Türkçe metin belgeleri için damgalama

Watermarking Tools for Turkish Texts

H. Mesut MERAL[2], Bülent SANKUR[1], A . Sumru ÖZSOY[2]

[1] Elektrik ve Elektronik Mühendisliği Bölümü, Boğaziçi Üniversitesi, Bebek, İstanbul
[2] Batı Dilleri ve Edebiyatları Bölümü, Boğaziçi Üniversitesi, Bebek, İstanbul
{mesut.meral, bulent.sankur, ozsoys}@boun.edu.tr

Özetçe

Metin damgalama, metinsel belgelerin içerik bilgileri ve iyelik haklarının güvenliğini yönelik bir doğal dil işleme konusudur. Bu çalışma Türkçe metinlere gizli bir kod, diğer bir deyişle, metin damgalama olanaklarını irdelemektedir. Bu amaçla biçimbilimsel ve sözdizimsel yapı değişimleri, ve eşanlamlı sözcük ve noktalama işareti değiştokuşları denenmiş, söz konusu araçların başarımları ölçülmüştür.

Abstract

Text watermarking is a recent subject of natural language processing aimed to the content security and authentication information of the text documents. This study explores possible text watermarking tools for the Turkish language. Various watermarking tools such as changes of morphological and syntactic structures, and swapping of synonyms and punctuations are investigated and their relative performance measured.

1. Giriş

Çoğulortam belgelerinin içine başka bir bilginin saklanması ve belgenin gizli damgalanması son onyılda gözde araştırma alanlarından biri olmuştur. İmge, video, audio, grafik, yazılım gibi belgelerin damgalanması için pekçok sayıda yöntem önerilmiş [1-3] ve bir kısmı da Internet ticaretini olanaklayan algoritmalar olarak kabul görmüştür. Ancak doğal dilde damgalama (DDD) alanında ilgi yeni uyanmaktadır [4-6]. DDD yöntemleri, çoğulortam belge damgalama araçlarından çok farklıdır.

DDD'nin amacı verili bir metinden bir takım dilbilimsel yöntemleri kullanarak farklı bir metin yaratmaktır. Farklı metin yaratılırken gözetilmesi gereken üç özellik şöyle sıralanır [4]: (i) anlambilimsel tutarlılık, diğer bir deyişle metnin anlamında ayrımsanabilir bir anlam farklılığının yaratılmamış olması; (ii) dayanıklılık, belge üzerinde doğal metin işleme ya da damgayı yok etme amaçlı değişikliklere rağmen gömülü mesajın korunması; (iii) kapasite ve uygulama alanı genişliği, metne yeterince uzun bir mesajı gömebilme ve metnin türünden bağımsızca damgalayabilme.

DDD'de metinde yaratılan her değişikliğe karşılık bir bit gömülmüş sayılır. Örneğin, bir sözdizimi aracı olarak, tümce yapısını etkenden edilgene çevirdiğimizde mantıksal "1" değerini, tersini yaptığımızda ise mantıksal "0" değerini koymuş oluruz. Bir sözlükbilim aracı olarak bir sözcüğün bir eşanlamlısı 1, diğeri ise 0 olarak kurala bağlanır. Daha çok eşanlamlı olabildiğinde, mesela 4 eşanlamlı sözcüğü 00, 01, 10 ve 11 örüntülerine karşı düşürmek mümkündür.

Bu çalışmada Türkçe için doğal dilde damgalama araçları irdelenecek ve bir örnek metin (Altan Öymen 25.12.2005 Radikal) üzerinde sınanacaktır. Bildirinin amacı bu alanda ilk açınsayıcı irdelemeyi yapmak ve DDD'nin Türkçe için de olabilirliğini sınamaktır. Bildirinin 2. Bölümünde metin damgalama araçları gözden geçirilecek, 3. Bölümde örnek metin üzerinde deneme örnekleri verilecek ve 4. Bölümde de sonuç ve vargılar tartışılacaktır.

2. Metin damgalama araçları

Metin damgalama için metin içerisine özgün metnin anlambilimsel yapısına zarar vermeyecek nitelikteki değişimleri yaratabilen başlıca dilbilimsel araçlar şöyle sıralanır:

2.1. Biçimbilimsel araçlar

Türkçe'de biçimbilim ve sözdizimin birbirinden kesin çizgilerle ayrılamaması nedeniyle çalışmada kullanılan biçimbilimsel araçlar 'ek düşürme' ve 'ek saklama' ile sınırlandırılmıştır. Aşağıdaki örnek (1) Türkçe'deki ek düşürme ve ek saklama işlemini göstermektedir.

