İNGİLİZCE TÜRKÇE ÇEVRİLMİŞ CÜMLELERDEN İSTATİSTİKİ YÖNTEMLER İLE KELİME GRUPLARININ ELDE EDİLMESİ

Yaşar Ayırkan

YÜKSEK LİSANS TEZİ
Bilgisayar Mühendisliği Anabilim Dalı
Danışman: Dr. Öğr. Üyesi Mehmet Ali Aksoy Tüysüz

İstanbul
T.C. Maltepe Üniversitesi
Fen Bilimleri Enstitüsü
Eylül, 2018

JÜRİ VE ENSTİTÜ ONAYI

Yaşar AYIRKAN'ın "İngilizce-Tükçe Çeviri Cümlelerden İstatistiki Yöntemlerle Kelime Gruplarının Bulunması" başlıklı tezi 28.09.2018 tarihinde aşağıdaki jüri tarafından değerlendirilerek "Maltepe Üniversitesi Lisansüstü Eğitim ve Öğretim Yönetmeliği"nin ilgili maddeleri uyarınca, Bilgisayar Mühendisliği Anabilim Dalında Yüksek Lisans/Doktora tezi **oy birliğiyle /** ~~**oy çokluğuyla**~~ olarak kabul edilmiştir.

	Unvanı. Adı ve soyadı	İmza
Üye (Tez Danışmanı) :	Dr. Öğr. Üyesi Mehmet Ali Aksoy TÜYSÜZ	.....
Üye :	Dr. Öğr. Üyesi Volkan TUNALI	.....
Üye :	Dr. Öğr. Üyesi Buket DOĞAN	.....

Prof. Dr. İlter BÜYÜKDIĞAN

Enstitü Müdürü

	ŞEKİL ONAY SAYFASI	Doküman No	FR-105
		İlk Yayın Tarihi	20.12.2017
		Revizyon Tarihi	
		Revizyon No	
		Sayfa	1/2

Revizyon Takip Tablosu

REVİZYON NO	TARİH	AÇIKLAMA
00	20.12.2017	İlk yayın.

ŞEKİL ONAY SAYFASI

10/26/2018

FEN BİLİMLERİ ENSTİTÜSÜ MÜDÜRLÜĞÜNE,

Aşağıda bilgileri bulunan lisansüstü öğrencinin tezi şekil yönünden tarafımca incelenmiş ve Enstitüye teslim edilmesi uygun bulunmuştur.

Anabilim Dalı Başkanı
Dr. Öğr. Üyesi Ali Akman

ÖĞRENCİ BİLGİLERİ	
ADI SOYADI	YAŞAR AYIRKAN
ÖĞRENCİ NUMARASI	111402202
ANABİLİM DALI	Bilgisayar Mühendisliği
PROGRAMI	(X) YÜKSEK LİSANS ( ) DOKTORA ( ) SANATTA YETERLİK
DANIŞMANI	Dr. Öğr. Üyesi Mehmet Ali Aksoy Tüysüz
TEZ BAŞLIĞI	İngilizce Türkçe çevrilmiş cümlelerden istatistiki yöntemler ile kelime gruplarının elde edilmesi
SAVUNMA TARİHİ	28.09.2018
e-posta	yasarayirkan@hotmail.com

İç Kapak	☒ Var ☐ Yok
Jüri Onay Sayfası	☒ Var ☐ Yok
Etik İlke ve Kurallara Uyum Beyanı	☒ Var ☐ Yok
İntihal Raporu	☒ Var ☐ Yok

Hazırlayan İlgili Birim	Kalite Koordinatörü Dr. Öğr. Üyesi Şafak GÜNDÜZ	Kurumsal Yetkili Prof. Dr. Belma AKŞİT

(Doküman No: FR-105; Yayın Tarihi 20.12.2017; Revizyon Tarihi: ; Revizyon No:00)

	ŞEKİL ONAY SAYFASI	Doküman No	FR-105
		İlk Yayın Tarihi	20.12.2017
		Revizyon Tarihi	
		Revizyon No	
		Sayfa	2/2

Teşekkür Sayfası	☒ Var ☐ Yok
Öz (Başlık-Öz-Anahtar Sözcükler)	☒ Var ☐ Yok
Abstract (Title-Abstract-Key Words)	☒ Var ☐ Yok
İçindekiler	☒ Var ☐ Yok
Çizelgeler Listesi	☐ Var ☐ Yok
Şekiller Listesi (varsa)	☐ Şekil yok ☒ Uygundur ☐ Uygun Değildir
Kısaltmalar Listesi	☒ Var ☐ Yok
Tablolar Listesi (varsa)	☐ Tablo yok ☒ Uygundur ☐ Uygun Değildir
Ekler Listesi (varsa)	☒ Ek yok ☐ Uygundur ☐ Uygun Değildir
Özgeçmiş	☒ Var ☐ Yok
Sayfa Genişliği	☒ Uygundur ☐ Uygun Değildir
Yazı Tipi	☒ Uygundur ☐ Uygun Değildir
Referans Kullanımı	☒ Uygundur ☐ Uygun Değildir
Kaynakça Yazımı	☒ Uygundur ☐ Uygun Değildir
Ekler (varsa)	☒ Ek yok ☐ Uygundur ☐ Uygun Değildir

Hazırlayan İlgili Birim	Kalite Koordinatörü Dr. Öğr. Üyesi Şafak GÜNDÜZ	Kurumsal Yetkili Prof. Dr. Belma AKŞİT

(Doküman No: FR-105; Yayın Tarihi 20.12.2017; Revizyon Tarihi: ; Revizyon No:00)

maltepe üniversitesi	ETİK İLKE VE KURALLARA UYUM BEYANI	Doküman No	FR-178
		İlk Yayın Tarihi	01.03.2018
		Revizyon Tarihi	
		Revizyon No	00
		Sayfa	1/1

Revizyon Takip Tablosu

REVİZYON NO	TARİH	AÇIKLAMA
00	01.03.2018	İlk yayın.

ETİK İLKE VE KURALLARA UYUM BEYANI

10/26/2018

Bu tezin bana ait, özgün bir çalışma olduğunu; çalışmamın hazırlık, veri toplama, analiz ve bilgilerin sunumu olmak üzere tüm aşamalarından bilimsel etik ilke ve kurallara uygun davrandığımı; bu çalışma kapsamında elde edilmeyen tüm veri ve bilgiler için kaynak gösterdiğimi ve bu kaynaklara kaynakçada yer verdiğimi; çalışmanın Maltepe Üniversitesinde kullanılan "bilimsel intihal tespit programı" ile tarandığını ve öngörülen standartları karşıladığını beyan ederim.

Herhangi bir zamanda, çalışmamla ilgili yaptığım bu beyana aykırı bir durumun saptanması durumunda, ortaya çıkacak tüm ahlaki ve hukuki sonuçlara razı olduğumu bildiririm.

Yaşar AYIRKAN

Hazırlayan İlgili Birim	Kalite Koordinatörü Dr. Öğr. Üyesi Şafak GÜNDÜZ	Kurumsal Yetkili Prof. Dr. Belma AKŞİT

(Dokuman No: FR-178; Yayın Tarihi: 01.03.2018; Revizyon Tarihi: ; Revizyon No:00)

İngilizce Türkçe Çevrilmiş Cümlelerden İstatistiki Yöntemler ile Kelime Gruplarının Elde Edilmesi

ORIJINALLIK RAPORU

%5	%4	%0	%4
BENZERLIK ENDEKSI	İNTERNET KAYNAKLARI	YAYINLAR	ÖĞRENCI ÖDEVLERI

BIRINCIL KAYNAKLAR

1	Submitted to The Scientific & Technological Research Council of Turkey (TUBITAK) Öğrenci Ödevi	%4
2	Submitted to Üsküdar Üniversitesi Öğrenci Ödevi	<%1
3	core.ac.uk İnternet Kaynağı	<%1
4	warisantamar.blogspot.com İnternet Kaynağı	<%1
5	www2.slp.ics.tut.ac.jp İnternet Kaynağı	<%1
6	docplayer.biz.tr İnternet Kaynağı	<%1

Alıntıları çıkart Kapat
Bibliyografyayı Çıkart üzerinde
Eşleşmeleri çıkar Kapat

TEŞEKKÜR

Tez çalışmasında bana desteklerini esirgemeyen danışmanım Dr. Öğr. Üyesi Mehmet Ali Aksoy Tüysüz hocama ve aileme teşekkürü bir borç bilirim.

Yaşar AYIRKAN

Eylül 2018

Babam’a

ÖZ

İNGİLİZCE TÜRKÇE ÇEVRİLMİŞ CÜMLELERDEN İSTATİSTİKİ YÖNTEMLER İLE KELİME GRUPLARININ ELDE EDİLMESİ

Yaşar Ayırkan
Yüksek Lisans
Bilgisayar Mühendisliği Anabilim Dalı
Danışman: Dr. Öğr. Üyesi Mehmet Ali Aksoy Tüysüz
Maltepe Üniversitesi Fen Bilimleri Enstitüsü, 2018

Bu çalışmada yapılmak istenen, birebir çevrilmiş Türkçe ve İngilizce metinler üzerinde, bigram yöntemi ile oluşturduğumuz eşleştirme algoritmasını kullanarak, istatistiksel olarak ikili kelime gruplarının karşılıklarını bulmaktır.

Bu çalışma içinde kullanılan algoritma ve bigram eşleştirme yönteminin başarılı olması durumunda; Türkçe İngilizce tercüme alanında ve özellikle bilgisayar destekli çeviri yazılımlarında (CATT – Computer Aided Translation Tool), kelime tamamlama, sonraki kelimeyi otomatik önerme, kelime doğrulama, konuşma ve yazım tanımlama gibi hem tercüme hem de makine çevirisi alanında kullanılabilir bir yöntem elde edilebileceği düşünülmektedir.

Anahtar Sözcükler: 1. Çeviri; 2. Bigram; 3. N-gram; 4. Doğal Dil İşleme 5. Makine Çevirisi 6. Eşleştirme Algoritması

ABSTRACT

OBTAINING VOCABULARY GROUP BY STATISTICAL METHODS FROM TURKISH ENGLISH TRANSLATED SENTENCES

Yaşar AYIRKAN
Master Thesis
Department Of Computer Enginering
Thesis Advisor: Dr. Öğr. Üyesi Mehmet Ali Aksoy Tüysüz
Maltepe University Graduate School Of Science And Engineering, 2018

In this study, we try to obtain vocabulary groups using the Turkish English translated texts by created our bigram translation algorithm with statistical methods.

It is expected to be useful for text translation especially for the human translators using computer software for the job. Also it can be used for spelling correction, suggestion in messengers, translation memories of CATT software etc.

