- Алгоритъмът TBL разполага с набор от правила.
- При първото поставяне на тагове се използва най-общото правило, засягащо таговете на повечето елементи.
- След това се избира по-тясно специализирано правило, което променя някои от таговете на елементите.
- После се прилага още по-специализирано правило, което засяга таговете на малък брой елементи.

Поставяне на тагове, базирано на трансформационни правила

Например:

Котката е под масата.

- При първото поставяне на тагове можем да определим следното:

  Котката/NN *е*/VB *под*/NN *масата*/NN.

- При следващото поставяне на тагове се взима под внимание думите като части на речта след *под* :

  Котката/NN *е*/VB *под*/PP *масата*/NN.

Поставяне на тагове, базирано на трансформационни правила

Основни етапи.

- Всяка дума се разглежда самостоятелно и се поставя подходящия таг.
- Алгоритъмът проверява всяка възможна трансформация и избира най-удовлетворяващия го таг.
- Поставят се тагове на данните според последното правило.

Тези три етапа се повтарят, докато се достигне до критерий за край, например не е направена никаква промяна при последната проверка.

Поставяне на тагове, базирано на трансформационни правила

Шаблони

- Абстрактни трансформации , чрез които се ограничава броя на преобразованията.
- Списъкът от шаблони е:
  - На предходната(следващата) дума се поставя таг ***z***.
  - На думата, която е две думи преди (след) дадената, се поставя таг ***z***.
  - На една от двете предходни (следващи) думи се поставя таг ***z***.

Поставяне на тагове, базирано на трансформационни правила

Шаблони

- На предходната дума се поставя таг ***z***, а на следващата – ***w***.
- На една от трите предходни (следващи) думи се поставя таг ***z***.
- На предходната (следващата) дума се поставя таг ***z***, а на следващата (предходната) – ***w***.

```
function TBL-Learn-Transforms (corpus) returns
transforms_queue
INITIALIZE_WITH_MOST_LIKELY_TAGS(corpus)
until end condition is met do
    templates< Generate-Potential-Relevant-Templates
    best_transform< Get_Best_Transform(corpus, templates)
    Apply_Transform(best_transform, corpus)
    Enqueue(best_transform_rule, transforms_queue)
End
Return (transforms_queue)
function Get_Best_Transform (corpus, templates) returns
transform
For each template in templates
    (instance, score)< Get_Best_Instance(corpus, template)
    if (Score>best_transform.Score) then
        best_transform< (instance, score)
end
return (best_transform)
```

Множествени тагове и съставни думи
Проблеми и техните решения

- Неопределени тагове - съществува многозначна при таговете за дадена дума и е невъзможно или доста трудно то да се елиминира.
- Множествени тагове – елиминират многозначността при таговете.. Такъв е случая с алгоритмите Penn Treebank и British National Corpus.

Множествени тагове и съставни думи
Проблеми и техните решения

- Общ неопределен таг включва прилагателно име заедно с глагол в минало свършено време заедно с минало причастие на глагола и прилагателно име заедно със съществително нарицателно име.

Множествени тагове и съставни думи
Проблеми и техните решения

- Съставните думи - съвкупността от тагове CLAWS, например, позволява предлози като *с изключение на* да се разглеждат като една дума чрез добавяне на пореден номер за всеки таг:

  с/SP1 *изключение*/SP2 *на*/SP3

Непознати думи
Опростен алгоритъм за представяне на нови думи

- Всяка непозната дума се представя недвусмислено чрез възможните тагове с еднаква вероятност.
- Алгоритъма трябва да разчита единствено на правилата за частите на речта, за да определи подходящия таг.

Непознати думи

Алгоритъм за представяне на нови думи

- Този алгоритъм е предложен от Baayen и Sproat (1996) и Dermatas и Kokkinakis (1995).
- Използва идеята за вероятното разпределение на таговете между непознатите думи, който е много близък до разпределянето на тагове между думи, които се срещат само веднъж в дадения текст.

Непознати думи

Алгоритъм за представяне на нови думи

- Hapax legomena (дума/форма, засвидетелствувана само веднъж) - думите, които се срещат само веднъж.
- Вероятността P(wi|ti) за една непозната дума се определя от средното разпределение на всички еднотипни думи в текста.

Непознати думи

Алгоритъм за представяне на нови думи

Използва спелуването на дадена дума.

Например думи, които завършват на *-и* е по-вероятно да са в множествено число, а думи, които завършват на *-ше* по принцип са глаголи в минало несвършено време.

Weischedel(1993) използвал четири вида специфични правописни правила и на базата на тях използва формулата:

P(wi|ti)=p(непозната_дума|ti)*p(главна_буква|ti)*p(окончания/сложни_думи|ti)

Клас-базирани N-грами

Клас-базирани N-грами

Клас-базираният N-грам е частен случай на N-грам, който използва резултатите от класовете части на речта, за да осигури добре определена оценка на вероятността на срещане на низовете думи. Основният клас-базиран N-грам определя условната вероятност на дума wi , която се базира на две вероятности: вероятността на класа, базирана на предходните класове и вероятността на отделна дума, взета от класа:

$$P(w_n | w_{1n-1}) = P(w_n | c_n) P(c_n | c_{1n-1})$$

Клас-базирани N-грами

Максималната оценка на вероятността на взета дума от класа е следната:

$$P(w|c)=C(w)/C(c)$$

$$P(ci|ci\text{-}1)=C(ci\text{-}1\ ci)/\Sigma cC(ci\text{-}1\ c)$$