Radyoloji Raporları için Türkçe Bilgi Çıkarım Sistemi

Ergin Soysal[a], Ilyas Cicekli[b], Nazife Baykal[c]

[a]*Tıp Eğitimi ve Bilişimi AD, Ankara Üniversitesi Tıp Fakültesi, Ankara*
[b]*Bilgisayar Mühendisliği Bölümü, Bilkent Üniversitesi, Ankara*
[c]*Enformatik Enstitüsü, Orta Doğu Teknik Üniversitesi, Ankara*

A Turkish Information Extraction System for Radiology Reports

Abstract: *Free texts are still the main source of information in medical domain. This format is widely used for both storage and exchange of information of individual patient. Nevertheless, this form of information is not as useful as structured and coded data for decision making in medical domain. Reuse of this information requires extensive efforts.*

This paper describes a system that processes free text radiology reports in order to extract and convert the information into a structured information model. The system uses natural language processing techniques in combination with ontology as domain knowledge in order to transform verbal descriptions into a target information model that can be used for computational purposes. Although the prototype mainly concentrates on abdominal radiology, just by adapting the ontology and the rule set, the system can be used in another field of medicine. We explain how information is extracted from reports using natural language processing techniques. The results of clinical testing of the prototype are presented and evaluated.

Key words: *Information Extraction, Radiology Reports, Natural Language Processing, Ontology, Turkish*

Özet: *Serbest metin tıp alanındaki en önemli veri kaynağıdır. Bu format hastaya ait bilginin hem saklanmasında hem de değişiminde yaygın olarak kullanılmaktadır. Ancak bu veri türü karar verme ve bilgi keşfetmede yapılandırılmış ve kodlanmış veri kadar kullanışlı değildir. Bu verinin tekrar kullanılabilmesi ise yoğun emek gerektir.*

Bu makale, Türkçe serbest metin radyoloji raporlarını yapılandırılmış bilgi modeline dönüştüren bir sistemi tanımlamaktadır. Sistem doğal dil işleme tekniklerini bir alan bilgisi olarak kullandığı ontoloji ile birleştirerek, sözel tanımları, bilgisayar mantığı içinde kullanılabilecek bir hedef bilgi modeline dönüştürmektedir. Her ne kadar prototip temel olarak abdominal radyoloji üzerine yoğunlaşmış olsa da, ontoloji ve kural kümelerini uyarlayarak tıbbın başka bir alanında da kullanmak mümkün olacaktır. Burada, sistemin yapısını tartışıp, ontoloji ve doğal dil işleme yöntemlerinin raporlardan bilgi çıkarılmasında nasıl kullanıldığını açıklamaktayız. Prototipin klinik değerlendirmesinin sonuçları sunulmaktadır.

Anahtar Kelimeler: Bilgi Çıkarımı, Radyoloji Raporları, Doğal Dil İşleme, Ontoloji, Türkçe

1. Giriş

Sağlık bilgi sistemleri ve elektronik sağlık kayıtlarının sağlık bilgilerinin erişilebilirliğini artırarak sağlıkta verimliliği artırması ve böylece sağlık maliyetleri azaltmasını ve sağlık bakım kalitesinin artırmasını umulmaktadır [1]. Sağlık kayıtlarında yapılandırılmamış serbest metin hala en çok karşılaşılan veri kaynağıdır. Bu format, hem bilginin saklanması, hem de aktarılması amacı ile yaygın olarak kullanılmaktadır. Radyoloji, patoloji, nükleer tıp ve daha birçok tıbbi disiplin, verinin iletilme yolu olarak nerdeyse tamamen yapılandırılmamış serbest metinlere dayanmaktadır. Bunun yanında, bugün yapılandırılmış ve kodlanmış veri saklamaya başlasak bile, geriye yönelik olarak elektronik olarak kayıtlı çok fazla serbest metin verisi bulunmaktadır. Her bir hastanın dosyası radyoloji ve patoloji raporları, gözlem notları ve hekim kanaatleri, yatış ve çıkış özet bilgileri gibi çok sayıda serbest metin olarak kaydedilmiş bilgiyi içerir. Bu serbest metin içerisinde kayıtlı bilginin sağlık bakım hizmetlerinin en iyileştirilmesi ve verimliliğinin artırılması, hasta bakım kalitesinin artırılması, tıbbi araştırmalar ve eğitim gibi amaçlarda kullanılması için değerli bir kaynaktır. Ancak yapılandırılmamış bu veri, bu amaç için kolayca kullanılabilir halde değildir. Her ne kadar gerekli bilgiye elektronik olarak sahipsek de, bilgisayar tarafından kullanılabilir durumda değildir. "Böbrek yakınması olmayan hastalarda patolojik olmayan böbrek kistlerinin görülme sıklığı nedir?", "Bizim ülkemizde sağ ve sol böbrek ortalama boyutları nedir?", "Renal parankim ekojenitesi kanser tanısı konmadan önce zaman içinde nasıl değişir?" gibi pek çok sorunun cevaplarını bu serbest metinlerden kolayca alabilir durumda değiliz.

