Context-Free Grammars and Constituency Parsing

Bağlamdan Bağımsız Dilbilgisi ve Çözümleme

2024-2025 Güz

Bağlamdan Bağımsız Dilbilgisi / Context Free Grammar (CFG)

- Bağlamdan bağımsız dilbilgisi (CFG), doğal dilin ve de programlama dillerinin sözdiziminin pek çok biçimsel modelinin yapı taşıdır.
- Sözdizimsel ayrıştırma (çözümleme), herhangi bir cümleye sözdizimsel bir yapının atanması işidir.
- Çözümleme ağaçları dilbilgisinin denetlenmesi amacıyla kullanılabilir.
  - Çözümlenemeyen /ayrıştırılamayan bir cümlede dilbilgisi hataları olur, okunması zorlaşır.
- Çözümleme ağaçları, biçimsel formda anlambilimsel (semantik) analiz için bir ara çözümleme aşamasıdır.
- Çözümleme ve cümleye atanan dilbilgisi yapısı bir metin analizi aracıdır.
  - Cümlenin öğelerindeki ilişkilerin modellemesinin yapışdığı metinsel veri bilimi uygulamalarıdır.

Bağlamdan Bağımsız Dilbilgisi

- Bağlamdan bağımsız dilbilgisi, cümle yapısı dilbilgisi (phrase-structure grammar), olarak da adlandırılır.
- Biçimselliği Backus-Naur Forma (BNF) eşdeğerdir.
- Bir dilbilgisini bileşenlerinin yapısına dayandırma fikri psikolog Wilhelm Wundt'a (1900) uzanır.
- Dilbilgisi kavramı, Chomsky'ye (1956) ve bağımsız olarak Backus'a (1959) kadar biçimselleştirilmemiştir.

İSİM ÖBEĞİ
NOUN PHRASE /NP

$$NP \rightarrow Det\ Nominal$$
$$NP \rightarrow ProperNoun$$
$$Nominal \rightarrow Noun \mid Nominal\ Noun$$
$$Det \rightarrow a$$
$$Det \rightarrow the$$
$$Noun \rightarrow flight$$

- CFG, her biri dilin sembollerinin gruplandırılıp sıralanışını ifade eden bir dizi kuraldan oluşur.
  - Ayrıca bu kuralların üretilip dönüştürüldüğü bir kelime ve sembol sözlüğünden meydana gelmektedir.
- Kuralların sıralanışı, bir NP'nin (isim öbeğinin) bir belirteç (Det) ve ardından bir isim (adsal) grubundan (Nominal) veya bir özel isimden oluşabileceğini ifade eder.
- Nominal ise bir veya daha fazla isimden (Noun) oluşabilir.
- Bağlamdan bağımsız kurallar hiyerarşik şekilde yerleştirilebilir.

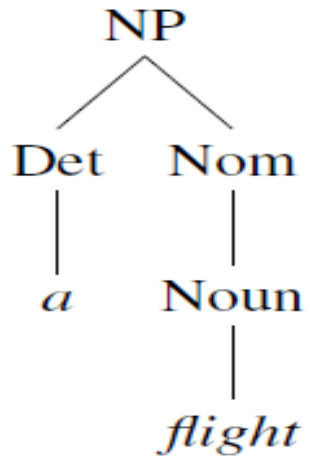
Bağlamdan Bağımsız Dilbilgisi

- Bir CFG'de kullanılan semboller iki sınıfa ayrılır.
  - Dildeki kelimelere ("the", "school") karşılık gelen sembollere *terminal* sembolleri denir.
    - ✓Sözlük (lexicon), terminal sembollerini oluşturan kurallar kümesidir.
  - Terminaller üzerindeki soyutlamaları ifade eden sembollere *terminal olmayan* semboller denir.
- Her CFG kuralında, (→)okun sağındaki öğe, bir veya daha fazla *terminal ve terminal olmayan* sıralanıştan oluşur; okun solunda ise genellemeyi ifade eden tek bir *terminal olmayan* sembol bulunur.
- Sözlük içindeki her kelimeyle ilişkilendirilen bir *terminal olmayan* sembol, bir sözcüksel (lexical) kategori veya konuşmanın bir parçasıdır.