Keywords: 1. Translate; 2. NLP; 3. N-gram; 4. Bigram; 5. Translation; 6. Translation Algorithm

İÇİNDEKİLER

JÜRİ VE ENSTİTÜ ONAYI v
İLKE VE KURALLARA UYUM BEYANI .....**Hata! Yer işareti tanımlanmamış.**
TEŞEKKÜR x
ÖZ xii
ABSTRACT xiii
İÇİNDEKİLER xiv
TABLOLAR LİSTESİ xv
ŞEKİLLER LİSTESİ xvi
KISALTMALAR xvii
ÖZGEÇMİŞ xviii
BÖLÜM 1. GİRİŞ 5
Tarihçe 5
Amaç 7
Önem 7
BÖLÜM 2. YÖNTEM 8
Araştırma Modeli 8
Algoritma 9
Veriler 14
Algoritmanın Verilere Uygulanması 14
Sonuçların Kontrolü 26
BÖLÜM 3. BULGULAR VE YORUMLAR 28
Bulgular 28
Yorumlar 29
BÖLÜM 4. SONUÇ 31
Özet 31
Öneriler 31
Kaynakça 33

TABLOLAR LİSTESİ

Tablo 1 – N-gram modelleri gruplama yöntemleri tablosu..... 8
Tablo 2 – Ayıklanan Noktalama İşaretleri..... 15
Tablo 3 – Türkçe Metinlerden Ayıklanan Kelimeler..... 16
Tablo 4 – İngilizce Metinlerden Ayıklanan Kelimeler..... 17
Tablo 5 – Türkçe bigram kelime grupları (TurkishContent_BiGrams.txt)..... 19
Tablo 6 – İngilizce bigram kelime grupları (EnglishContent_BiGrams.txt)..... 20
Tablo 7 – Zemberek ile 'evrensel' kelimesi kontrolü..... 22
Tablo 8 – Kelime ('evrensel') Çeviri Karşılıkları Dizisi (translatedWords)..... 23
Tablo 9 – İlk kelime çevirilerinin bulunduğu bigram listesi..... 24
Tablo 10 – Zemberek ile 'temel' kelimesi kontrolü..... 24
Tablo 11 – Kelime ('temel') Çeviri Karşılıkları Dizisi (translatedWords)..... 25
Tablo 12 – Bigram ('evrensel temel') için sonuç durumu..... 26
Tablo 13 – Uygulama Sonuçlarının Bir Bölümü..... 27
Tablo 14 – Öz/Abstract Verisi İçin Çeviri Sonuç Tablosu..... 28
Tablo 15 – Röportaj Verisi İçin Çeviri Sonuç Tablosu..... 28
Tablo 16 – Karışık Cümleler İçin Çeviri Sonuç Tablosu..... 29

ŞEKİLLER LİSTESİ

Şekil 1 – Bigram ile kelime gruplama 9
Şekil 2 – Bigram Çeviri ve Karşılık Bulma Algoritması diyagramı 11
Şekil 3 – Kelime köklerinin döngü ile kontrolü 22
Şekil 4 – Zemberek kütüphanesi ile kelimeyi inceleme 22
Şekil 5 – Kelime kökü (‘evren’) için sözlük sorgulama sonucu 23
Şekil 6 – Kelime (‘evrensel’) için sözlük sorgulama sonucu 23
Şekil 7 – Hedef bigram dokümanından sorgulama sonucu 24
Şekil 8 – Kelime (‘temel’) için Türkçe İngilizce sözlük sorgulama sonucu 25

KISALTMALAR

AI	**:** Artificial Intelligence (Yapay Zekâ)
CATT	**:** Computer Assisted Translation Tools
NLP	**:** Natural Language Processing (Doğal Dil İşleme)
TDK	**:** Türk Dil Kurumu
TM	**:** Translation Memory
YÖK	**:** Yükseköğretim Kurulu Başkanlığı

ÖZGEÇMİŞ

Yaşar AYIRKAN

Bilgisayar Mühendisliği Anabilim Dalı

Eğitim

Derece	*Yıl*	*Üniversite, Enstitü, Anabilim/Anasanat Dalı*
Ls.	2008	Anadolu Üniversitesi İşletme Fakültesi
Ön. Ls.	2003	İstanbul Üniversitesi Meslek Yüksek Okulu, Bilgisayar Programcılığı
Lise	1999	Haydarpaşa Teknik Lisesi, Bilgisayar Bölümü

İş/İstihdam

Yıl	*Görev*
2013-	Microsoft Sharepoint Takım Lideri, Bilgi Birikim Sistemleri
2011- 2012	Yazılım Geliştirme Uzmanı, Bilgi Birikim Sistemleri
2010- 2011	Yazılım Geliştirme Uzmanı, CSH
2005- 2009	Hastane Bilişim Yönetim Sistemi Danışman ve Proje Yöneticisi, Ergun Software Consultant (esc)
2003 – 2005	Web Developer

Kişisel Bilgiler

Doğum yeri ve yılı :1982 - KARS Cinsiyet :ERKEK
Yabancı diller : İngilizce
GSM / e-posta : 05306611827 / yasarayirkan@hotmail.com

BÖLÜM 1. GİRİŞ

Bu bölümde, araştırma konusunun tarihçesi, çeviri bellekleri, CATT (Computer Assisted Translation Tools) yazılımları hakkında genel bilgilendirme, bu yazılımların kullanım alanları ve çalışma prensipleri ile birlikte bu tezin amacı ve önemine yer verilmiştir.

Tarihçe

Dillerin birbirine otomatik çevirisi işlemlerinin bilgisayar tarafından yapılması için ilk çalışmaların 1950'li yılların başlarında başlanıldığı bilinmektedir (Aslan, 2016) (José B. Mariòo, 2006). Ancak 1960'ların ortalarına kadar, özellikle Amerika'da bu alana çok yoğun yatırım yapılmasına rağmen istenilen düzeyde başarı elde edilemeyince, çalışmalar bu tarihlerden itibaren azalmaya başlamış ve ilk başlarda özellikle devlet tarafından yapılan yatırımların kesilmesiyle de makine çevirisi alanındaki çalışmalar nerdeyse durma noktasına gelmiştir.

Ancak 1970'lerin ortalarından itibaren yine Amerika'da özellikle askeri amaçlı olarak Rusça İngilizce otomatik çeviri sistemleri kullanılmaya başlandı ve sonraki süreçte NASA bünyesinde bu sistemler daha geliştirilerek Sovyetler Birliğinin uzay çalışmalarıyla ortak projeler için kullanılmaya başlandı (Ahmet TARCAN, 2008).

1980'li yılların başından itibaren ise Martin Kay tarafından önerilen çeviri bellekleri yönteminin kullanılmaya başlanması (Aslan, 2016) (Christensen & Schjoldager, 2010), özellikle tercüme alanında yeni ve daha hızlı bir yöntemin oluşmaya başlamasına sebep olmuştur. Çeviri bellekleri yöntemi ile tercüme alanına bilgisayarların bir yardımcı unsur olarak katılımı sağlanmıştır.

TM (Translation Memory) yani çeviri belleklerinin temel çalışma mantığında, kaynak ve hedef metinlerdeki eşleşmeler (kelime grupları, cümleler, tanımlalar, isimlendirmeler vb..) bir veritabanına kayıt edilir. Bu eşleşmeler birebir olmak zorunda değil benzer eşleşmeler de çeviri belleğine dahil edilebilir (Barachi & diğerleri., 2007).

Tercüme esnasında ise önceden çevrilmiş olan bu eşleştirmeler çeviri belleğinden alınarak kullanılır. Bu yöntem 1990'lı yılların başından itibaren CATT (Computer Aided

Translation Tools) yazılım araçlarının geliştirilip, tercüme alanında aktif olarak kullanılmaya başlanmasına da alt yapı hazırlamıştır (Christensen & Schjoldager, 2010).

Çeviri bellekleri yöntemi sonraki yıllarda Avrupa Birliği Komisyonu Tercüme Genel Müdürlüğü tarafından 24 üye ülkenin dilleri arasındaki tercüme işlemlerinde kullanılmaya başlanmıştır.

1990'lardan itibaren ise özellikle internetin yaygınlaşması ile birlikte otomatik çeviri alanındaki çalışmalar tekrar hızlanmaya başlamıştır.

Günümüzde özellikle Avrupa dillerinin makine çevirisi ile birbirine otomatik tercümesi çok başarılı bir noktaya gelmiştir. Ancak aynı başarı diğer dillerde henüz yakalanamamıştır.

Türkçe için ise hem çalışmaların yetersiz olması hem de dilin kendi yapısından kaynaklı zorlukların (Kemal Oflazer, 2007) bulunması ve diğer dillere otomatik çevirisi noktasında bu zorlukların henüz aşılamaması yüzünden, otomatik makine çevirisi alanında istenilen düzeye henüz gelinememiştir.

Türkçe'ye bilgisayar destekli çeviri yazılım araçları olarak de çevrilen CATT yazılımları, 1990'lı yılların başından bu yana çevirmenler tarafından daha hızlı ve doğru bir çeviri için, çeviriye yardımcı yazılım araçları olarak kullanılmaktadır (Huang & diğerleri., 2015).

Temel çalışma mantığı şu şekildedir: Çeviri bellekleri (TM- Translation Memory) oluşturulup, oluşturulan bu bellekler ile çeviri anında birebir eşleşen ya da benzerlik bulunan metinler tekrar tercüme edilmek istendiğinde, çevirmene daha önce çevrilmiş halini sunarak çevirisine yardımcı olunmaktadır. CATT üzerinde çeviri bellekleri oluşturulurken, bellek üzerinde değişiklik yaparak bellekteki çeviri kalitesi de yükseltilebilir.

Piyasada hali hazırda kullanılan birçok CATT yazılımı mevcuttur. Birden fazla tercümanın aynı proje üzerinde birlikte çalışmasına da olanak sağlama gibi çeşitli özellikleri barındıran bu araçlar, genel olarak aşağıdaki ortak özelliklere sahiptirler:

- Çeviri Bellekleri ile yapılan çevirilerin saklanması/kaydedilmesi (TM - Translasion Memory)
- Terim Bazlı Çeviriler (Kurum, organizasyon, ürün vb. isimlendirmeler)

- Sözlük
- Kelime Doğrulama (checking spellings)

Amaç

Bu araştırmada temel amaç İngilizce ve Türkçe dillerinde birbirine karşılık gelen kelime gruplarının elde edilmesidir. Bu sayede, en azından ikili kelime grupları bazında iki dil arasında çeviri kolaylığı sağlanabileceği düşünülmektedir.

Sonraki süreçte bu yöntemin daha da geliştirilerek önce üçlü (trigram) kelime gruplarının karşılıklarının bulunması ve daha da ilerde daha uzun kelime grupları için bir alt yapı özelliği taşıması hedeflenmiştir (Sami Virpioja, 2007).

