

Doğal Dil İşleme(Natural Language Processing)

Sadi Evren SEKER

^aIstanbul Medeniyet University, Department of Business

Özet

Doğal dil işleme konusuna giriş içeriğinde olan bir yazı olmuştur. Doğal dil işleme ingilizce literatürde *natural language processing* ya da NLP olarak kısaltılmış hali geçmektedir. Doğal dil işleme süreçleri, doğal dil işlemeye neden gereksinim duyulmakta, karşılaşılan sorunlar nelerdir ve son kullanılan yerlere değinilerek giriş anlamında bir anlatım olmuştur.

Anahtar Kelimeler: Doğal dil işleme , Makine Çevrimi ,Kelime işleme ,internet arama motoru , Metin işleme , Finans ,Zamansal Mantıklar, Argüman Ağları

Summary

This article aims to make a brief introduction to natural language processing concept and terms. Some problem types and some application areas are introduced in the article.

Keywords: Natural Language Processing , Machine Translation, Word Processing ,NL/DB interface, AI Bots (Siri), Text Processing, Text Summarization, Argument Aggregation

1. Doğal Dil İşlemeye Genel Bakış

Diller bilindiği üzere ikiye ayrılır; makine dili(machine language) ve insanlar tarafından kullanılan doğal dildir(natural language). Bilgisayarların insanların dillerini anlaması , onlarla iletişime geçmeleri için doğal dil işleme bilimi kullanmaları gerekmektedir. Kısacası, bilgisayarların doğal dilleri işleme sürecidir. Bunu insan-bilgisayar etkileşimi (human computer interaction-HCI) (Preece, 1994.) biliminin altında da görülebilir veya hesaplamalı dil bilimi(computational linguistics-CL) olarak da görülebilir. Burada dilin hesaplanması işlem görmesi durumudur.

Veri bilimiyle doğal dil işleme arasındaki ilişki düşünüldüğünde, aslında bu konuyu metin işleme(Text Processing) ve metin madenciliğinin(text mining) (Seker, Metin Madenciliği(Text Mining), 2015) altında düşünmek gerekir. Çünkü veri biliminin genelde ilgilendiği kısım buralardır. Bunun sebebi , eğer veri kaynakları metin şeklindeyse bunların ilk önce analiz edilmesi ve işlenebilir hale getirilmesinde doğal dil işleme önemli bir rol oynamaktadır.

Doğal dil işleme girdilere göre ikiye ayrılır. Yazılı metinlere(text processing) göre ve ses (speech processing) üzerine yapılanlardır. Doğal dilin genellikle iki kaynağı bulunmaktadır. Birinin konuşması durumunda ona sesli yanıt verilmesi veya yazılı bir metnin analiz edilmesi olarak düşünülebilir. Doğal dil işleme çalışmalarını şu şekilde olmaktadır; ses üzerinden çok fazla çalışma yapılamamakla birlikte daha çok sesli anlatımların yazılı hale getirilip ve daha sonra yazılı hale getirilen metinler işleme dahil edilmeleriyle oluşmaktadır.

Doğal dil işlemenin çalışma seviyelerine bakıldığında dört ana madde görülmektedir. Kelime bilimi(Morphological-Lexical), söz dizimsel (syntactic), anlamsal(semantic) ve söylevdir(pragmatic-discourse). Doğal dil işleme aşamaları bu dört ana madde üzerinden anlatılmaya çalışılacaktır. Kelime bilimi(lexical), kelimelerin anlamlarının anlaşılmasıyla ilgilenir. Bir kelimenin kökünün hangi ekler alarak hangi anlama dönüştüğünü inceler ve bu sayede kelimenin ne olduğu anlaşılır. Söz dizimsel (syntactic), cümledeki kelimelerin dizilim şeklidir. Bir cümlede kelimelerin nasıl dizildiğiyle ilgilenir. Anlamsal(semantic), kullanılan cümlelerin anlamlarını inceler. Çünkü doğal dilin bilgisayar tarafından doğru işlenebilmesi için ilk, bu cümlelerin doğru bir şekilde anlaşılmasıdır. Söylev(pragmatics-discourse), bir konuşma sırasında kullanılan kelimeler ve anlamlarıyla ilgilenir. İlk önce bilgisayar tarafından konuşma anlaşılmalı ve nelerden bahsedildiği anlaşılan cümlelere doğru kelimeler kullanılarak cevap verilmesi durumudur.