Bağlamdan Bağımsız Dilbilgisi

- CFG iki farklı amaç için kullanılabilir:
  - Yeni cümleler oluşturulması (üretilmesi) için bir araçtır.
  - Herhangi bir cümleye bir yapı atamak, yani çözümlemek için bir araçtır.
- CFG bir üreteci (generator) olarak alındığında → okunun solundaki sembol, sağındaki sembol dizisiyle yeniden yazılır (yer değiştirilir) şeklinde okuyabiliriz
  - NP sembolünden başlayarak, *NP nonterminalini* yeniden yazmak üzere ilk kural kullanılabilir (*Det Nonimal*). Det terminal kategorisinden «»a» alınır. Daha sonra *Nominal nonterminali* Noun olarak yeniden yazılır. En sonunda, «a flight» POS olarak yeniden yazılır.
    - SONUÇ OLARAK: «a flight» dizgisi terminal olmayan sembollerden (NP) türetilmiştir.

İsim Öbeği Çözümleme Ağacı / Noun Phrase Parsing



- Bu çözümleme ağacı, NP düğümünün ağaçtaki tüm düğümlerine (Det, Nom, Noun, a, flight) aittir.
  - NP nonterminali , Det ve Nom düğümlerine ayrışır.
- CFG ile tanımlanan bir formal dil, belirlenmiş başlangıç sembolünden türetilebilen dizgeler kümesidir.
- Her dilbilgisinde, S olarak adlandırılan bir başlangıç sembolü olmalıdır.
- CFG, genellikle cümleleri tanımlamak için kullanıldığından, S «cümle düğümü /sentence node», yani başlangıç düğümü olarak tanımlanır.

Eylem Öbeği Kurallarının Oluşturulması

Herhangi bir cümle bir isim öbeği (NP – noun phrase) ve onu izleyen bir eylem öbeğinden (VB - verb phrase) oluşur:

S → NP VP I prefer a morning flight

Herhangi bir eylem öbeği, bir eylemin ardından gelen çeşitli öbeklerden oluşur.

VP → Verb NP prefer a morning flight

ya da

VP → Verb NP PP leave Boston in the morning

ya da

VP → Verb PP leaving on Thursday

PP → Preposition NP from Los Angeles

Sembollerden oluşan bir Sözlük (Lexicon)

Terminal olmayan bir sembolün alternatif olası açılımları | imi ile, veya sembolü ile gösterilir.

Noun → flights | flight | breeze | trip | morning

Verb → is | prefer | like | need | want | fly | do

Adjective → cheapest | non-stop | first | latest | other | direct

Pronoun → me | I | you | it

Proper-Noun → Alaska | Baltimore | Los Angeles | Chicago | United | American

Determiner → the | a | an | this | these | that

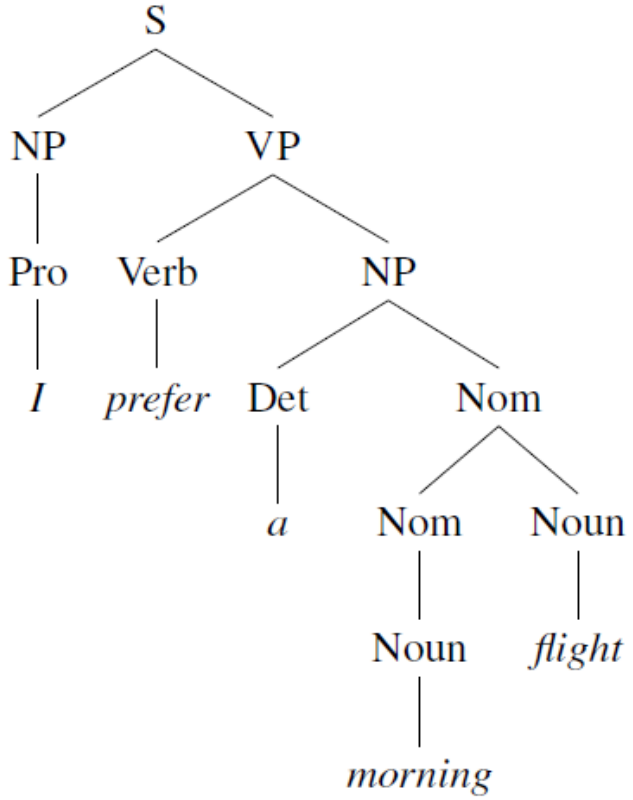
Preposition → from | to | on | near | in

Conjunction → and | or | but

CFG Kuralları ve Örnekleri

Dilbilgisi Kuralları	**Örnekleri**
S → NP VP	I + want a morning flight
NP → Pronoun	I
| Proper-Noun	Los Angeles
| Det Nominal	a + flight
Nominal → Nominal Noun	morning + flight
| Noun	flights
VP → Verb	do
| Verb NP	want + a flight
| Verb NP PP	leave + Boston + in the morning
| Verb PP	leaving + on Thursday
PP → Preposition NP	from + Los Angeles

CFG'ye ait Çözümleme Ağacı



Çözümleme ağacı[ ] (köşeli parantezlerle) ifade edilebilir.