Bu sayede, öncelikle Türkçe İngilizce tercüme alanında CATT yazılımlarına ikili kelime grupları bazında eşleştirme ile ikili kelime grubu çevirisi sunma, sonraki kelimeyi önerme, cümle tamamlama, kelime doğrulama gibi yeni özellikler katacak bir eklenti elde edebilmek amaçlanmaktadır. Sonraki süreçte de bu özelliği başka dillere de uygulayarak tercüme alanındaki yazılımlara katkı sağlanabileceği düşünülmektedir.

Önem

Bu çalışma ile bigram metodunun Türkçe İngilizce birebir çeviriler üzerinde denenmesi sağlanacak, kullanılan algoritmanın bu yöntem üzerindeki başarı oranı test edilecektir.

Hem elde edilecek başarı oranı hem de çalışma boyunca ortaya çıkan problemlere aranacak çözüm yöntemleri ile iki dilin birbirine otomatik çevirisi alanındaki çalışmalara katkı sunulmaya çalışılacaktır.

BÖLÜM 2. YÖNTEM

Bu bölümde, araştırma modeli, algoritma, veriler, algoritmanın verilere uygulanması ve sonuçların insan gücü kullanılarak kontrolü sonucu ortaya çıkan bilgilere yer verilmiştir.

Araştırma Modeli

Eşleştirme algoritması oluşturmak için n-gram modeli (Jurafsky, 2017) içinde bulunan bigram kelime gruplama yönteminden yararlanıldı. N-gram modelleme unigram, bigram, trigram, four-gram… vb. şeklinde ilerleyen gruplama yöntemlerine sahip, özellikle doğal dil işleme (NLP) uygulamalarında sıkça kullanılan bir modeldir (Chunyu Kit, 1998). Aşağıdaki tablo da n-gram modelindeki gruplama yöntemlerinin kullanımı ile ilgili bir tablo verilmiştir.

"*Bir kelime gruplama yöntemidir*" cümlesi için;

n	**Model**	**Örnek**
1	unigram	{'Bir', 'kelime', 'gruplama', 'yöntemidir'}
2	bigram	{'Bir kelime', 'kelime gruplama', 'gruplama yöntemidir'}
3	trigram	{'Bir kelime gruplama', 'kelime gruplama yöntemidir'}
4	four-gram	{'Bir kelime gruplama yöntemidir'}

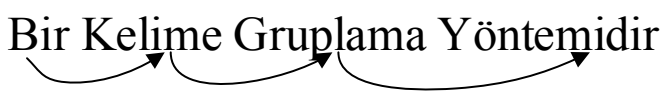
Tablo 1 – N-gram modelleri gruplama yöntemleri tablosu

Bu modelin doğal dil işlemede temel kullanım matığı; bir önceki kelimeye bakarak sonraki kelimenin ne olabileceğini bulmaya çalışmaktır.

Bigram yöntemi ile bu araştırmada temel amaç: birebir çevrilmiş İngilizce Türkçe metinleri karşılıklı parçalayıp, ikili kelime gruplarının tam karşılıklarını bulmaya çalışmaktır. Bu işlem esnasından birden fazla çıkan eşleştirmeler için ise istatistiksel olarak en fazla karşılık gelen ikilileri bulmak ve çıkan sonuçların doğruluklarını insn gücü kontrol etmektir.

Bigram yönteminde temel hedef metni kelime kelime parçalara ayırırken, çıkan her kelimeyi bir önceki ile ve bir sonraki gruplamaktır (Peter F. Brown, 1990) (Jurafsky, 2017) (José B. Mariòo, 2006).

"Bir Kelime Gruplama Yöntemidir" cümlesi için;



Şekil 1 – Bigram ile kelime gruplama

n adet kelime için; (n-1|n) (n-2|n-1) (n-3|n-2)… şeklinde gruplayabilmek için aşağıdaki bigram formülü kullanılır (José B. Mariòo, 2006) (Peter F. Brown, 1992).

$$p(w) = \prod_{i=1}^{k+1} p(w_i|w_{i-1})$$

p("Bir kelime gruplama yöntemidir") = p(Bir|kelime) p(kelime|gruplama) p(gruplama|yöntemidir)

Algoritma

Yukardaki bigram yapısı kullanılarak birebir çevirisi olan metinlerde uygulanmak üzere, aşağıda adımları teker teker açıklanmış ve diyagramı çizilmiş bir algoritma geliştirildi.

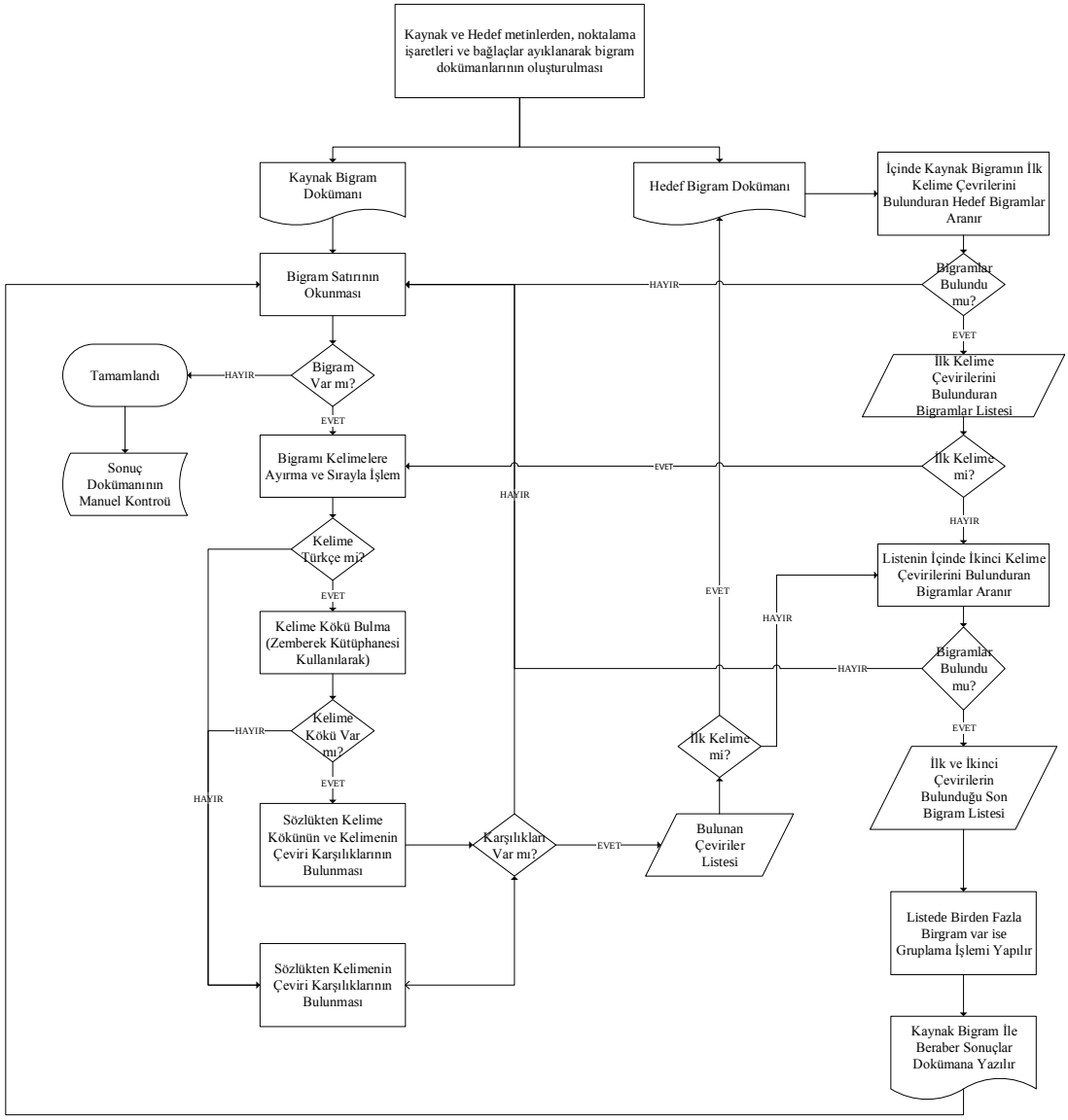
Algoritma birebir çevrilmiş, Türkçe (Kaynak) ve İngilizce (Hedef) olan metinler üzerinde kurgulandı. Türkçe metinlerdeki kelimelerin, Türkçe olup olmadıklarını, Türkçe ise kelime köklerini, tipini ve çoğul olup olmadığını bulmak için Zemberek (Zemberek-NLP, 2018) kütüphanesi kullanıldı

Algoritma temel olarak 6 bölümden oluşmaktadır:

1. Bigramlara parçalama: Kaynak ve hedef metinlerden, noktalama işaretleri ve bağlaçlar temizlendikten sonra, bigram metoduyla parçalanıp, satır satır bütün kelime gruplarının yazıldığı iki adet doküman elde edilmesi.

2. Sözlükten karşılıklarını bulma: Satır bazlı olarak, kaynak bigram kelimelerinin sözlükteki karşılıklarını bulma, bulunan kelimeleri dizilerde tutma.
3. Hedef bigamlar ile karşılaştırma: Bulunan sözlük karşılıklarının, hedef bigram dokümanında tam karşılıklarının bulunup listelenmesi.
4. Gruplama ve yazdırma: Bulunan sonuçların gruplandırılması ve bir sonuç dokümanına aktarılması.
5. Tekrarlama: İkinci adımdan başlayarak, bütün kaynak bigram satırları tamamlanana kadar bu işlemin tekrarlanması
6. Sonuçların kontrolü: Elde edilen sonuçların, bir sonuç dokümanına yazılması ve doküman üzerinde bulunan karşılıkların doğruluklarının insan gücü ile kontrolü.

Adımlarını kısa başlıklar halinde özetlemeye çalıştığımız algoritmamızın, diyagram şeması ve detaylı açıklamaları ise aşağıdaki gibidir.



Şekil 2 – Bigram Çeviri ve Karşılık Bulma Algoritması diyagramı

1. Bigramlara Parçalama:

   Öncelikle Kaynak ve Hedef dokümanların içerikleri tarafımızdan geliştirilen bir yazılım ile okunup, noktalama işaretleri, bağlaçlar, zamirler, soru kelimeleri vb. önceden belirlenen kelimeler ayıklanır. Bu ayıklama işlemi Türkçe ve İngilizce bazında, dillerin özelliklerine göre önceden belirlenmiş kelimeler üzerinden yapıldı. Bu kelimelerin listesi algoritmanın verilere uyarlanması aşamasında verilecektir.