Doğal dil işlemenin bir diğer kısmı üretim(Generation) kısmıdır. Burada bir dilin anlaşılması için gerekli zamanı yine bu dili üretmek için de kullanılmalıdır. Örneğin bilgisayarın ürettiği bir türkçe kelime olarak düşünülebilir.

2. Doğal Dil İşlemede Karşılaşılan Zorluklar

2.1. Belirsizlik-Muğlaklık Durumu (Ambiguity)

2.1.1. Sözcük Seviyesi(Lexical)

Diller ailesinde bitişik ,sondan eklemeli(Agglutinative) olarak bir dil ailesi bulunmaktadır. Yani bir kelimenin sondan eklemeli olarak ,birden fazla ek almasıyla bir çok anlama gelebilir. Bu sebepten dolayı bu kelimelerin anlamını bulmak zorlaşmaktadır. Örneğin Türkçe dili de sondan eklemeli bir dildir.

2.1.2. Sözdizimsel(Syntactic)

Cümlelerin içinde kelimelerin yer değiştirilmesiyle oluşturulan cümleler bulunmaktadır. Bu kelimelerin cümlede bulundukları yere göre anlamı değişebilmektedir. Kelimenin vurgusu değişebilir. Bu durumun bilgisayar tarafından anlaşılması gerekmektedir.

2.1.3. Kısmi Bilgi Yorumlanması(Interpretation of Partial Info)

Bir tercüme ,çeviri yapılması durumunda bir cümlede bulunan örneğin bir zamirin hangi kelimeyi karşıladığı bulunması durumudur. İngilizce zamirlerden “he,she,it” durumu gibi.

2.1.4. Bağlam(Context)

Bilgisayara girilen bir cümlelerin hangi alandan veya konudan bahsedildiği bilgisayar tarafında anlaşılmalıdır. Örnek olarak ‘Türkiye de ege bölgesinde dağlar denize diktir. ’cümlesi bilgisayara girildiğinde bunun coğrafi bir bilgi olduğu anlaşılmalı ve içerisindeki kelimelerin coğrafi terimler olabileceği varsayabilmelidir. Birşeyi ifade eden çok şey, çok şeyi ifade eden bir şey vardır cümlesi bu durumu açıklayan çok güzel bir ifadedir. Bir kişi bir cümle ile karşısındaki kişiye bir çok şey ifade edebilir ya da bir kelimeyle her şeyi açıklayabilir. İşte bu durumda bilgisayarın bu cümleyi nasıl anladığı önemli bir sorun olabilir ya da bir kelimeyi ifade eden bir çok kelime olabilir, bilgisayarın en doğru karşılığı bulması da burada yine önemli bir sorundur.

3. Dil Bilgisi Bilimi(Knowledge)

3.1. Ses bilimi(Phonology-Phonemes)

Kelimelerdeki vurgu durumudur. Her kelimenin, her harfin bir söyleniş şekli bulunmaktadır. Sesler ve bunlara bağlı fonetik olma durumları değişebilmektedir. Kısacası her kelime her zaman aynı şekilde ifade edilemez.

Türkçede her kelime yazıldığı gibi okunur fakat her ses her zaman aynı şekilde söylenmez. Örneğin ‘k’ harfi ele alındığında bazen ‘ke’ bazen de ‘ka’ şeklinde seslenilir. Yani yeri geldiğinde vurgular değişmektedir. Bu durum ingilizce de daha kompleks durumdadır.

3.2. Kelime Bilimi(Morphology-Morphemes)

Kelime bilimindeki en küçük parçalar olarak ifade edilebilir. Yapım ekleri ,çekim ekleri bunlara örnek verilebilir. Örneğin -cı,-ci gibi gelen yapım ekleri bir kelime parçalanırken bölünen en küçük parçadır.

3.3. Sözdizimi(Syntax-Phrases,sub-phrases)

Kelime gruplarının belirlenmesi ,tamlamalar sözdizimi için örnek olarak gösterilebilir. Tamlamadaki her kelime kendine ait bir anlamı varken birlikte de bir anlam ifade edebilir. Örnek olarak kırmızı evin kapısı isim tamlamasında her kelime tek başına bir nesneyi ifade ederken birlikte farklı bir şeyi ifade etmeleridir.

3.4. Anlamsal(Semantics-Representation)

Kelime gruplarının bilgisayarın anlayabileceği, kelimeleri birbiriyle karşılaştırabileceği gruplayabileceği eklemeler yaparak yeni kelime türetebileceği ,üretirken kullanacağı formüller halinde ele alınabilecek bir gösterim durumudur. Kısacası mantıksal gösterine ihtiyaç duyulmaktadır.