Örnek 1:

Ek saklama	selam-**lar-ım-ı** ve sevgi-**ler-im-i**
Ek düşürme	selam-**Ø-Ø-Ø** ve sevgi**lerimi**

Yukarıdaki örnekte gösterilen biçimbilimsel araç dilsel yapıda herhangi bir anlam değişikliğine yol açmamaktadır. Önerilen diğer biçimbilimsel araçlar aslında sözdizim içersinde değerlendirilebilir. Nitekim, [eylem+et-/isim+cI+ol-(yardım et-/yardımcı ol-)] ya da [-sIz/olmadan (yardımsız/yardım olmadan)] değişimleri başta

biçimbilimsel olarak alınmıştı, ancak sonradan sözdizim içersinde incelemesi daha uygun bulundu.

2.2. Sözdizimsel araçlar

Türkçe'deki anlamca birbirine benzer sözdizimsel yapıların çeşitlilik göstermesi metin damgalama için önemli ölçüde fırsatlar yaratmaktadır. Etken-edilgen çatı değişimi, yan tümce değişimi, birleşen öğe değişimi, ortaç yapıları değişimi, sözcük düzeni değişimi, özne-yüklem yer değişimi gibi sözdizimsel araçlar Türkçe metin damgalamada kullanılabilir. Aşağıdaki Tablo 1'de bu araçlar için birer örnek verilmektedir.

Tablo 1: Sözdizimsel değişim örnekleri

Araç	Örnek
Etken-edilgen çatı değişimi	İşçiler kum-**u** taşıdı → Kum taşı**n**dı
Birleşen öğe yer değişimi	Öğretmenler ve öğrenciler → Öğrenciler ve öğretmenler
Özne-yüklem yer değişimi	Çalışmanın temel savı budur → Bu çalışmanın temel savıdır.
Ortaç yapıları değişimi	Ne düşünüyorum biliyor musun? diye başlay**an** konuşmamın devamını dinlememişsin → Ne düşünüyorum biliyor musun? diye başla**dığ**ım konuşmamın devamını dinlememişsin.
Yan tümce değişimi	"500 bin dolayında" dedi. → 500 bin civarında olduğunu söyledi.
Sözcük düzeni değişimi	Biraz da sıkılmışsın sohbetimden → Biraz da sohbetimden sıkılmışsın

Bu sözdizimsel yapı değişimleri tümcenin özgün yapısını bozmamakta, tümceye de anlam açısından bir zarar vermemektedir.

2.3. Eşanlamlı sözcük değiştokuşu

Eşanlamlı sözcük değiştokuşu diğer diller için sıklıkla kullanılan araçlardan biridir. Tümcenin sözdizimsel yapısında herhangi bir değişikliğe neden olmaması, bu aracın kullanımını kolaylaştırmaktadır. Türkçe'de Eşanlamlı değiştokuşu İngilizce gibi dillerdekinden çok daha kısıtlı bir şekilde ortaya çıkmaktadır. Bu kısıtlılık durumunu Türkçe'deki anlamca benzer sözcükler arasındaki küçük kullanım farklarının dilsel yapının anlamının değişen oranlarda bozulmasına neden olmasıyla açıklayabiliriz. Örneğin, '*kesin*' sıfatı için bulunan karşılıklardan biri olan '*belirgin*' "bizim kesin/belirgin bir ayrılığımız var" tümcesinde anlamı etkilemezken, "kesin bir karar verdik" tümcesinde anlamı etkilemektedir. Hem Eşanlamlı dağarcığının çok kısıtlı olmasını göz önüne alarak, hem de yukarıda değinilen eşanlamlının yol açtığı dilsel değişiklikten kaçınmak koşuluyla ad, sıfat, belirteç, bağlaç ve eylem gibi değişik sözcük türlerine ait Eşanlamlı sözcük değiştokuşu da bir metin damgalama aracı olarak kullanılmıştır.

Tablo 2: Eşanlamlı değiştokuş örnekleri

Sözcük türü	Sözcük	Eşanlamlı
Ad	Hal	Durum
Sıfat	Memnun	Hoşnut
Eylem	İzlemek	Seyretmek
Bağlaç	Ama	Fakat
Zarf	Mutlaka	Kesinlikle
Belirteç	Beraber	Birlikte

Eş anlamlı değiştokuşlarında aday sözcüğün rastlanma sıklığı gözönünde bulundurulur. Eş anlamlı değiştokuşundaki işleklik, aday sözcüğün sözcük dağarcığındaki rastlanma sıklığı x değiştokuş başarı olasılığı değerine bağlıdır. Değiştokuş edilebilir dağarcık ise rastlanma olasılık toplamlarının, sözgelimi %p olana değin yeni aday sözcüklerin kullanılması ile oluşturulur.

