**ПЛОВДИВСКИ УНИВЕРСИТЕТ „ПАИСИЙ ХИЛЕНДАРСКИ“**

**Факултет по математика и информатика**

**РЕФЕРАТ ПО КОМПЮТЪРНА ЛИНГВИСТИКА**

**ТЕМА НА РЕФЕРАТА**

*Вероятностен парсер. Учене на вероятности от PennTreeBank: проблеми и решения. Бинаризация. Анотация с категорията на родителя. Разбиване на символите. Хоризонтална марковизация. Вертикална марковизация. Лексикализация. Депендентно парсиране. Пренареждане на най-добрите кандидати (re-ranking). Вероятност на дърво и вероятност на парсиране.*

**Изготвил:**

**Александър Каранейчев**

**Специалност: Информатика**

**Факултетен номер: 2001261008**

**Пловдив**

**2024 г.**

## Учене на вероятности от PennTreeBank

**Какво е PennTreeBank?**

PennTreeBank е ръчно анотиран корпус от английски език, който се използва за обучение на модели за езикова обработка. Корпусът съдържа над 4.5 милиона думи, анотирани с части на речта, синтактични зависимости и семантични категории.

**Как се учат вероятности от PennTreeBank?**

Вероятностите от PennTreeBank се учат чрез статистически анализ на анотираните данни. Изчисляват се вероятностите за:   
***Части на речта; Синтактични зависимости; Семантични категории***

**Как се използват вероятностите от PennTreeBank?**

Вероятностите от PennTreeBank се използват за:   
***Подобряване на точността на моделите за езикова обработка; Разработване на нови модели за езикова обработка; Предимства на ученето на вероятности от PennTreeBank; Подобряване на точността на моделите; Разработване на нови модели***

**Недостатъци на ученето на вероятности от PennTreeBank:   
*PennTreeBank е ръчно анотиран; PennTreeBank е ограничен по размер***

**Заключение:**Ученето на вероятности от PennTreeBank е полезен метод за подобряване на точността на моделите за езикова обработка.

## Бинаризация

**Какво е бинаризация?**  
Бинаризацията е процес на преобразуване на данни в двоичен формат, т.е. в последователност от 0 и 1.

**Защо се използва бинаризация?**  
Бинаризацията се използва за:   
***Улесняване на съхранението, Улесняване на обработката, Повишаване на точността***

**Как се извършва бинаризация?**  
Съществуват различни методи за бинаризация, в зависимост от вида на данните:  
***Текстови данни; Числови данни; Изображения***

**Предимства на бинаризацията:  
*Улесняване на съхранението; Улесняване на обработката; Повишаване на точността***

**Недостатъци на бинаризацията:  
*Загуба на информация; Увеличаване на сложността***

**Заключение:**  
Бинаризацията е полезен метод за преобразуване на данни в двоичен формат, което може да улесни съхранението, обработката и повишаване на точността на данните.

## Анотация с категория на родителя

**Какво е анотация с категория на родителя?**  
Анотацията с категория на родителя е метод за добавяне на информация към синтактично дърво. При анотация с категория на родителя към всеки възел в дървото се добавя етикет, съответстващ на категорията на неговия родител.

**Защо анотацията с категория на родителя е важна?**  
Анотацията с категория на родителя е важна, тъй като:  
***Подобрява точността на парсиране; Улеснява обучението на модели***

**Пример за анотация с категория на родителя:**

**S**

**/ \**

**NP VP**

**/ \ / \**

**N Det V NP**

**the saw the telescope**

В този пример:

* S е етикетът на корена на дървото, който съответства на категорията "изречение".
* NP е етикетът на левия възел, който съответства на категорията "именна фраза".
* VP е етикетът на десния възел, който съответства на категорията "глаголна фраза".
* N е етикетът на левия възел под NP, който съответства на категорията "съществително име".
* Det е етикетът на десния възел под NP, който съответства на категорията "определителен член".
* V е етикетът на левия възел под VP, който съответства на категорията "глагол".
* NP е етикетът на десния възел под VP, който съответства на категорията "именна фраза".

**Заключение:**  
Анотацията с категория на родителя е проста, но ефективна техника, която може да подобри точността на парсиране и да улесни обучението на модели.

## Разбиване на символите

**Какво е разбиване на символите?**  
Разбиването на символите е процес на разделяне на терминални символи в синтактично дърво на по-малки единици.

**Защо разбиването на символите е важно?**  
Разбиването на символите е важно, тъй като:  
***Подобрява точността на парсиране, Улеснява обучението на модели***

**Пример за разбиване на символите:**

**S**

**/ \**

**NP VP**

**/ \ / \**

**N Det V NP**

**the saw the telescope**

В този пример:

* "the" е разбито на "the" (определителен член)
* "saw" е разбито на "saw" (глагол)
* "telescope" е разбито на "tele" (представка), "scope" (съществително име)

**Варианти на разбиване на символите:**  
Съществуват различни варианти на разбиване на символите, които се различават по начина, по който се разделят терминалните символи.