3.5. Yorumsal (Pragmatics)

Sadece kelimelerin anlamları ya da sadece cümlede nerede oldukları değil aynı zamanda pratik anlamda hangi anlama geldikleri ,güncel hayatta nasıl kullanıldığının yorumlanmasıdır.

3.6. Söylev(Discourse)

Bir kelimenin hangi alanda ne anlama gelir bunun bilinmesi gerekir. Örnek olarak bir top kelime üzerinde düşünüldüğünde futbolda kullanılan ‘top’ kelimesi başka bir şey ifade ederken basketbol kullanılan ‘top’ kelimesi farklı bir şeyi ifade etmektedir. Ya da bambaşka bir alan olan askeri terminoloji de bambaşka bir ifade olur.

3.7. World Knowledge

İnsanların yıllardır edindikleri bilgi birikimi ;okudukları ,araştırdıkları geçmiş birikimler sayesinde kelimelerin ifade ettikleri kültürel anlamlar değişebilmektedir. Bahsedilen konuların değişmesiyle bir kelime farklı anlamlar ifade edebilir. İşte bu kelimelerin doğal dil işleme ile incelenip bilgisayarların anlamasını sağlamak başka bir problem olarak görülmektedir.

Dil bilgisi, kelimenin anlaşılmasından başlanarak, gelen eklerin yorumlanması, seslerin incelenip anlam yüklenilmesiyle hangi konuya dahil olduğu tahmin edilip insan bilgi birikimi ve kültürel anlam ile birlikte en yakın anlamı bulmaya çalışmaktadır.

4. Karşılaşılan Örnek Problemler

Koyun kelimesi ele alındığı; Koyun kelimesi tek başına anlamlı bir kelimedir. Yani ‘koyun’ kelimesi görüldüğünde ya da duyulduğunda direk akla gelen anlam hayvanlar aleminden bir hayvan olma ihtimalidir. Fakat ‘koyun’ kelimesi biraz parçalandığında koy-un ve buradan ilk kelime olan ‘koy’ ifadesi koymak olan bir fiil olur. Bu sefer koyun kelimesine tekrar bakıldığında bir nesneyi bir yere koy anlamında karşıdaki şahsa emir anlamını ifade eder. İşte burada da görülüyor ki bir kelime biraz incelendiğinde başka anlamlar barındırabilir. Bu kelimenin anlaşılır olması bu sebeple önemlidir. Bir kelime hakkında ; fiil ya da ek olarak türemiş bir kelime ya da söz öbeği şeklinde anlaşılması morfolojik seviyede birden fazla anlamını ortaya çıkarmaktadır.

Literatürde geçen art gönderim çözümleme sorunu (anaphora resolution problem) (Van der Sandt, 1992)

olarak da geçen başka bir problem ;

Ali okula otobüsle geldi. O çantasını unuttu .

Cümleleri üzerinde düşünüldüğünde ;ikinci cümlede bulunan ‘O’ ifadesi hangi ifadeyi temsil ediyor bu sorun ortaya çıkmaktadır. Bu durumun sorun olma sebebi ,bu cümlelerin başka bir dile çevrilme durumunda ‘o’ ifadesini kime göre çevirme yapılacaktır, bunun belirsizliğidir. Örnek üzerinden incelenmeye devam edildiğinde Ali kelimesi türkçede erken ismi olarak geçer ve bu ‘o’ ifadesi Ali ismini vurgulayan bir zamirdir. Bu durumda başka bir dile çevrildiğinde ‘O ’ kelimesi erkek(masculen) takısı almalıdır. İngilizcede olan ‘he’ karşılığı gibi.

Başka bir örnek üzerinden devam edildiğinde ;

Elektrik arızası cebe girdi. (Haber, 2015)

Cümlesi üzerinden düşünüldüğünde, bir insan tarafından bu cümle analiz edilebiliyor ve anlaşılıyor. Nedir bu, elektrik arızasının cep telefonuna indirilen bir uygulama sayesinde kolay bir şekilde ihbar edilebilmesidir. Fakat bunun bilgisayar için anlamına bakıldığında bu cümlelerin anlaşılabilmesi için bilgisayarın ciddi bir bilgi birikimine ihtiyacı vardır. Bilgisayarın cep kelimesinden kısaca ‘cep telefonu ’ndan bahsedildiğini , girmek kelimesinden bir uygulamanın cep telefonuna indirilme işlemini vurguladığını tahmin etmelidir.