2.4. Noktalama işaretleri değişimi

Metin damgalama için kullanılabilecek bir diğer araç noktalama işareti değişimidir. Çalışmada virgül, üç nokta ve tırnak işareti gibi dilsel yapının anlamsal bütünlüğünü bozmayacak noktalama işareti değişimleri kullanılmıştır. Örnek (2) bir noktalama işareti değişimini örneklemektedir.

Örnek 2:

Türkçe metin belgelerinde kullanılabilecek araçları araştıran bu çalışma**[,]** üç önemli sonuca ulaşmıştır.
Türkçe metin belgelerinde kullanılabilecek araçları araştıran bu çalışma**[ ]** üç önemli sonuca ulaşmıştır.

Türkçe'nin yazım kurallarına göre öznelerden sonra virgül kullanılmasına karşın virgüllü ya da virgülsüz yazım tümcenin anlamında genellikle büyük değişikliklere neden olmaz. Yukarıdaki örneğin de gösterdiği gibi iki tümce anlamca birbirine eştir. Tablo 3'te noktalama işaretleri değişimlerine ait örnekler gösterilmiştir.

Tablo 3: Noktalama işaretleri değişimleri

Noktalama işareti	Örnek
Virgül	Diyelim ki kapı çalımış **[,]** sohbete gelmişsin → Diyelim ki kapı çalımış **[]** sohbete gelmişsin
Tırnak İşareti	Yabancı olduğunuzu hissedip mutlaka "hoş geldin" derler.→ Yabancı olduğunuzu hissedip mutlaka **[']**hoş geldin**[']** derler.
Üç nokta	Evimiz uzak şehirlerde **[...]** Evimiz uzak şehirlerde **[.]**

3. Damgalama Deneysel Çalışmaları

Bu bölümde metin damgalama araçlarının Altan Öymen'in 25.12.2005 tarihli Radikal

Gazetesi'ndeki yazısından oluşan derlem üzerinde denenmesi sonucu elde edilen bulgular özetlenmektedir.

3.1. Derlem:

Sözkonusu yazı toplam 131 tümceden oluşmaktadır. Yazıdaki sözcük sayısı ise 1061'dir. Örnek deneylerde kullanılan ana derlem bu olsa da, çalışmanın arka planında başka derlemler de kullanılmıştır. 179.817 sözcükten oluşan 2005 yılı TBMM tutanakları, 192.87 sözcükten oluşan 13.10.2005 tarihli Radikal 2 gazetesi, 950.000 sözcükten oluşan Pusula Yayınevinin bilgisayar ve roman kitapları ve 11.000.000 sözcük içeren değişik gazetelerden elde edilmiş derlem bunlar arasındadır.

3.2. Sözdizimsel araçlar:

Sözdizimsel araçların kullanıldığı 1. deneyde yazıdaki ilk 100 tümce için, sözcük düzeni, yan tümce, özne-yüklem, etken-edilgen çatı, ortaç eki ve birleşen öğe değişimi olmak üzere 6 kategoride toplam 209 'sözdizimsel yapı' değişimi yapılmış, yapılan değişimlerde tümcedeki ifadenin anlambilimsel ve pragmatik açıdan zarar görmemesi dikkate alınmıştır. Deneye ilişkin bulgular Tablo 4'te özetlenmiştir.

Tablo 4: Sözdizimsel araçların başarısı

Değişim türü	**Rastlama sıklığı**	**Başarı sayısı**	**Başarı yüzdesi**
Sözcük düzeni	102	102	%100
Yan tümce değişimi	12	6	% 50
Ortaç eki değişimi	1	1	%100
Birleşen öğe değişimi	11	10	% 91
Özne-yüklem değişimi	9	9	%100
Etken-edilgen çatı	67	67	%100
Toplam	202	198	% 98

Tablo 4'ten anlaşılabileceği gibi sözdizimsel araçların kullanılma sıklığı oldukça yüksektir (tümce başına ortalama 2 sözdizimsel değişim). Bu yapı değişimleri özgün yapıyı bozmamakta, tümceye de anlam açısından bir zarar vermemektedir. Sözdizimsel araçların başarısız olduğu durumlar daha çok yan tümce değişimleridir. Örneğin, *'"Bir ihtiyacın var mı?" derler.* →*Bir ihtiyacımız olup olmadığını sorarlar.'* gibi bir yan tümce türü değişimi anlam ve yapı yönünden tümceye bir zarar vermezken, aynı tür değişim *'...mutlaka "hoş geldin" derler*→*...mutlaka hoşgeldiğimizi söylerler...'* gibi bir yapıda uygulanamamaktadır.