Bu bilginin insan eliyle çıkartılması ise uzun zaman alan maliyetli bir işlemdir. İşlenmesi gereken metin miktarı arttıkça, bilgisayar yardımı almak giderek daha zorunlu hale gelmiştir. Serbest metin formatındaki bilgi miktarı arttıkça, bu bilgiyi kullanma gereksinimi de giderek artmaktadır. Bu da bilgi çıkarımı (BÇ) ve doğal dil işleme (DDİ) tekniklerini gerekli kılar. BÇ, doğal dille yazılmış metinler içinde, önceden belirlenmiş olaylar ya da ilişki sınıflarına ait nesnelerin ve bunlara ait uygun parametrelerin belirlenmesidir.

Bu makale, radyoloji uzmanları tarafından serbest metin olarak hazırlanmış radyoloji raporlarından bilgi çıkarımı için geliştirilmiş prototip bir sistemin mimarisini tanımlar. Geliştirdiğimiz prototip Türkçe Radyoloji Bilgi Çıkarma Sistemi (TRBÇS), serbest metini, daha sonra bilgisayar tarafından işlenmeye hazır, yapılandırılmış veriye

dönüştürmektedir. Türkçe desteği bulunmasa da, benzer sorunların çözümünü hedefleyen başka sistemler de tasarlanmıştır [2-17].

RIMI sistemi [2] tıbbi raporların indekslenmesi amacına yönelik olarak geliştirilmiştir. Görüntü veritabanında, belli bir görüntüye ulaşmak için serbest metin raporlarını indeksler ve kullanır.

Haug [3] tarafından tanımlanmış olan SPRUS (Special Purpose Radiology Understanding System), ve Specialist [4,5] sistemleri raporlar içindeki terimleri kodlayarak standardize etmek ve bu yolla raporları doğrudan karar destek sistemleri içinden kullanabilmeyi amaçlar.

RadTRAC (Radiology Text Report Analyzer and Classifier) [6], sadece raporlar içerisinde bahsedilen neoplazmların tanımlanmasını hedefler. Sistem yeni keşfedilen ya da boyutu artan tümörleri saptamaya çalışır.

Do Amaral [7], semantik analizi ön planda tutan bir sistem tanımlamıştır. Serbest metin olarak kaydedilen metin içerisindeki anlamın, cümleleri oluşturan bileşenlerden çıkarılabileceğini var sayar. Ancak, bu değişik cümlelere dağılmış anlamların kaybolması sonucuna yol açmaktadır.

Hripcsak [9] serbest metin radyoloji raporlarındaki terimleri kodlamaya amaçlayan bir sistem tanımlamıştır. Metin içinde yer alan "pulmoner damarda göllenme" gibi bulguları uygun koda çevirmeye çalışır. Böylelikle bu kodlanmış veriler karar destek amacıyla kullanılabilecektir.

Linguistic String Project (LSP) [12] klinik notlardan bilgi çıkarmayı hedefler ve alt dil analizi prensibine dayanır. Analiz edeceği metine göre İngilizce'nin alt bir gramerini ve sözlüğünü kullanır. Metinden çıkardığı bilgileri ilişkisel bir veritabanına kaydetmeye çalışır.