5. Uygulama Alanları

Doğal dil işleme(DDI),günümüz teknolojilerinde bir çok alanda kullanılmaktadır bunlardan bazılarını şu şekilde sıralamak mümkündür.

5.1. Soru cevaplama(Question answering)

Doğal dil işleme konusunun ilk problemlerinden biridir. Verilen bir metin üzerinden düşünüldüğünde, bu metnin analiz edilerek anlamlı soru ve cevap çıkarılması ya da verilen sorular üzerinden o soruların cevaplarının bir metin üzerinden bulunması istenilir. Bu sorunun çözümü aynı zamanda arama motorlarının da bir problem olan ‘konuşarak (sesli)arama yapabilme sorunu’ çözülmesine de sebep olacaktır.

5.2. Otomatik tercüme(machine translation)

Bir makine tarafından bir metnin tercüme bir dilden başka bir dile çevirme işlemi yapılması istenilmektedir.

5.3. Bilginin Getirilmesi(Information retrieval)

Aranılan bir bilginin metin kaynakları arasında aranıp bulunması ve getirilmesi işlemidir.

5.4. Kelime işleme (Word Processing (Grammar checking, spell checking ,spell correction,Find/Replace))

Sıkça kullanılan bir işlem olan kelime işleme bir çok alanda kullanılan bir takım işlemlerden oluşur. Metin üzerinde yazımsal hataları kontrol etmek için dil bilgisi kontrol işlemi (Grammar checking) kullanılır. Ya da bir kelime için heceleme kontrol ve düzeltme (spell checking and spell correction) işlemleri kullanılır. Kemal Oflazer (Oflazer, (1996))tarafından yazılan bir makale de ,word programında bir kelime arama (Find/replace) işlemi yapıldığında, bul ve değiştir işleminin uygulanması istenildiğinde; örneğin bir metinde okul kelimesinin ev ile değiştirilmesi istenildiğinde ,metinde bulunan bütün okul kelimelerinin ev kelimesi ile değiştirildiği görülmektedir. Fakat burada okullar kelimesi düşünüldüğünde bu kelimenin evlar olarak değiştiği görülecektir. Bu kelime hatasının evler olarak düzeltilme durumu da hece düzeltme-kontrol(spell checking-correctoin) işlemi uygulanmalıdır.

5.5. Doğal Dil İşleme-Veri Tabanı Arayüzü (NL/DB interface)

Veri tabanı üzerinden sorgulama işlemi yapıldığında kullanılan bir işlemidir. Bir üretim departmanında soru olarak gelen ‘Bu yıl ki satış oranı nedir?’ sorusunu veri tabanında uygulamak için bir sorgu cümlesine çeviren bir sistemdir.

5.6. İnternet Arama Motorarı (Web Search)

Doğal dilde arama olarak internet arama motorları sayesinde arama yapma işlemidir. Örnek olarak ,İstanbul kayseri arası kaç km’dir? Everest Dağ’ının yüksekliği ne kadardır? Gibi sorulara cevap aranabilmesi.

5.7. Anti-Intelligence (AI) Bots(Siri)

Konuşularak verilen bir soru üzerine makine tarafından yine konuşarak karşılık verilmesi. İstenilen sorulara arama yaparak bularak karşılık vermesi ve günümüz teknolojilerinde sıkça rastlanan akıllı sistemlerde görülen durumdur. Bu durumun en bilinen örneği SİRİ uygulamasıdır.

5.8. Metin İşleme(Text processing)(Categorization,clustering,IE)

Bir metin üzerinde yapılan kategorize etme, bölütleme ve kümeleme işlemidir. Bir önceki konu olan text mining (Seker, Metin Madenciliği(Text Mining), 2015)konusunda bu konuları daha detaylı görmek mümkündür.

5.9. Finance,Chat Rooms,Telefon Cevaplama

Finansal raporlama ,sohbet odaları ve çokça rastlanılan telefon cevaplama sistemleri. Arama sonucunda karşı tarafta bir makinenin olması ve sorulan sorulara aldığı cevaplar ile yönlendirme yapması ya da sohbet odalarında sohbet diyaloglarına karşılık veren makinelerin bulunması.