3.3. Eşanlamlı sözcük değişimi:

Eşanlamlı sözcük değişimi için ad, sıfat, eylem, zarf, bağlaç ve edat olmak üzere 6 farklı sözcük türü irdelenmiştir. Eşanlamlı değişim deneyinde derlemin ilk 100 tümcesi içim toplam 142 eşanlamlı sözcük değişimi yapılmıştır. Deneyin sonuçları Tablo 5'te özetlenmiştir.

Tablo 5:Eşanlamlı sözcük değişimi başarısı

Sözcük türü	Değişim sayısı	Başarılı değişim	Başarı %
Ad	59	58	% 98.5
Sıfat	19	19	% 100
Eylem	19	19	% 100
Bağlaç	16	16	% 100
Zarf	13	13	% 100
Edat	16	16	% 100
Toplam	142	141	% 99,9

Tablo 5'ten de anlaşıldığı gibi eşanlamlı sözcük değişimi her zaman başarılı olamamaktadır. Başarısız sonuç doğuran eşanlamlı adaylarının ya birden fazla anlamı vardı ya da eşanlamlıların arasında metne bağlı küçük anlambilimsel ve kullanım farkları bulunmaktaydı. Örneğin, *[zaman-vakit*] eş anlamlı değiştokuşunda metin içersinde farklı eğilim oratya çıkmıştır. '*...geldiği zaman*→*...geldiği vakit'* değişimi başarılı (anlam değişmemekte) iken '*..yapıldığı zamanın özeliklerini...* →*...yapıldığı vaktin özelliklerini...*' değişimi başarısız (anlam değişmekte) olmuştur.

3.4. Noktalama işaretleri değişimi:

Yazıda geçen ilk 100 tümce için toplam 91 noktalama işareti değişimi yapıldı. Değişim için kullanılan noktalalama işaretleri üç nokta, virgül ve tırnak işaretlerinden oluşmaktadır. Aşağıdaki Tablo 6'da noktalama işaretleri değişimine ait bulgular yer almaktadır.

Tablo 6:Noktalama işareti değişimi başarısı

Noktalama türü	**Değişim sayısı**	**Başarı sayısı**	**başarı %**
Virgül	64	57	% 89
Üç nokta	13	13	% 100
Tırnak işareti	14	13	% 93
Toplam	91	83	% 91

Noktalama işaretlerindeki başarısızlık durumu en çok virgül/Ø değişimlerinde ortaya çıkmıştır. Bu durum virgülün Türkçe yazımda bir çok işlevi olmasından kaynaklanmaktadır. Kimi işlevler virgülün kaldırılmasına izin verirken, kimi kullanımlar aksi yönde eğilim göstermektedir. Örneğin '*...gittiğimiz kahvaltı salonunun resmi adı**[,]** 'Konak Kahvaltı Salonu'ydu.*' tümcesindeki özneden sonra kullanılan virgül kaldırılmağa izin vermekte, bir başka deyişle tümcenin anlamında bir değişime neden olmamaktadır. Ancak, *'Aslında 'lokanta' demek yanlış**[,]** Van'daki adları 'kahvaltı*

salonu'dur.' tümcesindeki iki önerme gücü olan öbeği birleştiren virgülün kaldırılması mümkün değildir, bir başka deyişle tümcenin anlamı bozulmaktadır.

3.4. Biçimbilimsel yapı değişimleri:

Biçimbilimsel araçların kullanıldığı deneyde söz konusu yazıda geçen ilk 100 tümce kullanılmıştır. Bu 100 tümce için toplam 11 biçimbilimsel araç değişimi yapılmıştır. Bu değişimlere ait bulgular aşağıdaki Tablo 7'de özetlenmiştir.

Tablo 7: Biçimbilmsel araçlar

Biçimbilimsel araç türü	Değişim sayısı	Başarı sayısı	başarı %
Ek saklama	5	5	% 100
Ek düşürme	6	5	% 83,5
Toplam	11	10	% 91

Biçimbilimsel araçların rastlanma sıklığı az olsa da başarı oranları yüksektir. Bu aracın başarısız olduğu durum ikinci bölümde örneği verilen ek düşürme yapısıdır. Örneğin, *'...Van Kalesi'ni ve civarını tavsiye ederim.* → *Van kalesi ve civarını tavsiye ederim.'* değişimi hiçbir sorun oluşturmazken, 'İçinde 12 mezar odası, bir açık hava mabedi ve çivi yazılı kitabeleri...' → 'İçinde 12 mezar oda, bir açık hava mabet ve çivi yazılı kitabeleri..' değişimi mümkün gözükmemektedir.