Ayıklanmış dokümanlar bigram yöntemi (Jurafsky, 2017) (Tahasildar, 2015) ile parçalanarak yeni iki adet bigram listesi bulunan, kaynak ve hedef dokümanı elde edilir. Elde edilen dokümanlara bigramlar satır satır yazılır.

2. Sözlükten karşılıklarını bulma:

   Kaynak doküman üzerinde satır satır parçalanmış kelime gruplara okunmaya başlanır. Her bir satırda bir bigram yani iki kelime bulunur, bu kelimeler bir diziye atılarak teker teker üzerlerinde işlem yapılmaya başlanır.

   Kaynak doküman Türkçe olduğu için yapılan işlemde kelime kökünü bulmak amacıyla Zemberek kütüphanesi kullanıldı. Hedef dil olarak kullanılan İngilizce için morfolojik analiz yapılmadı.

   İlk kelime Türkçe-İngilizce sözlükten kelimenin tam karşılığı bulunur, birden fazla karşılığı olabilir. Hem kelimenin kendisi hem de kelime kökü bazında ne kadar karşılığı var ise sözlükten çekilir ve bir diziye aktarılır. İlk kelimenin sözlükte bir karşılığı yok ise (özel isim olabilir, kısaltma olabilir, kullanılan sözlükte bulunmayabilir vb.) bu bigram satırı için işlem sonlandırılır ve sonraki satıra geçilerek, yeni kelime grubu için çeviri işlemine baştan başlanır.

   Bigramda bulunan ikinci kelimenin çevirisine geçmeden önce, ilk kelime için bulunan sonuçların listesi hedef bigram üzerinde sorgulanarak, içinde ilk kelime çevirilerinin bulunduğu yeni bir hedef bigram dizisi elde edilir. Daha sonra ikinci kelimenin, aynı birince kelimede yapılan işlemler gibi sözlükte karşılıkları bulunur ve bir diziye atılır, eğer ikinci kelimenin de sözlükte karşılığı yok ise yine bu bigram satırı için işlem sona erer ve sonraki satır için baştan başlatılır.

3. Hedef bigramlar ile karşılaştırma

   Kaynak bigram dokümanından alınan her bigram satırındaki ikili kelime grupları, iki aşamalı olarak hedef bigramlar ile karşılaştırılır;

   a. İlk kelime çevirisini hedef bigram dokümanında arama

      Sözlükten karşılığı bulunan bigramın ilk kelimesi için oluşan çeviri sonuç dizisindeki bütün kelimeler için, hedef bigram dizisinde sorgulama yapılır. İçinde bu kelimelerden herhangi bir tanesini bulunduran bütün bigramlar için yeni bir dizi oluşturularak, kelime grubumuz için olası sonuç bigramları dizisi elde edilir. İkinci kelimenin çeviri sonuçları artık bu yeni ve daha kısa bigram dizisi üzerinde sorgulanacaktır.

b. İkinci kelime çevirisini oluşturulan yeni hedef bigram dizisinde arama

İlk kelime için hedef dokümanından elde edilen yeni bigram dizisi üzerinde sorgulama yapmak için, öncelikle ikinci kelimenin ve kökünün sözlükten karşılıkları aynı ilk kelimedeki gibi çekilir ve bir diziye aktarılır. İlk kelime için bulunan yeni hedef bigram dizisi üzerinde doğrudan ikinci kelime çeviri sonuçları ile sorgulama yapılır. İçlerinde yalnızca ikinci kelime çevirilerinden herhangi biri bulunan bigramlar seçilerek, yeni bir sonuç bigram dizisi bulunmaya çalışılır.

Bulunan bu son bigram dizisi artık bizim için kaynak dokümandan alınan bigram kelime grubunun olası doğru karşılıklarıdır.

4. Gruplama ve yazdırma

Bulunan son hedef bigram dizisindeki aynı ikililer gruplanarak diziye son hali verilir. Kaynak dokümandan alınmış bigram ve son hali verilerek gruplanmış sonuçlar bir excel dokümanına birlikte kayıt edilir.

Kayıt aşamasında hangi karşılıklardan kaç tane bulunduğu da yazılarak olası doğru sonucun istatistiksel olarak da tahmin edilmesi sağlanmaya çalışılır.

5. Tekrarlama

Kaynak dokümanda bulunan her bir bigram satırı için, ikinci adımdan (sözlükten karşılıklarını bulma) başlayarak tekrar eder. Bulunan sonuçların hepsi aynı dokümana yazılır ve sonuçların doğruluğunun insan gücü kontrolü için hazır bir sonuç dokümanı elde edilir.

6. Sonuçların kontrolü

Tüm sonuçların yazıldığı dokümanda, kaynak bigram, karşılığı olarak bulunan hedef bigramlar ve bunların doküman içinde geçen toplam sayıları kayıt edilir.

Bu çalışmada bulunan bigram karşılıklarından en fazla olan ilk dört karşılığı sonuç dokümanına kayıt edilmiştir.

Çıkan sonuçlar; bu doküman üzerinde doğru ya da yanlış, birden fazla karşılığı bulunan sonuçlar için hangisinin doğru olduğu el ile işaretlenerek, bulunan kelime gruplarının karşılıklarının netleştirilmesi sağlanır.

Veriler

Bu çalışmada içerikten bağımsız olarak birden fazla metin türü ile çalışılmış ve testler yapılmıştır. Türkçe İngilizce çevirisi bulunan metinler içerisinde, dil öğreniminde kullanılmak üzere hazırlanmış, karşılıklı sayfalar şeklinde çevirisi bulunan; çocuk hikâye kitapları, birebir çevrilmiş rastgele cümlelerden oluşan uzun metinler, tez başlıklarında kullanılan Türkçe ve İngilizce öz/abstract bölümleri ve İngilizce çevirisi bulunan Türkçe köşe yazılarından yararlanılmıştır.

Ayrıca özellikle kelimelerin doğru karşılıklarını bulup karşılaştırma yapabilmek ve geliştirilen uygulamada kullanmak için altmış bir bin yüz doksan üç kelimelik bir İngilizce Türkçe sözlük kullanılmıştır.

Algoritmanın Verilere Uygulanması

Geliştirilen algoritma C# dili kullanılarak bir Windows console uygulamasına dönüştürüldü. Bu uygulamada yukarda "Veriler" kısmında belirtilen, çeşitli Türkçe İngilizce çevirilerin içinden, birbirini karşılayan kelime grupları ve karşılıkları bulunup bir Microsoft Excel dokümanına aktarılması sağlandı. Ve sonuç dokümanına yazılan sonuçların doğrulukları insan gücü ile kontrol edildi.

İngilizce yapılmış bir röportajın birebir Türkçeye çevrilmiş bir paragrafı için uygulamanın nasıl çalıştığı, bigramları ve sonuçları aşağıdadır.

Öncelikle hem İngilizce hem de Türkçe için txt uzantılı birer içerik dokümanı oluşturuluyor (EnglishContent.txt, TurkishContent.txt). Bu dokümanlara içerikler ekleniyor.

EnglishContent.txt içeriğine aşağıdaki İngilizce paragrafı ekliyoruz.

"*Universal basic income is a bad idea. Some of its proponents are free-market extremists who are dreaming of replacing the welfare state (such as it is in the United States but of course much stronger in Europe) with a universal basic income. This would be a huge reduction in redistribution, and would have terrible consequences for social mobility. Some proponents are those who would like to increase redistribution, but this is a terrible way of redistributing income, because you are giving it to lots of people who don't need redistribution. Good redistribution is targeted redistribution."*

TurkishContent.txt içeriğine ise yukardaki cümlenin, röportajı yapanlar tarafından yapılmış aşağıdaki Türkçe çevirisini ekliyoruz.

"*Evrensel Temel Gelir (ETG) kötü bir fikir. Bunun bir kısım savunucuları refah devletini (ABD'de ve çok daha etkin biçimde Avrupa'da olduğu gibi) evrensel temel gelir ile değiştirmek isteyen vahşi serbest-piyasa yanlılarıdır. Bu, yeniden tahsisi fevkalade azaltır, sınıflar arası geçişkenliği öldürür. Öte yandan, Evrensel temel gelirinin bir takım savunucuları ise yeniden tahsisi artırmak isteyenlerdir. Fakat Evrensel temel gelir, yeniden tahsisi artırmanın berbat bir yolu, çünkü yeniden tahsise ihtiyaç duymayan birçok insana temel gelir sağlıyorsunuz. En iyi yeniden tahsis, hedefi belli yeniden tahsistir.*"

Türkçe paragraf içinde "*Evrensel Temel Gelir*" cümlesi kısaltılarak ETG olarak yazar tarafından yazılmıştır. Ancak daha açıklayıcı olması için bu örnekte bu kısaltmanın uzun hali tercih edilerek olduğu gibi paragrafa yazıldı.

Hazırlanan bu iki doküman, bigramlarına ayrılmadan önce içlerinden noktalama işaretleri, bağlaçlar, zamirler, soru kelimeleri vb. aşağıda listeleri verilmiş belirli bazı kelimeler ile birlikte bir karakterden uzun boşluklarda ayıklanıyor.

Ayıklanan Noktalama İşaretleri			
Nokta	.	Eğik Çizgi	/
Virgül	,	Ters Eğik Çizgi	\
Soru İşaret	?	Kaşlı Ayraç - Sola	{
Noktalı virgül	;	Kaşlı Ayraç - Sağa	}
Kesme İşareti	'	Köşeli Ayraç - Sola	[
Orta Kısa Çizgi	-	Köşeli Ayraç - Sağa	]
Yay Ayraç-Sola	(	Alt Kısa Çizgi	_
Yay Ayraç - Sağa	)	Küçüktür	<
İki Nokta	:	Büyüktür	>
Ünlem	!	Eğik Çizgi	/

Tablo 2 – Ayıklanan Noktalama İşaretleri

Ayıklama işlemi esnasında noktalama işaretleri ve bir karakterden uzun boşluklar tek karakterlik bir boşluk olarak çevrilmiştir (replace).