**Заключение:**  
Разбиването на символите е проста, но ефективна техника, която може да подобри точността на парсиране и да улесни обучението на модели.

## Хоризонтална марковизация

**Какво е хоризонтална марковизация?**  
Хоризонталната марковизация е подход за моделиране на синтактични зависимости в рамките на вероятностните парсери. При хоризонталната марковизация се предполага, че вероятността на даден възел в синтактичното дърво зависи от вероятността на неговия родител и от вероятностите на неговите леви братя.

**Пример за хоризонтална марковизация:**

**S**

**/ \**

**NP VP**

**/ \ / \**

**N Det V NP**

**the saw the telescope**

В този пример:

* Вероятността на NP зависи от вероятността на S и от вероятността на Det.
* Вероятността на VP зависи от вероятността на S и от вероятността на NP.
* Вероятността на N зависи от вероятността на NP.
* Вероятността на Det зависи от вероятността на NP.
* Вероятността на V зависи от вероятността на VP.
* Вероятността на NP зависи от вероятността на VP.

**Предимства на хоризонталната марковизация:***Проста и ефективна; Лесна за обучение; Добре приложима за задачи, при които е важно да се моделират зависимости между думи, които са близо една до друга в изречението.*

**Недостатъци на хоризонталната марковизация:  
*Не*** *може да моделира зависимости между думи, които са далеч една от друга в изречението; Може да не е подходяща за задачи, при които е важно да се моделират дълги зависимости.*

**Заключение:**  
Хоризонталната марковизация е прост и ефективен подход за моделиране на синтактични зависимости. Тя е лесна за обучение и е добре приложима за задачи, при които е важно да се моделират зависимости между думи, които са близо една до друга в изречението.

## Вертикална марковизация:

**Какво е вертикална марковизация?**  
Вертикалната марковизация е подход за моделиране на синтактични зависимости в рамките на вероятностните парсери. При вертикалната марковизация се предполага, че вероятността на даден възел в синтактичното дърво зависи от вероятността на неговия родител и от вероятностите на неговите деца.

**Пример за вертикална марковизация:**

**S**

**/ \**

**NP VP**

**/ \ / \**

**N Det V NP**

**the saw the telescope**

В този пример:

* Вероятността на NP зависи от вероятността на S и от вероятностите на N и Det.
* Вероятността на VP зависи от вероятността на S и от вероятностите на V и NP.
* Вероятността на N зависи от вероятността на NP.
* Вероятността на Det зависи от вероятността на NP.
* Вероятността на V зависи от вероятността на VP.
* Вероятността на NP зависи от вероятността на VP.

**Предимства на вертикалната марковизация:**  
*Може да моделира зависимости между думи, които са далеч една от друга в изречението; Тя е добре приложима за задачи, при които е важно да се моделират дълги зависимости.*

**Недостатъци на вертикалната марковизация:***По-сложна от хоризонталната марковизация; По-трудна за обучение.*

**Заключение:**  
Вертикалната марковизация е по-сложен, но по-мощен подход за моделиране на синтактични зависимости. Тя може да моделира зависимости между думи, които са далеч една от друга в изречението, и е добре приложима за задачи, при които е важно да се моделират дълги зависимости.

## Лексикализация

**Какво е лексикализация?**  
Лексикализацията е процес на включване на информация за лексиката в модели за парсиране.

**Защо лексикализацията е важна?**  
Лексикализацията е важна, тъй като:  
***Подобрява точността на парсиране; Улеснява обучението на модели***

**Пример за лексикализация:**

**S**

**/ \**

**NP VP**

**/ \ / \**

**N Det V NP**

**the saw the telescope**

В този пример:

* "the" е лексикализирано като определителен член
* "saw" е лексикализирано като глагол в минало време
* "telescope" е лексикализирано като съществително име

**Варианти на лексикализация:**  
Съществуват различни варианти на лексикализация, които се различават по начина, по който се включва информация за лексиката в моделите за парсиране.

**Заключение:**  
Лексикализацията е проста, но ефективна техника, която може да подобри точността на парсиране и да улесни обучението на модели.

## Депендентно парсиране

**Какво е депендентно парсиране?**  
Депендентното парсиране е подход за синтактичен анализ, който се фокусира върху отношенията на зависимост между думите в изречението. При депендентното парсиране се изгражда дърво на зависимостите, в което всеки възел е дума, а дъгите между възлите представляват зависимости.