RADA [15] torasik görüntü veritabanlarında bilgi bulunmasını kolaylaştırmak için serbest metin raporların indekslenmesi amacına yönelik olarak tasarlanmış bir sistemdir.

GENIES [16], biyoinformatik alanındaki makaleler içerisinde geçen moleküler yolları tanımlamaya yönelik olarak tasarlanmıştır.

MEDSYNDIKATE [17], tıbbi metinlerden bulguları çıkarmak için tasarlanmıştır. Aynı metinleri kullanarak, yarı otomatik olarak domain bilgisini de edinebilmeyi hedeflemektedir.

2. Gereç ve Yöntem

Ardışık 756 abdominal USG incelemesine ait rapor çalışıldı. İncelenen bu raporlarla içerisinde kullanılan terimler manuel olarak süzülerek rapor sözlüğü, ontolojisi ve yer alan bilgilere ait kurallar çıkarıldı. Bu temel üzerinde, başlıca 5 ana bileşenden oluşan TRBÇS tasarlandı (Şekil 1, Tablo 1) ve Python programlama dili kullanılarak geliştirildi.

Morfolojik Analiz

Morfoloji, gramerin en küçük birimi olan morfemler (morpheme) üzerinde çalışır. Örneğin "adamlar" kelimesi, "adam" adılı ve adılı çoğul yapan "lar" çoğul eki olmak üzere 2 morfemden oluşmaktadır. Morfemler, önek, kök ya da sonek olarak birleşerek kelimeleri oluşturur. Türkçe ekler açısından son derece zengin bir dildir.

PC Kimmo [18] benzeri, sonlu durum makineleri (finite state automata – FSA) modelini uygulayan morfoloji analiz bileşeni, radyoloji raporlarından çıkarılan kendi sözlüğünü kullanmaktadır. Bu bileşen, verilen bir kelimenin olası morfem kombinasyonlarını hesaplamaktadır. Bileşen aynı zamanda Türkçeye özgü sesli uyumu, sert sessiz yumuşatması, sessiz çiftleştirme gibi fonetik kuralları da hesaplayabilmektedir. Bu bileşen, otomatik yazılım hatalarını düzeltebilen bir bileşenle birlikte çalışır. Yazım düzeltici, yer değiştirmiş iki karakter, unutulmuş bir karakter ya da fazladan yazılmış bir karakter gibi basit yazım hatalarını otomatik olarak düzeltebilir.

Ontoloji

Ontoloji, hiyerarşik varlık sınıfları ile bunlara ait nitelikler, niteliklerin alabileceği olası değer kümesi için kontrollü bir sözlük, tıbbi terimlere ait terminoloji ve hiyerarşik lokalizasyon bilgilerinden meydana gelmektedir (Şekil 2). Radyolojik bir görüntüdeki görünebilir tüm yapılar VisibleStructure (GörünürYapı) ana sınıfından türetilir. Türetilen tüm alt sınıflar, ana sınıfa ait tüm özellikleri devralır. Patient (hasta), Device (Cihaz) ve Examination (İnceleme) raporda mevcut bilginin modellenmesindeki diğer ana sınıflardır.