[S [NP [Pro I]] [VP [V prefer] [NP [Det a] [Nom [Nom [N morning]][Nom [N flight]]]]]]

Düzgün Diller

- Bir CFG ye ait tüm dizgilerin oluşturduğu küme düzgün bir dil tanımlar (L_0).
- Formal (biçimsel) bir G dilbilgisi tarafından türetilemeyen cümleler, o dilbilgisi tarafından oluşturulan L diline ait değildir.
- Belirli bir cümlenin herhangi bir doğal dilin parçası olup olmadığının belirlenmesi genellikle bağlama bağlı olmasıdır.
- Dilbilimde, doğal dilleri modellemek için biçimsel (formal) dillerin kullanılmasına üretici dilbilgisi *(generative grammar)* denir.
  - Herhangi bir L dili, dilbilgisi tarafından "üretilen" tüm olası cümleler kümesi tarafından tanımlanır.

Bağlamdan Bağımsız Dilbilgisinin Tanımı

- G dilbilgisi 4 parametreli olarak (4-tuple) tanımlanır. Bunlar:

  N, terminal olmayan sembollerin sonlu kümesi

  $\sum$ terminal sembollerinin sonlu kümesi (N ile ortak elemanı yoktur)

  P kuralların sonlu dizilişidir ve $A \rightarrow \beta$ formundadır.

  A bir nonterminal semboldür,

  β tüm sembollerin herhangi bir dizilişi olan dizgidir.

  $(\sum \cup N)^*$ olarak gösterilir.

  S başlangıç sembolüdür ve bir nonterminaldir.

Bağlamdan Bağımsız Dilbilgisinin Tanımı

- A, B ve S gibi büyük harfler terminal olmayan sembolleri gösterir.
- S başlangıç sembolüdür ve nonterminaldir.
- α, β, γ gibi küçük küçük harfleri $(\sum \cup N)^*$ dan türetilmiş dizgileri gösterir.
- u, v, w gibi küçük Latin harfleri terminal sembollerinin bir dizilişi olan dizgileri gösterir.
- Bir G dilbilgisi tarafından oluşturulan L(G) dili, S başlangıç non terminalinden başlayıp adım adım türetmelerle verilen giriş cümlesinin elde edilmesi dir.

$$L(G) = \{ w \mid w \in \Sigma^* \quad ve \quad S \Rightarrow w^* \}$$

Treebank (Ağaç Bankası)

- Her cümlenin bir çözümleme ağacıyla açıklandığı gövdeye ağaç bankası denir.
- Sözdizimsel bildirimlerin dilbilimsel incelemelerinde ve çözümlemelerinde çok kullanılır.
- Ağaç bankalarının (treebanks) oluşturulması için her cümle üzerinde bir çözümleyici (parser) çalıştırılır.
  - Daha sonra ortaya çıkan ayrıştırmanın (çözümlemenin) elle düzeltilmesinden oluşturulur.