5.10. Bilgi Deposu Olarak İnternet(Internet as a repository)

İnternet ortamının en büyük bilgi kaynağı ve ulaşılabilirlik açısından en kolay yol olduğu bilinmektedir. Düşünün her gün milyonlarca kişi tarafından gerek kendi sitelerinde ,gerek bloglarında gerekse kendi kişisel sayfalarında bir şeyler paylaşmaktadırlar. Paylaşılan bu bilgilerin anlaşılıp ,modellenip ve kümelenmesi durumunda bütün internet ağı üzerinde yani insanların ürettiği en büyük bilgi kaynağında ,sonuçta insanlık tarihi bir çok bilgi kaynağı üretmiştir ,geçmiş zamanlarda ,ansiklopediler ,yazıtlar gibi ve bu kaynakların hemen hemen hepsi internet kaynağına çevrilmiştir ,bilgisayarlar tarafından doğal dil işleme yöntemi sayesinde ortaya çıkması diğer insanların bu bilgiye çok kolay bir şekilde ulaşmaları demektir.

5.11. Metin Özetleme(Text summarization)

İnternet üzerindeki bir milyon tane kaynak üzerinden düşünüldüğünde bunlar üzerinden yapılan bir sonucunun bu sayfalar üzerinden bulunması ve özetleme yapılmasıdır.

5.12. Argümanların Birleştirilmesi(Argument aggregation)

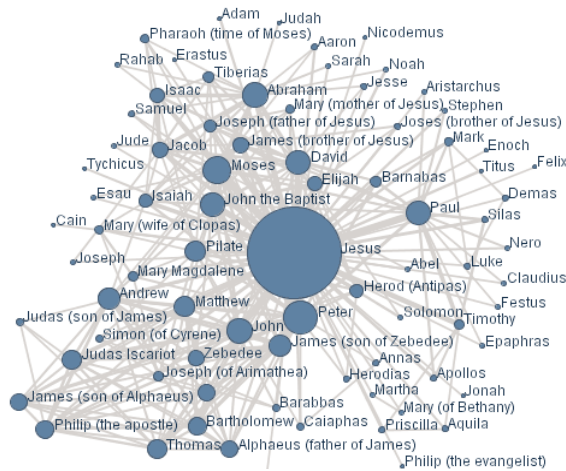
Sosyal ağlar üzerinden yapılan tartışmalar sonucu ortaya çıkan argümanların analiz edilmesi tekrarlı olanların ya da boş olanların temizlenmesi ve kalan argümanlardan anlamlı verilerin üretilmesiyle sonucun bir argümanda birleştirilmesi gerekmektedir. Bu sayede başlatılan bir tartışmaya daha kolay dahil olunması sağlanır. Daha önce hangi argümanlar konuşulmuş neler tartışılmış bunların bilinmesi durumudur.

Güzel bir örnek olarak, sosyal medya da sıkça görülen bir örneği paylaşmakta yarar var.

“A ,B’yi beğendiysen ama C’yi beğenmediyse, C ise D ve E’yi beğendiysen ve hem D hem de E ,A’yı beğendiysen ama A’nın D’nin varlığından haberi bile yoksa, E’nin, C’nin D’yi ve E’yi beğendiğini bilebilmesi için F, G’nin B ile konuşmasını teşvik etmeli mi ki bu sayede C , H’nin çevresini biraz da olsa etkileyebilsin.”

Bu sorunun size sosyal medya da sorulduğunu düşünün ve bunu bilgisayar tarafından sorulması durumunda cevap çıkmasının aranması durumu ortaya çıkacaktır.

6. Metin Madenciliği



Şekil 1-Metin Madenciliği Grafik-1

Yukarıdaki şekilde, Eski Ahit'ten yani Tevrat olarak bilinen kısımdan çıkarılmıştır. Önce metinde ismi geçen kişiler çıkarılmış ve daha sonra da bu kişiler arasındaki ilişkiler doğrultusunda grafiksel görünümü ortaya çıkmıştır.

7. Zamansal Mantik

Bir metindeki zamansal olarak olayların sıralanması hangi olay hangi olaydan sonra ortaya çıkmaktadır bunun incelenmesi ve bunu görsel olarak ortaya konulması işlemidir. Bir roman düşünüldüğünde bu romandaki tüm olayların bir zaman göre sıralanması ,hangi olayın hangi olaydan sonra geldiği konular birbirleriyle ilişkilendirilir.

Bir örnek üzerinde anlatılması gerekirse,

“ Ali odaya girdikten sonra masadaki elmayı yedi.”