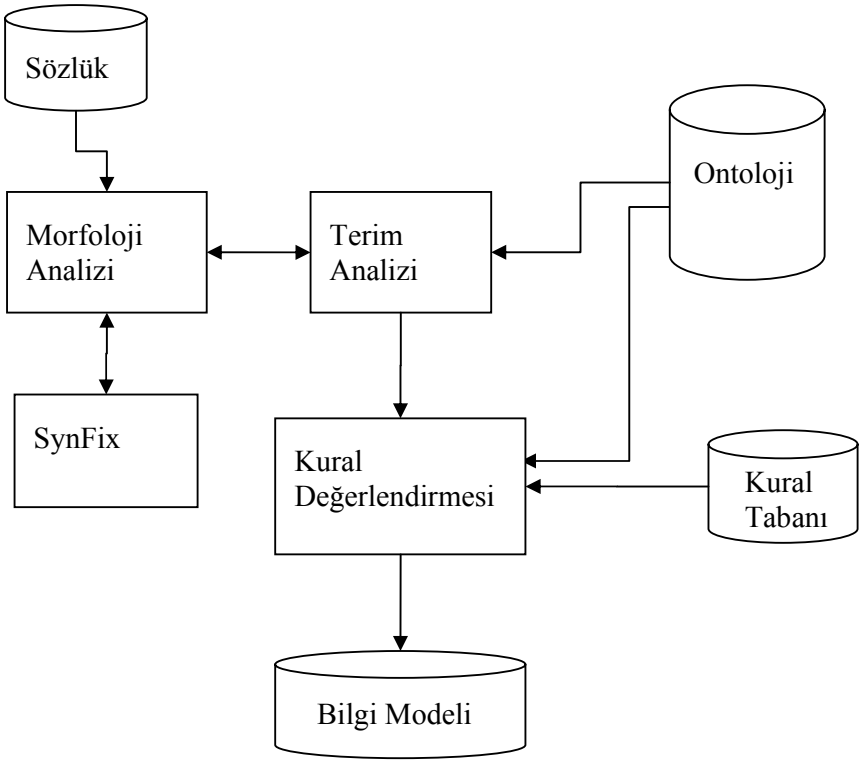
Terim Analizi

Cümledeki tüm kelimeler morfemlerine ayrıldıktan sonra cümle terim analizi bileşenine geçer. Bu bileşen tarafından ontolojinin terminoloji kısmının kullanılarak ve konu bağlamı (context) göz önüne alınarak terimler belirlenir. Ulamalı bir dil olan Türkçe nedeniyle terimlerin tanınmasında morfolojik yapılarda önemli bir rol oynamaktadır.

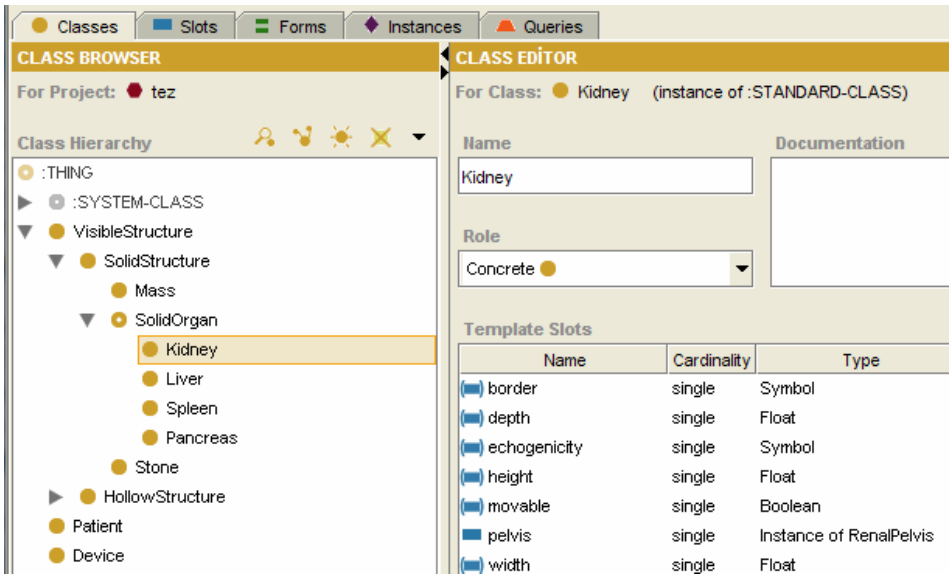
Tanımlanan terimleri ontolojik rolleri (nesne, nitelik, nitelik değeri), daha sonra kurallar işletilirken gerekli olmaktadır.

Kurallar

Eğitim seti olarak kullanılan raporlar alan uzmanı tarafından incelenerek kural setleri oluşturuldu. Kuralların çoğunluğu birkaç kısıta sahiptir. Kurallar ve kurallara ait kısıtlar ontoloji ile sıkı bir ilişki içindedir. Hem kuralların ifadesinde hem de kısıtlarında ontoloji sınıfları ve bu sınıflara ait niteliklere başvurulur. Her kuralın "bu nesne bu niteliğe sahip mi?" ya da "bu nesnenin bu niteliği bu değeri alabilir mi?" gibi ek kısıtları bulunabilir. Ayrıca morfolojik yapılar da sıklıkla kurallar içinde kullanılmıştır.