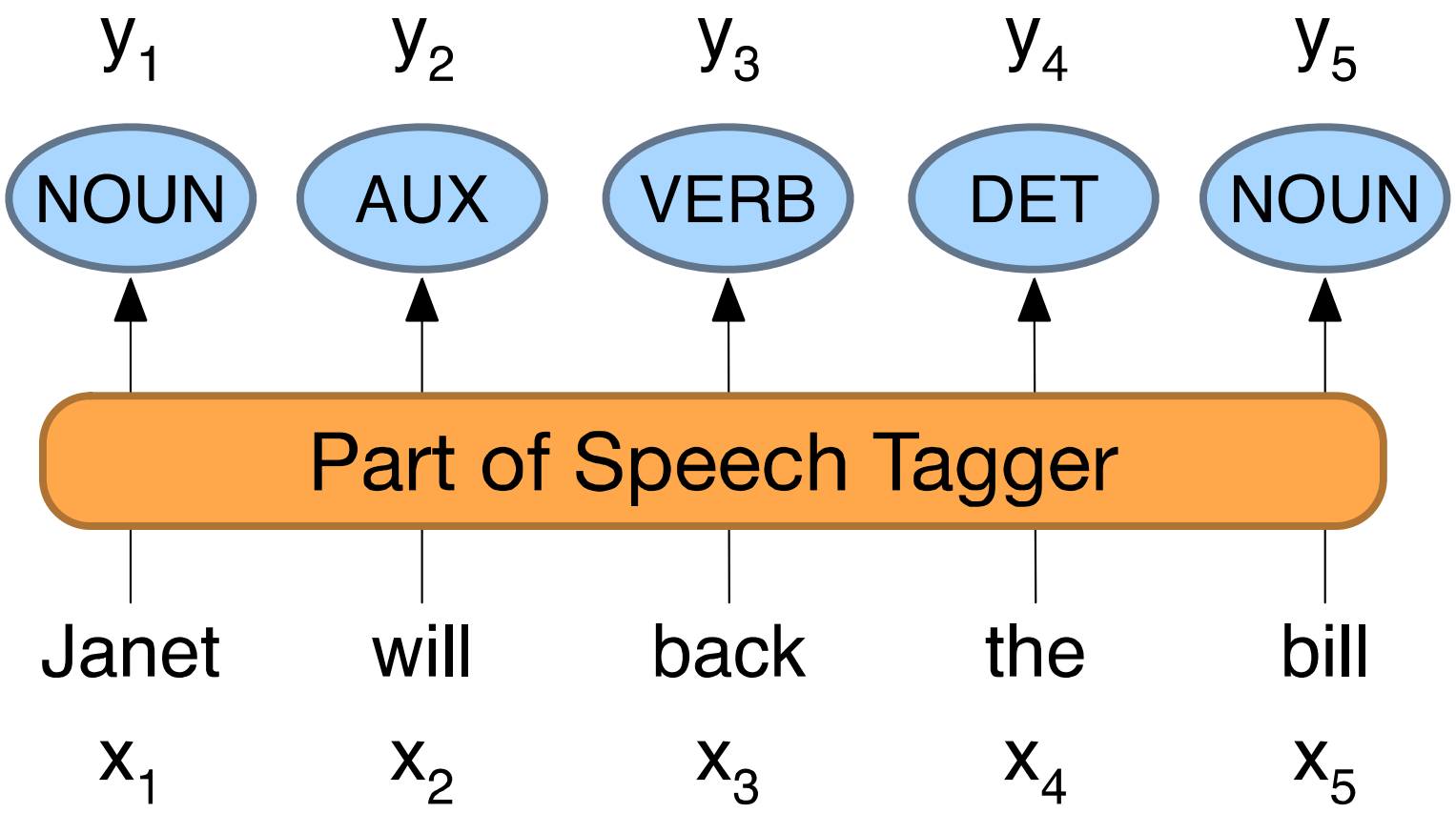
https://en.wikipedia.org/wiki/Treebank

Part-of-Speech Etiketleme / Tagging

- Kelime türü etiketleme, bir metindeki her kelimeye bir kelime türü atanması sürecidir.
- Etiketleme, bir anlam ayrımı görevi gerçekleştirir.
  - Kelimelerin çoğu belirsizdir (ambigious).
  - Her kelime olası birden fazla kelime türü içerebilir.
- Amaç, ilgili sözcük için doğru etiketi bulmaktır.
- Örneğin, kitap sözcüğü bir fiil (book that flight) veya bir isim (hand me that book) olarak kullanılabilir.

Part-of-Speech Etiketleme / Tagging

$x_1,\ldots,x_n$ olarak sözcüklerin herhangi bir sıralanışı iken $y_1,\ldots,y_n$ şeklinde POS etiketlerine dönüşür.



"Universal Dependencies" Tagset

Nivre et al. 2016

	Tag	Description	Example
Open Class	**ADJ**	Adjective: noun modifiers describing properties	*red*, *young*, *awesome*
	ADV	Adverb: verb modifiers of time, place, manner	*very*, *slowly*, *home*, *yesterday*
	NOUN	words for persons, places, things, etc.	*algorithm*, *cat*, *mango*, *beauty*
	VERB	words for actions and processes	*draw*, *provide*, *go*
	PROPN	Proper noun: name of a person, organization, place, etc..	*Regina*, *IBM*, *Colorado*
	INTJ	Interjection: exclamation, greeting, yes/no response, etc.	*oh*, *um*, *yes*, *hello*
Closed Class Words	**ADP**	Adposition (Preposition/Postposition): marks a noun's spacial, temporal, or other relation	*in, on, by under*
	AUX	Auxiliary: helping verb marking tense, aspect, mood, etc.,	*can, may, should, are*
	CCONJ	Coordinating Conjunction: joins two phrases/clauses	*and*, *or*, *but*
	DET	Determiner: marks noun phrase properties	*a, an, the, this*
	NUM	Numeral	*one, two, first, second*
	PART	Particle: a preposition-like form used together with a verb	*up, down, on, off, in, out, at, by*
	PRON	Pronoun: a shorthand for referring to an entity or event	*she, who, I, others*
	SCONJ	Subordinating Conjunction: joins a main clause with a subordinate clause such as a sentential complement	*that*, *which*
Other	**PUNCT**	Punctuation	; , ()
	SYM	Symbols like $ or emoji	$, %
	X	Other	asdf, qwfg