Türkçe Metinlerden Ayıklanan Kelimeler		
ve	olan	bize
veya	gibi	size
ile	çok	üzere
her	var	diye
ya	ayrı	göre
da	elde	mümkün
de	tam	nasıl
en	hem	neden
az	biri	nerede
çok	ki	kim
bir	iyi	ne
şey	hale	için
için	önce	arada
fakat	sonra	lütfen
ayrıca	neden	fazla
ama	sebep	ilişkin
daha	biz	ilgili
kadar	siz	değil
hepsi	ben	kez
bu	sen	defa
şu	bana	ancak
o	sana	lakin

Tablo 3 – Türkçe Metinlerden Ayıklanan Kelimeler

İngilizce Metinlerden Ayıklanan Kelimeler			
and	ever	than	did
or	wherever	any	may
he	whoever	some	might
she	whatever	few	can
it	such	got	could
I	if	just	be
you	unless	who	been
we	nonetheless	vs	will
they	many	why	but
them	more	while	able
him	much	hence	must
his	most	henceforth	not
its	here	whats	yet
their	there	havent	shall
your	over	dont	was
our	all	doesnt	were
my	one	hasnt	let
her	each	didnt	less
me	this	couldnt	theyll
what	that	wont	cant
where	for	should	other
which	because	shouldnt	no
how	since	isnt	each
when	as	arent	already
has	well	wasnt	to
have	also	werent	of
do	so	would	nor
does	then	wouldnt	the
am	want	youll	hell
are	these	those	shell
is			

Tablo 4 – İngilizce Metinlerden Ayıklanan Kelimeler

Uygulama, paragrafın üzerinde nokta (.) işaretine kadar cümle bazlı alarak işlem yapmaya başlıyor. Öncelikle cümleden noktalama işaretlerini ayıklıyor ve noktalama işareti olmayan kelime bazlı bir diziye çevirip, sonrada cümlenin Türkçe veya İngilizce olup olmamasına göre içinden, yukarıdaki tablolarda belirtilen kelimeleri ayıklayarak,

yeni bir dizi elde ediyor ve elde edilen bu dizi, bigramlarına ayırmaya hazır hale gelmiş oluyor.

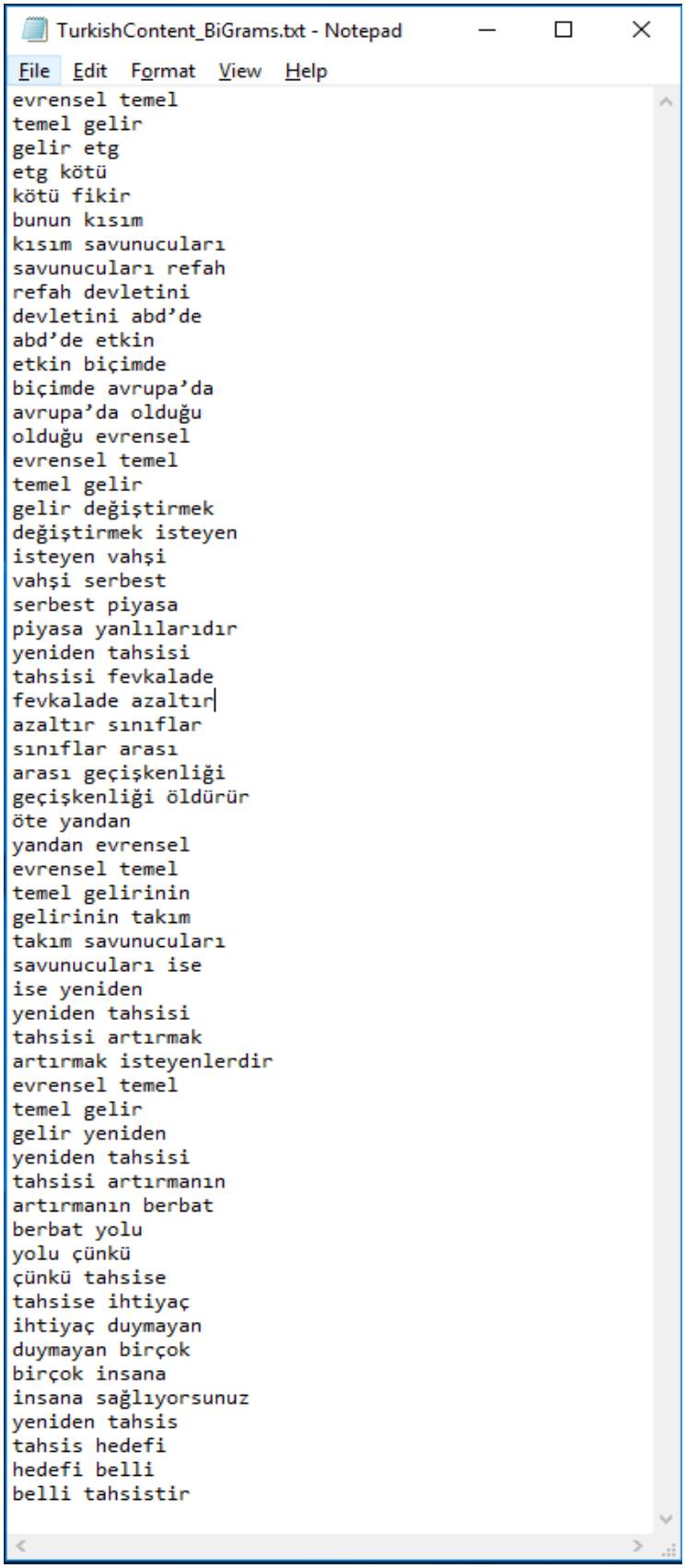
Örnek metinlerin ilk cümlelerini ele alırsak;
Türkçe için : "*Evrensel Temel Gelir (ETG) kötü bir fikir.*"
İngilizce için : "*Universal basic income is a bad idea.*"

Öncelikle noktalama işaretlerini ayıkladığımızda:
"*Evrensel Temel Gelir ETG kötü bir fikir*" → parantez '()' işaretleri ve nokta '.' ayıklandı.
"*Universal basic income is a bad idea*" → nokta (.) işareti ayıklandı.

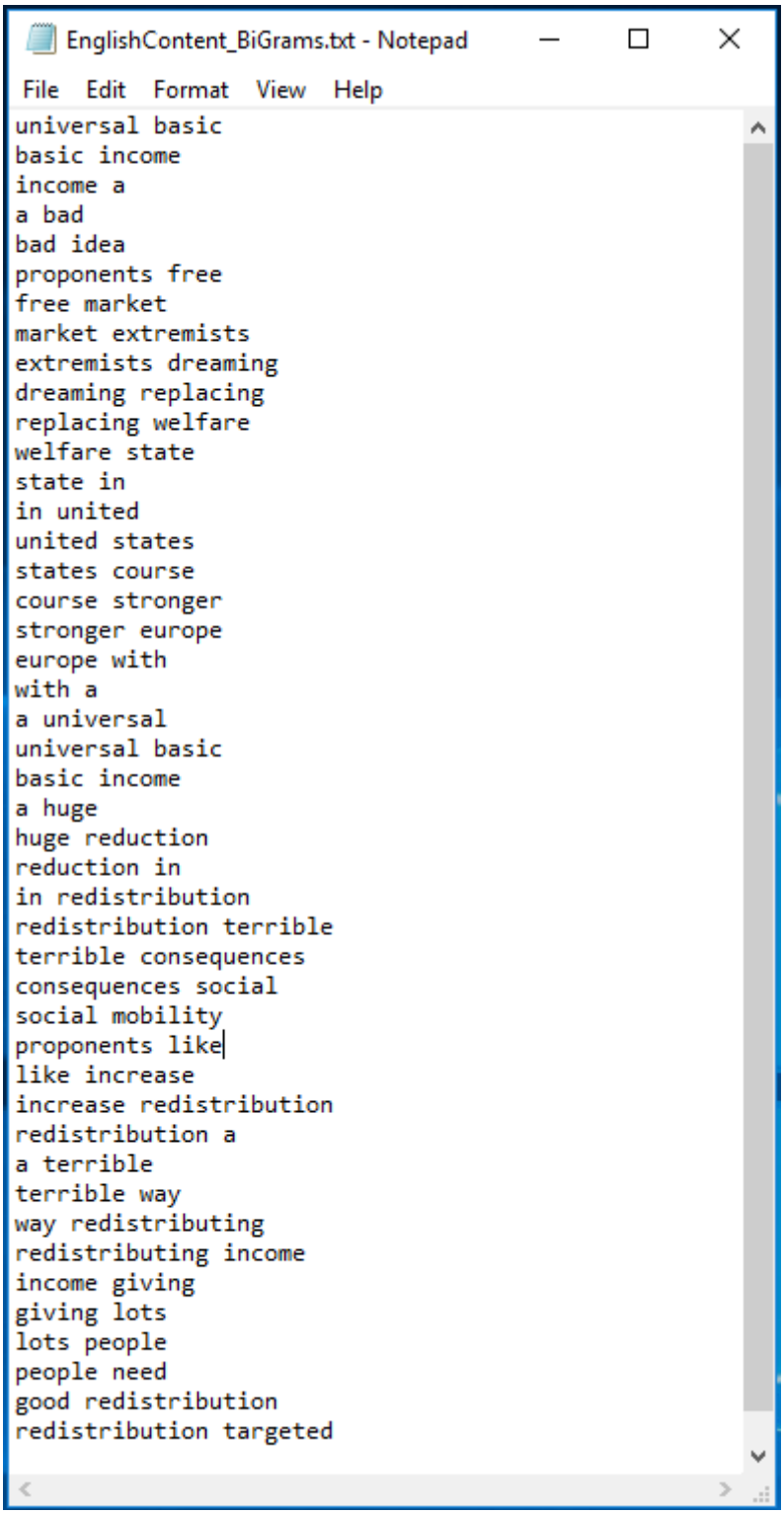
Önceden belirlediğimiz kelimeleri ayıkladığımızda ise;
"{*Evrensel Temel Gelir ETG kötü fikir}*" → 'bir' kelimesi ayıklandı.
"{*Universal basic income bad idea}*" → "is" ve "a" kelimeleri ayıklandı.

Şimdi de cümlelerin son hallerini bigramlarına ayırıyoruz.
Türkçe → {'evrensel temel', 'temel gelir', 'gelir ETG', 'ETG kötü', 'kötü fikir'}
İngilizce → {'universal basic', 'basic income', 'income bad', 'bad idea'}

Yukarıdaki gibi bütün metni bigramlarına ayırıp, oluşan bigram kelime gruplarını yazmak için iki yeni txt uzantılı doküman oluşturuluyor. Türkçe için TurkishContent_BiGrams.txt, İngilizce için EnglishContent_BiGrams.txt dokümanları oluşturulur ve bigram kelime grupları bu dokümanlara satır satır yazılır. Her bigram, bir satır olarak dokümana eklenir. Sonuç olarak yukarıda verilen iki paragraf için aşağıdaki gibi iki yeni bigram dokümanı oluşur.



Tablo 5 – Türkçe bigram kelime grupları (TurkishContent_BiGrams.txt)



Tablo 6 – İngilizce bigram kelime grupları (EnglishContent_BiGrams.txt)

Metin içeriklerinden yukarıdaki gibi dillerine göre ayrılmış iki adet bigram kelime grubu elde ettikten sonra, sıradaki işlem ise bu kelime gruplarını karşılaştırıp, doğru eşleştirmeyi bulmaya çalışmak.