Şekil 1 – TRBÇS'nin major bileşenleri



Şekil 2 – Protégé yazılımı kullanılarak Ontoloji geliştirildi.

Bilgi Modeli

Tasarlanan sistemin önemli problemlerinden birisi, çıkarılması hedeflenen bilginin daha sonra karar destek sistemleri ya da bilgi keşfi amaçlarıyla kullanılabilirliğidir. Çıkarılan bilgi için hedef bilgi modeli, alan uzmanı görüşleri (klinisyen ve radyolog) ve Türkiye Ultrasonografi Derneği (TUD) kılavuzlarına başvurularak oluşturuldu. Tüm model ontoloji içerisine sınıflar ve nitelikler olarak bütünleştirildi.

Değerlendirme

Sistemin performansını değerlendirme amacı ile daha önce hiç kullanılmamış rastgele seçilen 50 rapor kullanıldı. Ölçüm seti test edilmeden önce sistem yapılandırması sabitlendi. Bir alan uzmanı altın standart olarak kabul edildi ve uzman tarafından çıkarılan bilgi ile sistem bulguları karşılaştırılarak, çıkarılan bilgiler Tablo 2 deki değerlere göre sınıflandırıldı.

Bilgi çıkarım sistemleri değerlendirmesinde geri çağırma oranı (recall) ve duyarlık (precision) sıklıkla kullanılan ölçütlerdir [19].

Cümle
Karaciğer vertikal yüksekliği 14 cm'dir.
Terim Analizi (İsimlendirilmiş Oluşum Tanımlaması)
[Karaciğer] [vertikal yükseklik] +ACC [14 cm] +COP.
Uyan kural
<GörünürYapı O> <O:Nitelik N> +ACC [Değer] +COP
Karşılanması gereken kural kısıtları
obj_niteligi_mi(Obje, Nitelik) – (Karaciğer, Vertikal Yükseklik) obj_nitelik_deger_uygun_mu(Obje, Nitelik, Değer) – (Karaciğer, Vertikal Yükseklik, 14cm)
Çıkarılmış Bilgi
Karaciğer.yükseklik = 14cm

Tablo 1 – Kuralların ve kısıtların uygulandığı örnek bir çümle

Bir BÇ sisteminin *geri çağırma oranı* (GÇO), doğru çıkarım sayısının var olan toplam çıkarım sayısına oranı olarak tanımlanır. Tablo 2'ye göre şu şekilde formülize edilebilir:

$$GÇO = \frac{DP}{DP + YN}$$

Duyarlık ise doğru çıkarım sayısının, sistem tarafından yapılan tüm çıkarımlar içerisindeki oranı olarak tanımlanır. Tablo 2'ye göre şu şekilde formülize edilebilir:

$$Duyarlik = \frac{DP}{DP + YP}$$

Oluşturulan tablo üzerinden sistemin GÇO ve duyarlık değerleri hesaplandı.

		AU tarafından çıkarılan	
		Evet	Hayır
tarafından çıkarılan	Evet	**DP**	**YP**
	Hayır	**YN**	**DN**

Tablo 2 – Değerlendirme tablosu. (AU: Alan Uzmanı, DP: Doğru pozitif, YP: Yanlış pozitif, YN: Yanlış negatif, DN: Doğru negatif)

3. Bulgular

Her rapor 15.18 ± 0.25 cümleden ve 107.23 ± 1.8 kelimeden oluşmuştur. Tüm kelimelerin ‰7'sinda yazım hataları vardı. Bu hataların %92.8'si TRBÇS tarafından düzeltilebilen basit hatalara sahipti: eksik tek harf (%26.13), fazla bir harf (%42.34 – sıklıkla aynı harfin tekrarı), yanlış bir harf (%16.21%) ve ardışık iki harfin yer değiştirmesi (8.1%).

GÇO ve duyarlık değerleri sırasıyla %92 ve %97 olarak hesaplandı.