**Пример за депендентно парсиране:**

**S**

**/ \**

**NP VP**

**/ \ / \**

**N Det V NP**

**the saw the telescope**

В този пример:

* "the" е зависимо от "saw"
* "saw" е главен възел
* "telescope" е зависимо от "saw"

**Предимства на депендентното парсиране:***По-просто от конституентното парсиране; По-лесно за обучение; По-ефективно за задачи, при които е важно да се извличат зависимости между думите в изречението.*

**Недостатъци на депендентното парсиране:***Не е толкова точно, колкото конституентното парсиране; Не е толкова добре приложимо за задачи, при които е важно да се извлича информация за структурата на фразата.*

**Заключение:**  
Депендентното парсиране е по-прост и по-ефективен подход за синтактичен анализ, в сравнение с конституентното парсиране. То е добре приложимо за задачи, при които е важно да се извличат зависимости между думите в изречението.

## Пренареждане на най-добрите кандидати (re-ranking)

**Какво е пренареждане на най-добрите кандидати (re-ranking)?**  
Пренареждането на най-добрите кандидати е техника, използвана в машинния превод за подобряване на качеството на превода. При пренареждането на най-добрите кандидати се генерират множество преводи на дадено изречение, след което се подреждат по отношение на тяхното качество.

**Как работи пренареждането на най-добрите кандидати?**  
Пренареждането на най-добрите кандидати се състои от два основни етапа:  
***Генериране на множество преводи; Подреждане на преводите***

**Предимства на пренареждането на най-добрите кандидати:**

*Може да подобри качеството на превода; Може да се използва за персонализиране на превода; Може да се използва за ускоряване на превода.*

**Недостатъци на пренареждането на най-добрите кандидати:***Може да бъде computationally expensive; Може да не е толкова ефективно за езици с ограничени ресурси.*

**Заключение:**  
Пренареждането на най-добрите кандидати е ефективна техника за подобряване на качеството на машинния превод.

## Вероятност на дърво и вероятност на парсиране

**Вероятност на дърво:**  
Вероятността на дърво е мярка за вероятността от дадено синтактично дърво да е правилното парсиране на дадено изречение. Тя се изчислява като произведение на вероятностите на всички правила, използвани за изграждане на дървото.

**Пример за изчисляване на вероятност на дърво:**

**S**

**/ \**

**NP VP**

**/ \ / \**

**N Det V NP**

**the saw the telescope**

Вероятността на това дърво е:

**P(S) \* P(NP | S) \* P(Det | NP) \* P(N | NP) \* P(VP | S) \* P(V | VP) \* P(NP | VP) \* P(telescope | NP)**

**Вероятност на парсиране:**  
Вероятността на парсиране е мярка за вероятността от дадено изречение да бъде правилно парсирано. Тя се изчислява като сума от вероятностите на всички възможни синтактични дървета за това изречение.

**Пример за изчисляване на вероятност на парсиране:**

**P(the saw the telescope) = P(S1) + P(S2) + ...**

където S1, S2, ... са всички възможни синтактични дървета за изречението "the saw the telescope".

**Връзка между вероятност на дърво и вероятност на парсиране:**  
Вероятността на парсиране е сума от вероятностите на всички възможни синтактични дървета за дадено изречение. Вероятността на дърво е мярка за вероятността от дадено синтактично дърво да е правилното парсиране на дадено изречение.

**Заключение:**  
Вероятността на дърво и вероятността на парсиране са важни концепции в машинното парсиране. Те се използват за оценка на качеството на парсиране и за избор на най-доброто парсиране за дадено изречение.

## Източници:

**За "Хоризонтална марковизация":**

* Wikipedia: Horizontal Markovization
* Dependency Parsing with Horizontal Markovization
* A Comparison of Horizontal and Vertical Markovization for Dependency Parsing

**За "Вертикална марковизация":**

* Wikipedia: Vertical Markovization
* Dependency Parsing with Vertical Markovization
* A Comparison of Horizontal and Vertical Markovization for Dependency Parsing

**За "Лексикализация":**

* Wikipedia: Lexicalization:
* Lexicalization in Dependency Parsing
* The Role of Lexicalization in Dependency Parsing

**За "Депендентно парсиране":**

* Wikipedia: Dependency Parsing
* Dependency Parsing: A Survey
* A Gentle Introduction to Dependency Parsing

**За "Пренареждане на най-добрите кандидати":**

* Wikipedia: Re-ranking
* Re-ranking for Machine Translation
* Neural Re-ranking for Machine Translation

**За "Вероятност на дърво и вероятност на парсиране":**

* Wikipedia: Tree Probability
* Wikipedia: Parsing Probability
* Parsing with Probabilistic Context-Free Grammars

**Допълнителни източници:**

* Speech and Language Processing: <https://web.stanford.edu/~jurafsky/slp3/> by Daniel Jurafsky and James H. Martin
* Foundations of Statistical Natural Language Processing: <https://web.stanford.edu/~jurafsky/slp3/> by Christopher Manning, Prabhakar Raghavan, and Hinrich Schütze
* The Google AI Blog: <https://ai.googleblog.com/>
* The Stanford NLP Group: <https://nlp.stanford.edu/>