Bu işlem için kaynak bigram dokümanından aldığımız her bir satırdaki kelime grubu kelime bazlı olarak bir diziye atılıyor. Sonrada dizideki her bir kelime için sözlükteki karşılıkları bulunur ve bulunan bu karşılıklar bir listeye atılır. Oluşan karşılık listesi sırasıyla hedef bigram dokümanı içinde aranarak, içinde eşleşebileceği kelimeler bulunan bigramları bulur. Bu şekilde yeni ve daha kısa bir bigram listesi oluşturmaya çalışır.

Aynı anda bir bigram kelime grubu hedef listedeki birden fazla bigram kelime grubu ile eşleşebilir. Bu durumda sonuç dokümanına; bulunan bütün sonuçlar ve kaç defa eşleştikleri (parantez içinde) beraber yazılır.

Bir bigram bir kez hedef dokümanda aranmış ve üzerinde işlem yapılmışsa, kaynak bigram içinde tekrar edilse bile, aynı işlemleri tekrar yapması engellenir. Böylece her bir bigram için yalnızca tek sefer işlem yapılır.

Yukarıdaki örnekte kaynak olarak Türkçe, hedef olarak da İngilizce bigram dokümanını alıp satır satır işlem yapıyoruz.

Örneğin ilk Türkçe bigramdan başlarsak; "*evrensel temel*"

Öncelikle kelime grubunu bir diziye atıp, kelime kelime işlem yapmaya başlıyoruz. {'*evrensel'*, '*temel'}* şeklinde iki kelimelik bir dizimiz mevcut. Buradan ilk kelimeyi ('*evrensel'*) alıyoruz. Kaynak dokümanımız Türkçe olduğu için öncelikle zemberek kütüphanesi yardımıyla aşağıdaki özelliklerini kontrol ediyoruz. Eğer kelime Türkçe ise (Zemberek kontrolü ile) aşağıdaki özellikler kontrol edilir;

- Kelime kökü var mı?
- Kök var ise tipi nedir? (İsim, zamir, soru, bağlaç, edat… vb.)
- Kök tipine göre kelimenin çoğul eki alıp almadığı.

Aynı zamanda kelimenin birbirinden farklı birden fazla kökü de olabilir, bu yüzden döngü ile kelime kökleri kontrol edilir ve her bir kök için ayrı işlem yapılır.

İlk ele aldığımız "*evrensel*" kelimesine zemberek ile yakından bakarsak;

```
if (zemberek.kelimeDenetle(_kelime))
{
    foreach (var wordRoot in zemberek.kelimeCozumle(_kelime))
    {
```

wordRoot {{Icerik: evrensel Kok: evren tip:ISIM} Ekler:ISIM_KOK + ISIM_ILISKILI_SEL}

Şekil 3 – Kelime köklerinin döngü ile kontrolü

Name	Value	Type
zemberek.kelimeCozumle("evrensel")	{net.zemberek.yapi.Kelime[1]}	net.zemberek.yapi.Kelime[]
[0]	{{Icerik: evrensel Kok: evren tip:ISIM} Ekler:ISIM_KOK + ISIM_ILISKILI_SEL}	net.zemberek.yapi.Kelime
Static members		
Non-Public members		
Non-Public members		
_ekler	Count = 2	System.Collections.Generic.List<net.zemberek.yapi.ek.Ek>
[0]	{ISIM_KOK}	net.zemberek.yapi.ek.Ek
[1]	{ISIM_ILISKILI_SEL}	net.zemberek.yapi.ek.Ek
Raw View		
_icerik	{evrensel}	net.zemberek.yapi.HarfDizisi
Length	8	int
Non-Public members		
Results View	Expanding the Results View will enumerate the IEnumerable	
_kok	{evren ISIM }	net.zemberek.yapi.Kok
Frekans	184	int
Indeks	6453	int
KisaltmaSonSeslisi	0 '\0'	char
Static members		
Non-Public members		
_icerik	"evren"	string
_tip	ISIM	net.zemberek.yapi.KelimeTipi
asil_Renamed_Field	null	string
frekans	184	int
indeks	6453	int
kisaltmaSonSeslisi	0 '\0'	char
ozelDurumlar	{net.zemberek.yapi.kok.KokOzelDurumu[0]}	net.zemberek.yapi.kok.KokOzelDurumu[]
_tip	ISIM	net.zemberek.yapi.KelimeTipi

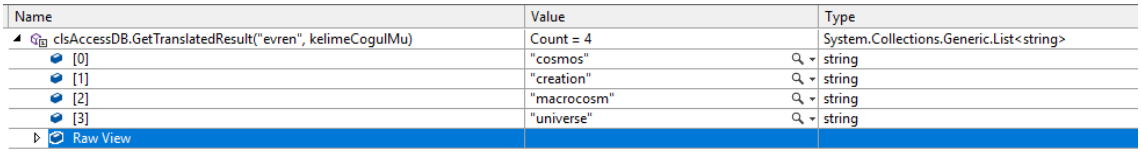
Şekil 4 – Zemberek kütüphanesi ile kelimeyi inceleme

Bu kontrollerden sonra "*evrensel*" kelimesi için zemberek kütüphanesi ile aşağıdaki gibi bir sonuç ortaya çıkar;

Kelime	evrensel
Kelime Türkçe mi?	True
Kelime Kökü Var mı?	True
Kök Sayısı	1
Kök	evren
Tip	ISIM
Kelime Çoğul mu	False
Kelime ile Kelime Kökü Aynı mı?	False

Tablo 7 – Zemberek ile 'evrensel' kelimesi kontrolü

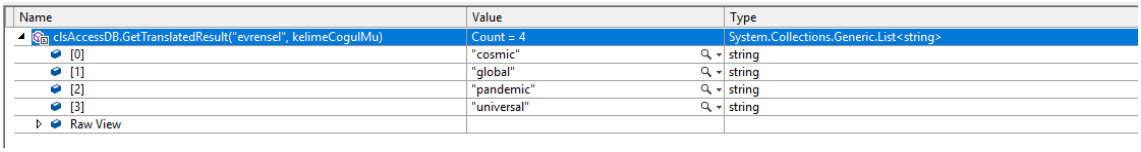
Yukardaki özelliklere göre '*evren'* kökünü Türkçe İngilizce sözlükte arattığımızda aşağıdaki sonuçlar dönecektir.



Name	Value	Type
clsAccessDB.GetTranslatedResult("evren", kelimeCogulMu)	Count = 4	System.Collections.Generic.List<string>
[0]	"cosmos"	string
[1]	"creation"	string
[2]	"macrocosm"	string
[3]	"universe"	string
Raw View		

Şekil 5 – Kelime kökü (‘evren’) için sözlük sorgulama sonucu

Kelime kökü ve kelime birbirinden farklı olduklarından dolayı, ‘*evrensel’* kelimesinin kendisi içinde sözlükten sorgulama yapıp karşılıklarını buluyoruz.



Name	Value	Type
clsAccessDB.GetTranslatedResult("evrensel", kelimeCogulMu)	Count = 4	System.Collections.Generic.List<string>
[0]	"cosmic"	string
[1]	"global"	string
[2]	"pandemic"	string
[3]	"universal"	string
Raw View		

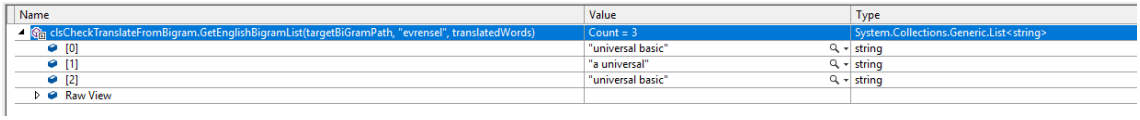
Şekil 6 – Kelime (‘evrensel’) için sözlük sorgulama sonucu

Sözlükten hem kelime, hem de kelimenin kökü için dönen bütün sonuçları tek bir diziye atıp, hedef bigram dokümanı üzerinde sonraki işlemimize bu dizi üzerinden devam ediyoruz. Bu durumda elimizde bigram kelime grubunun ilk kelimesi ‘*evrensel’* için sözlükten gelen aşağıdaki gibi bir karşılık dizisi (translatedWords) oluşur.

Kelime – Kelime Kökü	*‘evrensel’, ‘evren’*
İngilizce Karşılıkları	{*‘cosmos’, ‘creation’, ‘macrocosm’, ‘universe’, ‘cosmic’, ‘global’, ‘pandemic’, ‘universal’* }

Tablo 8 – Kelime (‘evrensel’) Çeviri Karşılıkları Dizisi (translatedWords)

Yukardaki çeviri dizisi (*translatedWords*), hedef bigram dokümanında (Tablo 6 – İngilizce bigram kelime grupları (EnglishContent_BiGrams.txt)) taranarak, içinde çevirilerden herhangi birisinin bulunduğu bigram kelime grupları sorgulanır ve ilk kelime çevirilerinin bulunduğu bigram listesi elde edilir.



Şekil 7 – Hedef bigram dokümanından sorgulama sonucu

```
universal basic
a universal
universal basic
```

Tablo 9 – İlk kelime çevirilerinin bulunduğu bigram listesi

İlk kelime çevirileri için hedef dokümandan ilgili bigram kelime grupları bulunup listelendikten sonra, kaynak bigramın ikinci kelimesi için işleme başlanır.

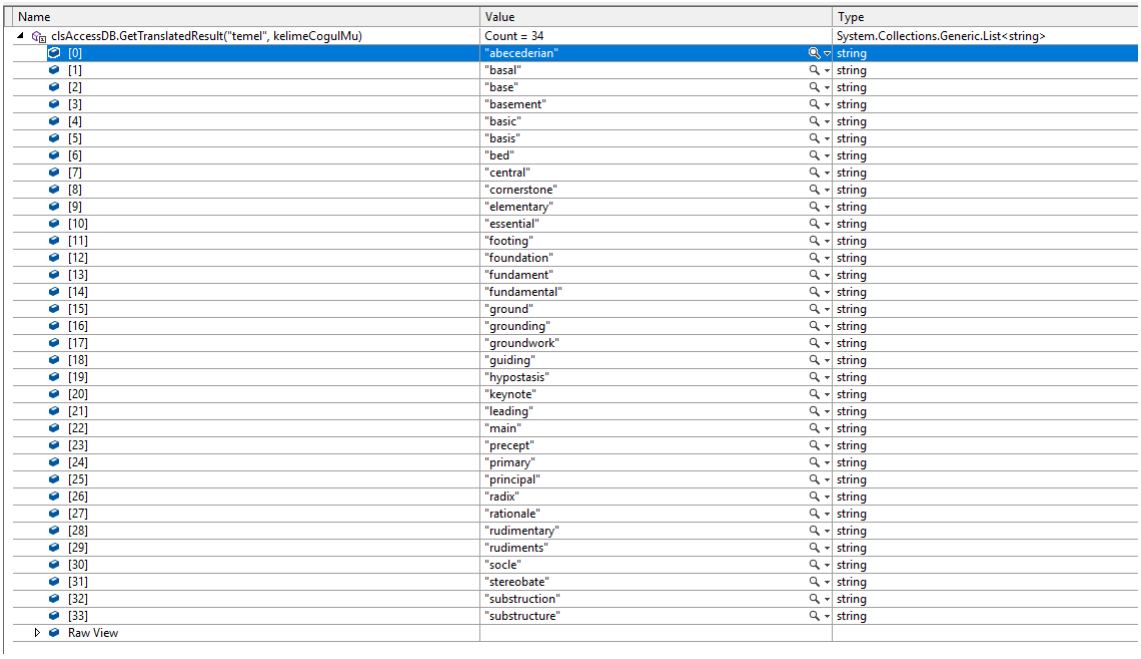
Kaynak bigram kelime grubumuz (‘*evrensel temel’*) ikinci kelimesi ‘*temel’* içinde yine ilk kelimedeki gib; Zemberek kütüphanesi ile gerekli kontroller yapıldıktan sonra sözlükten karşılıkları bulunur.