4. Tartışma

Türkçe ekler açısından son derece zengin bir dildir. Zaman, iyelik, ilgeçler gibi pek çok gramer fonksiyonu kelime ekleri olarak ifade edilir. İngilizce böyle karmaşık morfemlere sahip olmadığı için DDİ sistemleri genellikle morfolojik analiz öngörülmemiştir [20]. Örneğin,

Doktor hastayı muayene etti. (Doktor hasta + ACC muayene et +PERS3SG +PAST)

Doktoru hasta muayene etti. (Doktor + ACC hasta muayene et +PERS3SG +PAST)

Morfolojik yapı göz ardı edildiğinde önemli bilgi kaybı ile karşılaşılmaktadır. Bu nedenle TRBÇS her aşamada morfolojik kurallardan yararlanmaktadır.

Rapor içinde herhangi bir bulgudan bahsederken, bulgunun tanımlanması sıklıkla birden fazla cümleye yayılmaktadır. Anlamdaki sürekliliği sağlayabilmek ve değerleri doğru nesnelere atayabilmek için TRBÇS kural kısıtları, ontoloji ve bağlam gibi birden fazla mekanizma kullanır.

Bilgi çıkarımı sırasında en önemli zorluklardan birisi de belirsizliktir [21]. Değer dağıtımı belirsizliği (örn. "Parankim ve sinüs ekosu tabiidir."), sözel ifadenin modele çevrilme belirsizliği (örn "Barsak duvarlarında aşikar duvar kalınlığı izlenmedi", barsak.duvar.kalınlık = normal olarak değerlendirilmeli), gizli nesne ya da nitelik belirsizliği (örn. "Parankimi homojendir", böbrek.parankim.eko_yapısı = homojen) gibi oluşabilen belirsizliklerin hemen tümü kural/kısıt – ontoloji – bağlam mekanizmaları ile çözülebilmiştir.

TRBÇS doğrudan serbest metinler üzerinde çalışır ve radyologlar tarafından yapılması gerekecek ek bir işleme ya da kodlamaya ihtiyaç

duymaz. Bu yöntem sistemin kabul edilebilirliğini artıran bir etmendir. Ayrıca bu yaklaşım geçmişe dönük olarak da raporların işlenebilmesini sağlayacaktır [15]. TRBÇS, raporların oluşturulması sırasında hiçbir değişiklik talep etmemektedir. Bu bazı sistemlerde uygulanması beklenen yapılandırılmış veri almaya yönelik uygulamalar [3] bulunmamaktadır.

5. Sonuç

TRBÇS, Türkçe radyoloji raporlarında alan bilgisini ontoloji olarak sisteme yerleştirmiştir. Ayrıca, anlam çözümlemesini geliştirmek için Türkçe morfoloji bilgisini kurallara ve terim analizi içine bütünleştirmiştir. Geliştirilen prototip abdominal USG raporlarında yüksek GÇO ve duyarlık değerleriyle sistemin gerçek raporlarda kullanılabilirliğini göstermiştir.

6. Teşekkür

Çalışmamız süresince destek olan ve kimliksizleştirilmiş radyoloji raporlarına erişimimizi sağlayan Prof.Dr.Serdar Akyar, Prof.Dr.Mustafa Özmen ve Prof.Dr.Utku Şenol'a teşekkür ederiz.

7. Referanslar

[1] J.M. Corrigan, M.S. Donaldson, and L.T. Kohn, *Crossing the quality chasm: A new health system for the 21st century*, Washington, DC: National Academy Press, 2001.

[2] C. Berrut, "Indexing medical reports: The rime approach," *Information Processing & Management*, vol. 26, 1990, pp. 93-109.

[3] P.J. Haug, D.L. Ranum, and P.R. Frederick, "Computerized extraction of coded findings from free-text radiologic reports. Work in progress.," *Radiology*, vol. 174, Feb. 1990, pp. 543-548.

[4] A.T. McCray and S. Srinivasan, "Automated access to a large medical dictionary: Online assistance for research and application in natural language processing.," *Computers and Biomedical Research*, vol. 23, 1990, pp. 179–198.

[5] A.T. McCray, S. Srinivasan, and A.C. Browne, "Lexical methods for managing variation in biomedical terminologies," *ANNUAL SYMPOSIUM ON COMPUTER APPLICATIONS IN MEDICAL CARE*, 1994, pp. 235–235.