Bu cümle incelendiğinde;

Dışarıda	Girdi	Odanın içerisinde		
Masadan Uzakta		Yaklaştı	Masanın Yanında	
Elması yok			Aldı	Elması Var
				Yedi

Şekil 2-Zamansal Mantık

şeklinde bir tablo ortaya çıkmıştır.Bu tablo üzeri

Bir morfoloji(morphology) örneği verilmesi gerekirse,

* uygarlaştıramayabileceklerimizdenmişsinizcesine

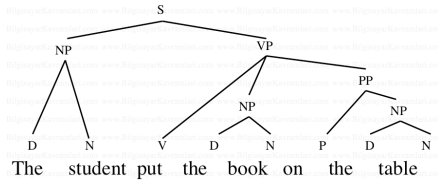
uygar/civilized las/BECOME tir/CAUS ama/NEG yabil/POT ecek/FUT ler/3PL imiz/POSS-1SG

den/ABL mis/NARR siniz/2PL cesine/AS-IF

Bu kelime üzerinden analiz yapıldığında ilk önce kelimenin kökü olan ‘uygar’ kelimesi bulunup daha sonra bu kelimeye gelen eklerin çeşitleriyle ayrıldığı görülmektedir.

8. Parçalama(Parsing)

Aşağıdaki örnek üzerinden cümlelerin parçalanma durumu incelenecektir.



Şekil 3-Parçalama İşlemi

The student put the book on the table.(öğrenci kitabı masanın üzerine bıraktı)

-S(sentence-cümle)→ ilk adım bu bir cümledir.

-NP(Noun phrase-Kelime grubu)

VP(Verb phrase-Fiil grubu)→ kelime gruplarının ayrılması

-D ve N(article(belirli bir nesne) ve noun(isim))→kelime grubundaki ifadenin ayrıştırılması

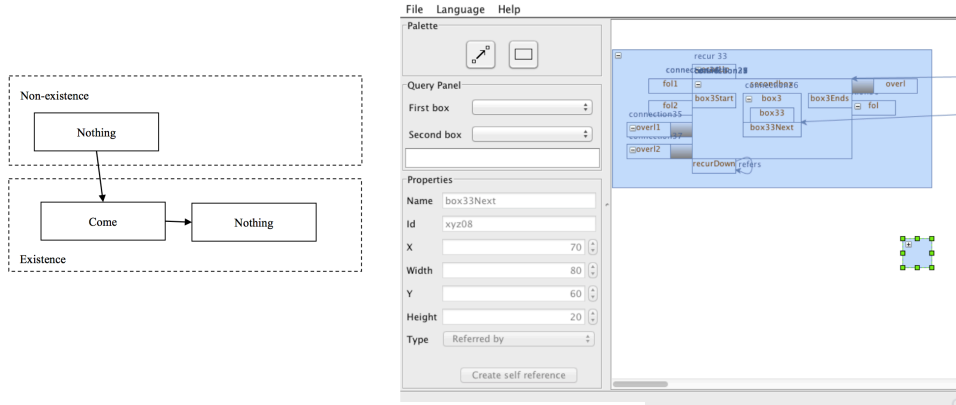
-V-Np-D-N-PP-P-Np-D-N-Son durumun anlaşılması

Buradan da görüldüğü gibi bir parçalama ağacı oluşturuluyor ve bu ağaç üzerinden cümlelerin anlamı çıkarılmaya çalışılıyor.

9. Mantık

Mantıksal bir şekilde gösterilmesi de söz konusudur.

Nothing comes from nothing. Cümlesi düşünüldüğünde ” hiçbir şeyden hiçbir şey çıkmaz” anlamı çıkarılmaktadır. Cümle incelendiğinde ilk önce var olmayan bir durumdan(non-existence),var olan(existence) bir duruma dönüştüğü düşünülmektedir (Seker, Temporal logic extension for self referring, non-existence, multiple recurrence and anterior past events, 2015) .



Şekil 4- Zamansal mantık gösterimi

10. Argüman Ağları

- SEKER, S. E. (2015). Computerized Argument Delphi Technique. *IEEE Access* ,v. 3, 368 - 380.
- Seker, S. E. (2015). Metin Madenciliği(Text Mining), YBS Ansiklopedi v. 2, is. 3, pp. 30-32.
- SEKER, S. E. (2015,). Veri Ambarları, YBS Ansiklopedi, v. 2, is. 4, pp. 6 - 13.
- Van der Sandt, R. (1992). Presupposition projection as anaphora resolution. . *Journal of semantics* , 333-377.