“*temel*” kelimesi için zemberek kütüphanesi ile aşağıdaki gibi bir sonuç ortaya çıkar;

Kelime	temel
Kelime Türkçe mi?	True
Kelime Kökü Var mı?	True
Kök Sayısı	1
Kök	temel
Tip	ISIM
Kelime Çoğul mu	False
Kelime ile Kelime Kökü Aynı mı?	True

Tablo 10 – Zemberek ile ‘temel’ kelimesi kontrolü

‘temel’ kelimesi için sözlükten dönen karşılıklar dizisi;



Name	Value	Type
clsAccessDB.GetTranslatedResult("temel", kelimeCogulMu)	Count = 34	System.Collections.Generic.List<string>
[0]	"abecederian"	string
[1]	"basal"	string
[2]	"base"	string
[3]	"basement"	string
[4]	"basic"	string
[5]	"basis"	string
[6]	"bed"	string
[7]	"central"	string
[8]	"cornerstone"	string
[9]	"elementary"	string
[10]	"essential"	string
[11]	"footing"	string
[12]	"foundation"	string
[13]	"fundament"	string
[14]	"fundamental"	string
[15]	"ground"	string
[16]	"grounding"	string
[17]	"groundwork"	string
[18]	"guiding"	string
[19]	"hypostasis"	string
[20]	"keynote"	string
[21]	"leading"	string
[22]	"main"	string
[23]	"precept"	string
[24]	"primary"	string
[25]	"principal"	string
[26]	"radix"	string
[27]	"rationale"	string
[28]	"rudimentary"	string
[29]	"rudiments"	string
[30]	"socle"	string
[31]	"stereobate"	string
[32]	"substruction"	string
[33]	"substructure"	string
Raw View		

Şekil 8 – Kelime (‘temel’) için Türkçe İngilizce sözlük sorgulama sonucu

‘temel’ kelimesi için sözlükten 34 adet sonuç bulunur ve elimizde bu kelime için aşağıdaki gibi yeni bir çeviri dizisi (translatedWords) oluşur.

Kelime – Kelime Kökü	*‘temel’*
İngilizce Karşılıkları	*{‘abecederian’, ‘basal’, ‘base’, ‘basement’, ‘basic’, ‘basis’, ‘bed’, ‘central’, ‘cornerstone’, ‘elementray’, ‘essential’, ‘footing’, ‘foundation’, ‘fundament’, ‘fundamental’, ‘ground’, ‘grounding’, ‘groundwork’, ‘guiding’, ‘hypostasis’, ‘keynote’, ‘leading’, ‘main’, ‘precept’, ‘primary’, ‘principal’, ‘radix’, ‘rationale’, ‘rudimentary’, ‘rudiments’, ‘socle’, ‘stereobate’, ‘substruction’, ‘substructure’}*

Tablo 11 – Kelime (‘temel’) Çeviri Karşılıkları Dizisi (translatedWords)

Bulunan ikinci kelime çeviri dizisini, ilk kelimeden farklı olarak, hedef dokümanda değil, ilk kelime için hedef dokümandan sorgulanan yeni bigramlar listesinde (Tablo 9 – İlk kelime çevirilerinin bulunduğu bigram listesi) bulunan bigram kelime gruplarında sorguluyoruz.

Bu şekilde hem ilk hem de ikinci kelimenin aynı anda bulunduğu hedef dokümandaki bigramları, kaynak dokümandaki bigramın karşılıkları olarak gösterebiliriz.

İkinci kelime (‘temel’) için sözlükte 34 adet karşılığı tek tek, Tablo 9’daki ilk kelime çevirilerinin bulunduğu bigram listesinde arattığımızda, bulunan bigramları sonuç dokümanına yazılmaya hazır ve kaynak bigramın tam karşılığı olabilecek bigram kelime gruplarıdır.

Kaynak dokümandaki bütün bigram kelime grupları tamamlanıncaya kadar bu işlemler tekrar eder. Bulunan bütün karşılık bigramları ise bir sonuç dokümanına, ilgili kaynak bigram ile beraber yazılarak insan gücü ile kontrol için hazır hale gelir.

Sonucun doğru ölçülebilmesi için, karşılığı bulunamayan kaynak bigram kelime grupları da sonuç dokümanına yazılır.

Sonuçların Kontrolü

Kaynak bigram kelime grupları ve var ise karşılık gelen hedef bigram kelime grupları ile birlikte bir sonuç dokümanına yazılır.

Bu çalışmada sonuç dokümanı olarak xsl/xslx uzantılı, aşağıdaki başlıkları barındıran Microsoft Excel dokümanı kullanıldı.

Turkish	**English1**	**English2**	**English3**	**English4**	**Result**
evrensel temel	universal basic(2)				

Tablo 12 – Bigram (‘evrensel temel’) için sonuç durumu

Sonuç dokümanında ilk kolon kaynak bigram, sonraki 4 kolon hedef bigramdaki karşılıkları barındırır, parantez içindeki rakam ise bulunan karşılık sayısını belirtir ve sıralamada bu rakama göre büyükten küçüğe doğru olur. En fazla bulunan 4 farklı bigram karşılığı sonuç dokümanına yazılır.

Son kolondaki “**Result**” alanı ise, bulunan karşılıkların doğruluklarının göz ile kontrolü için konulmuştur. Kontrolü yapacak kişi, bulunan değerin doğru olup olmadığına göre sonuca “TRUE” yada “FALSE” olarak işaretler.

Turkish	English1	English2	English3	English4	Result
evrensel temel	universal basic(2)				TRUE
temel gelir	basic income(2)				TRUE
gelir etg					
etg kötü					
kötü fikir	bad idea(1)				TRUE
bunun kısım					
kısım savunucuları					
savunucuları refah					
refah devletini	welfare state(1)				TRUE
devletini abd’de					
abd’de etkin					

Tablo 13 – Uygulama Sonuçlarının Bir Bölümü

Yukarıdaki tabloda kısa bir özet olarak verilen sonuçlara bakarak, verilen örnekteki iki paragraf için bigram yöntemi ile karşılık bulma olasılığını %40 - %50 arasında olduğu, bulunan sonuçlar doğruluk oranlarının ise %95 - %100 oranına yakın olduğu söylenebilir.

Ancak yalnızca bu iki paragraf için ulaşılan sonuç oranları bize bir genelleme yapmak için yeterli değildir. Bunun sebepleri ise sonuç bölümünde ayrıca belirtilmektedir.

BÖLÜM 3. BULGULAR VE YORUMLAR

Çalışmaya başlanıldığında, ilk karşılaşılan temel sorun; açık kaynaklardan yeterince Türkçe İngilizce birebir çeviri bulunamaması ya da bulunanların kısa metinler halinde olmasıydı. Daha detaylı sonuçlar elde etmek için bulunan bu veriler, birleştirilerek araştırmada kullanılmak üzere uzun metinler elde edilmeye çalışıldı.

Sonuç olarak; hem birbiriyle alakasız paragraf ve cümlelerden oluşan çok uzun metinler ile, hem de açık kaynaklardan doğrudan erişilebilecek ancak daha kısa metinler halinde bulunan, Türkçe İngilizce çeviriler üzerinde çalışıldı. Seçilen verilerin cümle bazında birebir çeviri olmasına özen gösterildi.

Bu bölümde; yapılan test ve çalışmalar sonucunda elde edilen örnek bazı bulgular ve bu bulgular üzerinde yapılan yorumlara yer verilmiştir.

Bulgular

Üzerinde çalışılan metinler, bu metinlerdeki bigram sayıları, karşılaştırma sonuçları ve doğruluk oranları ile ilgili bilgiler aşağıda tablo olarak sunulmuştur.

221 Türkçe ve 272 İngilizce kelimeden oluşan bir yüksek lisans tezinin, öz ve abstract bölümleri için sonuç tablosu aşağıdaki gibidir.

Örnek Veri	Kaynak Bigram Sayısı (Türkçe)	Hedef Bigram Sayısı (İngilizce)	Bulunan Karşılık Sayısı	Karşılık Bulma Oranı (%)	Karşılıklardaki Doğruluk Sayısı	Doğruluk Oranı (%)
Tez Öz/Abstract	157	150	12	8	12	100

Tablo 14 – Öz/Abstract Verisi İçin Çeviri Sonuç Tablosu

1610 Türkçe ve 1954 İngilizce kelimeden oluşan bir röportaj için sonuç tablosu

Örnek Veri	Kaynak Bigram Sayısı (Türkçe)	Hedef Bigram Sayısı (İngilizce)	Bulunan Karşılık Sayısı	Karşılık Bulma Oranı (%)	Karşılıklardaki Doğruluk Sayısı	Doğruluk Oranı (%)
Röportaj	1209	1005	110	11	63	57

Tablo 15 – Röportaj Verisi İçin Çeviri Sonuç Tablosu

Birbirinden bağımsız 9136 farklı cümlenin bir araya getirildikten sonra oluşan metnin birebir çevirileri için sonuç tablosu aşağıdaki gibidir. Bu cümlelerde 86584 Türkçe ve 107065 İngilizce kelime bulunmaktadır.