[6] D. Zingmond and L.A. Lenert, "Monitoring Free-Text Data Using Medical Language Processing," *Computers and Biomedical Research*, vol. 26, Oct. 1993, pp. 467-481.

[7] M.B.D. Amaral and Y. Satomura, "Structuring medical information

into a language-independent database," *Informatics for Health and Social Care*, vol. 19, 1994, p. 269.

[8] C. Friedman, P. Alderson, J. Austin, J. Cimino, and S. Johnson, "A general natural-language text processor for clinical radiology," *J Am Med Inform Assoc*, vol. 1, Mar. 1994, pp. 161-174.

[9] G. Hripcsak, C. Friedman, P.O. Alderson, W. DuMouchel, S.B. Johnson, and P.D. Clayton, "Unlocking Clinical Data from Narrative Reports: A Study of Natural Language Processing," *Ann Intern Med*, vol. 122, May. 1995, pp. 681-688.

[10] A.M. Rassinoux, J.C. Wagner, C. Lovis, R.H. Baud, A. Rector, and J.R. Scherrer, "Analysis of medical texts based on a sound medical model.," *Proceedings of the Annual Symposium on Computer Application in Medical Care*, 1995, p. 27.

[11] P. Zweigenbaum, B. Bachimont, J. Bouaud, J. Charlet, and J.F. Boisvieux, "A multi-lingual architecture for building a normalised conceptual representation from medical language," *ANNUAL SYMPOSIUM ON COMPUTER APPLICATIONS IN MEDICAL CARE*, 1995, pp. 357–361.

[12] N. Sager, M. Lyman, N.T. Nhan, and L.J. Tick, "Medical language processing: applications to patient data representation and automatic encoding," *Methods of information in medicine*, vol. 34, 1995, pp. 140-146.

[13] M.L. Gundersen, P.J. Haug, T.A. Pryor, R. van Bree, S. Koehler, K. Bauer, and B. Clemons, "Development and Evaluation of a Computerized Admission Diagnoses Encoding System," *Computers and Biomedical Research*, vol. 29, Oct. 1996, pp. 351-372.

[14] D.A. Evans, N.D. Brownlow, W.R. Hersh, and E.M. Campbell, "Automating concept identification in the electronic medical record: an experiment in extracting dosage information," *Proceedings: A Conference of the American Medical Informatics Association / ... AMIA Annual Fall Symposium. AMIA Fall Symposium*, 1996, pp. 388-392.

[15] D.B. Johnson, R.K. Taira, A.F. Cardenas, and D.R. Aberle, "Extracting information from free text radiology reports," *International Journal on Digital Libraries*, vol. 1, Dec. 1997, pp. 297-308.

[16] C. Friedman, P. Kra, H. Yu, M. Krauthammer, and A. Rzhetsky, "GENIES: a natural-language processing system for the extraction of molecular pathways from journal articles," *Bioinformatics*, vol. 17, 2001, pp. S74–82.

[17] U. Hahn, M. Romacker, and S. Schulz, "MEDSYNDIKATE a natural language system for the extraction of medical information from findings reports," *International Journal of Medical Informatics*, vol. 67, Dec. 2002, pp. 63-74.
[18] E.L. Antworth, "PC-KIMMO: a two-level processor for morphological analysis," *Occasional Publications in Academic Computing*, vol. 16, 1990, pp. 0–88312.
[19] G. Hripcsak, G.J. Kuperman, C. Friedman, and D.F. Heitjan, "A Reliability Study for Evaluating Information Extraction from Radiology Reports," *Journal of the American Medical Informatics Association*, vol. 6, Apr. 1999, pp. 143-150.
[20] C. Friedman and S.B. Johnson, "Natural Language and Text Processing in Biomedicine," *Biomedical informatics: computer applications in health care and biomedicine*, Springer, 2006, pp. 312-343.
[21] M. Temizsoy and I. Cicekli, "An Ontology-Based Approach to Parsing Turkish Sentences," *Machine Translation and the Information Soup*, 1998, pp. 124-135.