4. Deney Sonuçları ve Tartışma

Tablo 8 tüm deneylerin sonucunu göstermektedir. Deneylerde kullanılan DDD araçlarının başarıları birbirlerine yakın çıkmıştır. Ancak deneylerde kullanılan araçların rastlanma sıklıkları farklı özellikler göstermektedir. Buna göre, en sık rastlanan DDD aracı sözdizimsel yapı değişimleri arasından sözcük düzeni değişimidir. Sözcük düzeni değişimi tümce başına ortalama 1 sıklığında gerçekleşebilirken, ikinci en sık rastlanan sözdizimsel değişim oranı olan etken-edilgen değişimi tümce başına ortalama 0,67 sıklığında gerçekleşebilmektedir. Sözdizimsel araçların tümünü göz önüne aldığımızda bu değerin tümce başına ortalama 2 olduğunu görürüz.

Sözdizimsel araçlardan sonra eç çok rastlanan değişimi eş anlamlı değiştokuşları oluşturmuştur. Ancak, sadece bir kategorideki (ad) eş anlamlı değiştokuşu 0,5 sıklık değerinin üzerinde gerçekleşmiştir. Diğer eş anlamlı değiştokuş kategorilerinin başarı oranları yüksek olsa da rastlanma sıklıkları düşüktür. Noktalama işaretleri değişiminde de sadece virgül değişimi 0.5 sıklık değerinin üzerinde gerçekleşmiştir. Diğer noktalama işareti değişimi kategorileri ise başarı oranları yüksek olsa da düşük sıklık değerleriyle ortaya çıkmıştır. Biçimbilimsel araçlarında aynı şekilde, başarı oranları yüksek olsa da rastlanma değerleri düşüktür.

Tablo 8: Tüm Sonuçlar

		50 tümce		100 tümce	
değişim		*Başarılı*	*başarısız*	*başarılı*	*başarısız*
Noktalama işareti değişimi	*Virgül / Ø*	22	4	57	7
	Üç nokta / tek nokta	5	0	13	0
	Çift tırnak / tek tırnak	7	0	14	1
Sözdizimsel yapı değişimi	*Sözcük düzeni*	49	0	102	0
	Özne-yüklem	6	0	9	0
	Ortaç eki	0	0	1	0
	Etken / edilgen yapı	42	0	67	0
	Yan tümce türü	1	5	12	6
	Birleşen öğe	4	0	10	1
Biçimbilimsel değişimler	*Ek düşürme*	1	0	5	1
	Ek saklama	3	0	5	0
Eşanlamlı sözcük değişimi	*Ad*	18	0	58	1
	Sıfat	11	0	19	0
	Eylem	10	0	19	0
	Bağlaç	8	0	16	0
	Zarf	6	0	13	0
	Edat	6	0	16	0

Türkçe metin belgelerindeki DDD olanaklarını sınayan bu açınsayıcı çalışma Türkçe için bit gömme esasına dayalı bir DDD modelinin gerçekleştirilebilirliğini tartışmış, konu üzerinde yapılacak ileriki çalışmalar için bir temel oluşturmağı amaçlamıştır.

5. Kaynakça

1. Hartung, F., Kutter, M., "Multimedia watermarking techniques", *Proceedings of the IEEE*, Vol. 87, No. 7, s. 1079-1107, Temmuz 1999.
2. Swanson, M. D., Kobayashi, M., Tewfik, A. H., "Multimedia data embedding and watermarking technologies", *Proceedings of the IEEE*, Vol. 86, No. 6, s. 1064-1087, Haziran 1998.
3. I.. Cox, M.L. Miller, J.A. Bloom, Digital Watermarking, Morgan Kaufman, 2002.
4. M. S. Khankhalli, K.F. Hau, Watermarking of electronic text documents, Electronic Commerce Research, 2, 169-187, 2002.
5. M. Topkara, C.M. Taskiran, E.J. Delp, Natural language watermarking, SPIE Conf. On Security, Steganography and Watermarking of Multimedia Contents VI, San Jose 2005.
6. C.M. Taskiran M. Topkara, E.J. Delp, Attacks on linguistic steganography systems using text analysis, SPIE Conf. On Security, Steganography and Watermarking of Multimedia Contents VII, San Jose 2006. no: 3&4, pp. 313-336, 1996.