Örnek Veri	Kaynak Bigram Sayısı (Türkçe)	Hedef Bigram Sayısı (İngilizce)	Bulunan Karşılık Sayısı	Karşılık Bulma Oranı (%)	Karşılıklardaki Doğruluk Sayısı	Doğruluk Oranı (%)
Karışık Cümleler	65287	58807	21758	37	18929	87

Tablo 16 – Karışık Cümleler İçin Çeviri Sonuç Tablosu

Yorumlar

Çalışma boyunca elde edilen bütün bulgular değerlendirildiğinde; çeviri işlemi yapılırken kullanılan yöntem, çevirinin aslına uygun bir şekilde bire bir çevrilmesi yada sadece anlam bazlı olarak çevrilmesi, çevrilen metnin uzunluğu, belirli bir konu üzerindeki bir çeviri olması, çeviride kullanılan dilin sadeliği (çok fazla kısaltma kullanılmaması vb.) gibi, çevrilen metnin özellikleri ve çevrilme yöntemi, bigram yönetimiyle ikili kelime gruplarının doğru karşılıklarının bulunması ve istatistiksel olarak hem karşılık değerinin hem de bulunan karşılıkların doğruluk oranının yükselmesinin doğrudan etkilediği görülmüştür

Bulgular bölümündeki örnek çeviri sonuç tablolarına bakıldığında ise, özellikle kullanıcıların kendilerinin çevirdikleri, tez çalışmalarındaki öz abstract bölümlerindeki kelime grupları karşılık oranlarının %10 altında olduğu görülüyor.

Yapılan gözlemlere göre bunun iki temel sebebi var; birincisi ve en önemlisi, çevirinin birebir yapılmayıp, genellikle sadece anlam bazında benzer olmalarına dikkat edilecek şekilde çeviri yapılmış olmasıdır. İkincisi ise, kullanılan algoritmadan kaynaklı olduğu düşünülmektedir. Algoritmanın daha iyi bir sonuç üretebilmesi, eldeki çeviri metinlerini uzunluğu ve birebir çeviri olması ile doğru orantılıdır ve tezlerdeki bu bölümler genelde birer sayfayı geçmeyecek kısa metinlerden oluşmaktadır.

Daha uzun metinlerde ise, özellikle profesyonel çevirmenler tarafından çevrilen makale, köşe yazısı, röportaj vb. içeriklerin çevirilerin de kullanılan dilin sadeliği, çevirinin cümle cümle ya da paragraf bazlı olması, hem kaynak hem hedef metinde

kullanılan kısaltmaların iki tarafta da kullanılması ya da kullanılmaması da sonuç bulma oranında etkili olan sebeplerden bazılarıdır.

Ayrıca karşılık bulmak için kullanılan sözlük, bu sözlüğün kelime dağarcığı da sonuçları etkileyen başka bir faktör olarak karşımıza çıkmaktadır. Özellikle klasik sözlük kullanımı, yani kelime ve kelime kökü ile onaylanmış karşılıkları getiren bir sözlük kullanımı yerine, Google Translate ya da Microsoft Bing gibi yapay zeka ürünü sözlükler kullanmanın, bigramların kelime gruplarının hem karşılıklarını hem de bulunan bu karşılıkların doğruluk oranını etkilediği görülmüştür.

Yapay zekâ ürünü bir çeviri uygulamasını sözlük olarak kullanıldığında, Türkçe dili için kelime kökü ya da ekleri ile uğraşmadan doğrudan bir sonuç bulunup getirilebilmektedir. Ancak bu çeviriler onaylanmış bir karşılık sunmadığı için doğruluğundan emin olunamamakta, aynı zamanda uygulamada bulunan karşılıkların doğruluk oranında da bir düşüş gözlemlenmektedir.. Özellikle Türkçe için klasik ve kelime hazinesi geniş bir sözlük kullanmak hem karşılık bulma oranını hem de bulunan karşılıkların doğruluk oranını pozitif yönde etkiliyor.

Bulgular kısmının son tablosunda da (Tablo 16) görüldüğü gibi, bire bir karşılıklı çevrilmiş uzun metinler üzerinde, algoritmanın başarı oranı yükselmektedir.

Ancak açık kaynaklardaki Türkçe İngilizce çeviri metinlerinin az olması, olanların yeterince uzun olmaması sebebiyle, çok daha farklı metinler üzerinde algoritma test edilememiştir.

BÖLÜM 4. SONUÇ

Bu bölümde, araştırmada ulaşılan temel sonuçlar için bir özet, araştırma sonunda elde edilen bulgular ve araştırma boyunca karşılaşılan problemler dikkate alınarak elde edilen öneriler alt bölümleri yer almaktadır.

Özet

Çalışma boyunca kullanılan metot ve algoritma ile baştan itibaren hedef alınan Türkçe İngilizce kelime gruplarının karşılıklarının elde edilmesi noktasında ilerleme sağlandığı gözlemlenmiştir.

Türkçe'nin dil olarak zorluğu dikkate alındığında, kullanılan bigram yöntemi ile öncelikle çevirilerin ayıklanması, kelime gruplarının elde edilmesi, Türkçe için bu kelime gruplarındaki her bir kelime için zemberek kütüphanesi kullanılarak, kelime köklerinin bulunması, bunların bir sözlük ile karşılıklarının alınıp, İngilizce karşılığında bulunan kelime grupları ile karşılaştırılması ve sonuçta iki taraflı olarak kelime gruplarının doğru karşılıklarının bulunması noktasında, hem bigram yöntemi hem de kurguladığımız algoritma için aşağıdaki gibi iki tespit yapılabilir;

1. Türkçe İngilizce birebir çevirilerde bu yöntem ile bulunan kelime gruplarının başarı oranının %50 den daha düşük olması başarılı bir sonuç olarak değerlendirilemez.
2. Ancak bulunan bu kelime gruplarının karşılıklarının doğruluk oranı ise yaklaşık olarak %90 civarında bulunması, başarılı bir sonuç olarak değerlendirebilir.

Özetle, her ne kadar karşılık bulma oranı düşük gibi görünse de elde edilen karşılıkların doğruluk oranlarının yüksek olduğu görülmüştür.

Öneriler

Mevcut araştırmamızda istatistiki olarak bulunan ikili kelime grup karşılıkları ve bunların doğruluk oranları bu araştırmanın geliştirilebileceğini göstermiştir.

Çalışma aşamasında özellikle Türkçe'nin yapısından (Kemal Oflazer, 2007) (Turhan, 1997) kaynaklı karşılaşılan problemler dikkate alındığında bu uygulamanın geliştirilmesinden önce Türkçe'nin mevcut yapısı üzerinde hala ek çalışmalara ihtiyaç bulunduğu düşünülmektedir. Uygulama içinde kullandığımız Zemberek kütüphanesi

(Zemberek-NLP, 2018) bu konuda yol almış görünmekle beraber daha iyi sonuçlar için bu tarz çalışmaların çoğalması ve/veya kalitesinin artması gerektiği değerlendirilmektedir.

Uygulama aşamasında en çok zorlanılan nokta Türkçe İngilizce birebir çeviri olarak yeterince makale/kaynak bulma zorluğudur. Erişebilecek birebir çeviri metin örneğinin artması bu konudaki çalışmaları da çeşitlendirebilir.

Özellikle çalışmada kullanılan metodun daha fazla kelime barındıran gruplara uygulanabilmesi için algoritmanın geliştirilmesine ihtiyaç bulunmaktadır. Doğruluk oranındaki başarı dikkate alındığında aynı algoritmanın bigram dışındaki farklı n-gram metotları içinde geliştirebileceği düşünülmektedir.

KAYNAKÇA

Ahmet TARCAN, Fahri ÇAKAR. "Bilgiayarlı Dil Tanımlamada Dilbilimsel Yaklaşımlar ve Bir Yazılım Denemesi," **Elektronik Sosyal Bilimler Dergisi**. 7, 26: 64-70, 2008.

Aslan, Erdinç. "Geçmişten Günümüze Çeviri Teknolojileri," *Uluslararası Eğitim ve Sosyal Bilimler Kongresi*: 419-424. Antalya: 2016.

Barachi, Sergio, Oliver Bender, Francisco Casacuberta, Jorge Civera, Elsa Cubel, Shahram Khadivi, Antonio Lagarda, Hermann Ney, Jesus Tomas, Enrique Vidal & Juan-Miguel Vilar. "Statistical Approaches to Computer-Assisted Translation," **Computational Linguistic**. 35, 1, 2007.

Christensen, Tina Paulsen & Anne Schjoldager. "Translation-Memory (TM) Research: What Do We Know and How Do We Know It?," **Hermes – Journal of Language and Communication Studies**. s. 44, 2010.

Chunyu Kit, Yorick Wilks. "The Virtual Corpus Approach to Deriving Ngram Statistics from Large Scale Corpora," *Proceedings of 1998 International Conference on Chinese Information Processing*: 223-229. 1998.

Huang, Guoping , Jiajun Zhang, Yu Zhou & Chengqing Zong. "A New Input Method for Human Translators: Integrating Machine Translation Effectively and Imperceptibly ," *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence* : 1163-1169. 2015.

José B. Mariòo, Rafael E. Banchs, Josep M. Crego, Adrià de Gispert, Patrik Lambert, José A. R. Fonollosa, Marta R. Costa-jussà. "N-gram-based Machine Translation," **Computational Linguistics**. 32, 4: 527-549, 2006.

Jurafsky, Dan. "Language Modelling," **University Of Maryland Department Of Computer Science**. *http://www.cs.umd.edu/class/fall2017/cmsc723/slides/slides_07.pdf*. 2017. (2018).

Kemal Oflazer, İlknur Durgar El-Kahlout. "Exploring Different Representational Unitsin English to Turkish Statistical Machine Translation," *StatMT '07 Proceedings of the Second Workshop on Statistical Machine Translation*: 25-32. Prague: 2007.

Peter F. Brown, John Cocke, Stephen A. Della Pietra, Vincent J. Della Pietra, Fredrick Jelinek, John D. Lafferty, Robert L. Mercer, Paul S. Roossin. "A Statistical Approach To Machine Translation," **Computational Linguistics**. 2, 16: 79-85, 1990.

Peter F. Brown, Peter V. Desouza, Robert L. Mercer, Vincent J. Della Pietra, Jenifer C. Lai. "Class-Based n-gram Models of Natural Language ," **Computational Linguistics**. 18, 4: 467-479, 1992.

Sami Virpioja, Jaakko J. Väyrynen, Mathias Creutz, Markus Sadeniemi. "Morphology-Aware Statistical Machine Translation Based on Morphs Induced in an Unsupervised Manner," *Proceedings of the Machine Translation Summit XI*: 491-498. Copenhagen: 2007.

Tahasildar, Smruti. "Language detection and translation using n-gram and statistical machine translation approach," **International Journal for Research in Engineering Application & Management (IJREAM)**. 3, 1, 2015.

Turhan, Çiğdem Keyder. "An English to Turkish Machine Translation System Using Structural Mapping," *ANLC '97 Proceedings of the fifth conference on Applied natural language processing*: 320-323. 1997.

Zemberek-NLP. "Zemberek Kütüphanesi," *https://github.com/ahmetaa/zemberek-nlp*. 2018.