Etiketlenmiş Cümleler

There/PRO were/VERB 70/NUM children/NOUN there /ADV . /PUNC

Preliminary/ADJ findings/NOUN were/AUX reported/VERB in/ADP today/NOUN ’s/PART New/PROPN England/PROPN Journal/PROPN of/ADP Medicine/PROPN

Segmentasyon / parçalanma (noktalama işaretleri gibi) ile ilgili karar alınmalıdır

Aynı kelime farklı cümlelerde farklı etiketlere sahip olabilir.

POS Etiketleme Algoritmaları

- Denetimli makine öğrenme algoritmaları / Supervised Machine Learning Algorithm:
  - Hidden Markov Models
  - Conditional Random Fields (CRF)/ Maximum Entropy Markov Models (MEMM)
  - Neural sequence models (RNNs or Transformers)
  - Large Language Models (like BERT), finetuned (ince ayarlama)
- Elle etiketlenmiş eğitim setlerinin çoğunluğu İngilizcedir (%97)
- Bilgi (information) kaynakları çoğunlunlukla aşağıdaki süreçlerden geçerek elde edilir:
  - İnsan tarafından (human created) oluşturulmuş özelikler: HMMs ve CRFs
  - Öğrenmenin gerçekleştirilmesi: Sinirsel dil modelleri

POS Etiketleme

- POS etiketleme aşağıdaki 3 yoldan biri ile yapılabilir:
  - Hidden Markov Modelleri veya Conditional Random Fields (Koşullu Rastgele Alanlar) gibi klasik denetlenen makine öğrenme algoritmaları
  - Sıfırdan eğitilmiş sinir dizisi modelleri
  - İnce ayarlanmış büyük dil modelleri olan sinir modelleri.
- Yeterli eğitim verisiyle, yaklaşık olarak tümünde eşit performans elde edilir.
- Hemen hemen tümü benzeri bilgi kaynaklarını kullanır
  - HMM'ler ve CRF'ler, insan tarafından oluşturulan özellikler sinir modelleri tarafından bu özelliklerin kullanılması ile öğrenmeyi gerçekleştirmek üzere üretilir.

Named Entities / Özel İsimler

PER (Person): “Marie Curie”
LOC (Location): “New York City”
ORG (Organization): “Stanford University”
GPE (Geo-Political Entity): "Boulder, Colorado"

NER / Name Entity Recognition output

Citing high fuel prices, [ORG **United Airlines**] said [TIME **Friday**] it has increased fares by [MONEY **$6**] per round trip on flights to some cities also served by lower-cost carriers. [ORG **American Airlines**], a unit of [ORG **AMR Corp.**], immediately matched the move, spokesman [PER **Tim Wagner**] said. [ORG **United**], a unit of [ORG **UAL Corp.**], said the increase took effect [TIME **Thursday**] and applies to most routes where it competes against discount carriers, such as [LOC **Chicago**] to [LOC **Dallas**] and [LOC **Denver**] to [LOC **San Francisco**].

BIO Tagging

B: token that *begins* a span

I: tokens *inside* a span

O: tokens outside of any span

# of tags (where **n** is #entity types):

1 O tag,

n B tags,

n I tags

total of *2n+1*

Words	BIO Label
Jane	B-PER
Villanueva	I-PER
of	O
United	B-ORG
Airlines	I-ORG
Holding	I-ORG
discussed	O
the	O
Chicago	B-LOC
route	O
.	O

Türkçe Dil Deposu /TDD

- Türkiye Açık Kaynak Platformu'nun temel projelerinden biridir.
- Türkçe metinlerin işlenmesi için
  - gerekli veri kümelerinin hazırlanması,
  - bu veri kümelerinin dağıtım altyapısının oluşturulması,
  - yüksek performanslı kütüphanelerin oluşturulması,
  - bu kütüphanelere dayanan kullanıcı dostu ve çevrim içi araçların sunulması amaçlanır.

https://tdd.ai/?language=tr