İngilizce-Türkçe İstatistiksel Makine Çevirisinde Biçimbilim Kullanımı

Onur GÖRGÜN, Olcay Taner YILDIZ

Bilgisayar Mühendisliği Bölümü
IŞIK ÜNİVERSİTESİ
İstanbul, TÜRKİYE

SIU 2012

Sunum Planı

Giriş

- Kaynak dilde verilen cümlenin hedef dildeki en olası karşılığının tespit edilmesi
- Paralel dermeceyi oluşturuan kaynak-hedef dildeki cümleler arasında kelime, kelime grubu veya sözdizilim ağacı hizalama.
- Hizalama sonuçları üzerinden istatistiksel bir çeviri modeli oluşturma ve çeviri olasılıklarının çıkarımı.
- Artı: Düşük seviyeli insan gücü gereksinimi (?).
- Eksi: Sözcük dizilimi sorunsalı!

İstatistiksel Makine Çevirisi

- Kaynak dilde verilen cümlenin hedef dildeki en olası karşılığının tespit edilmesi.

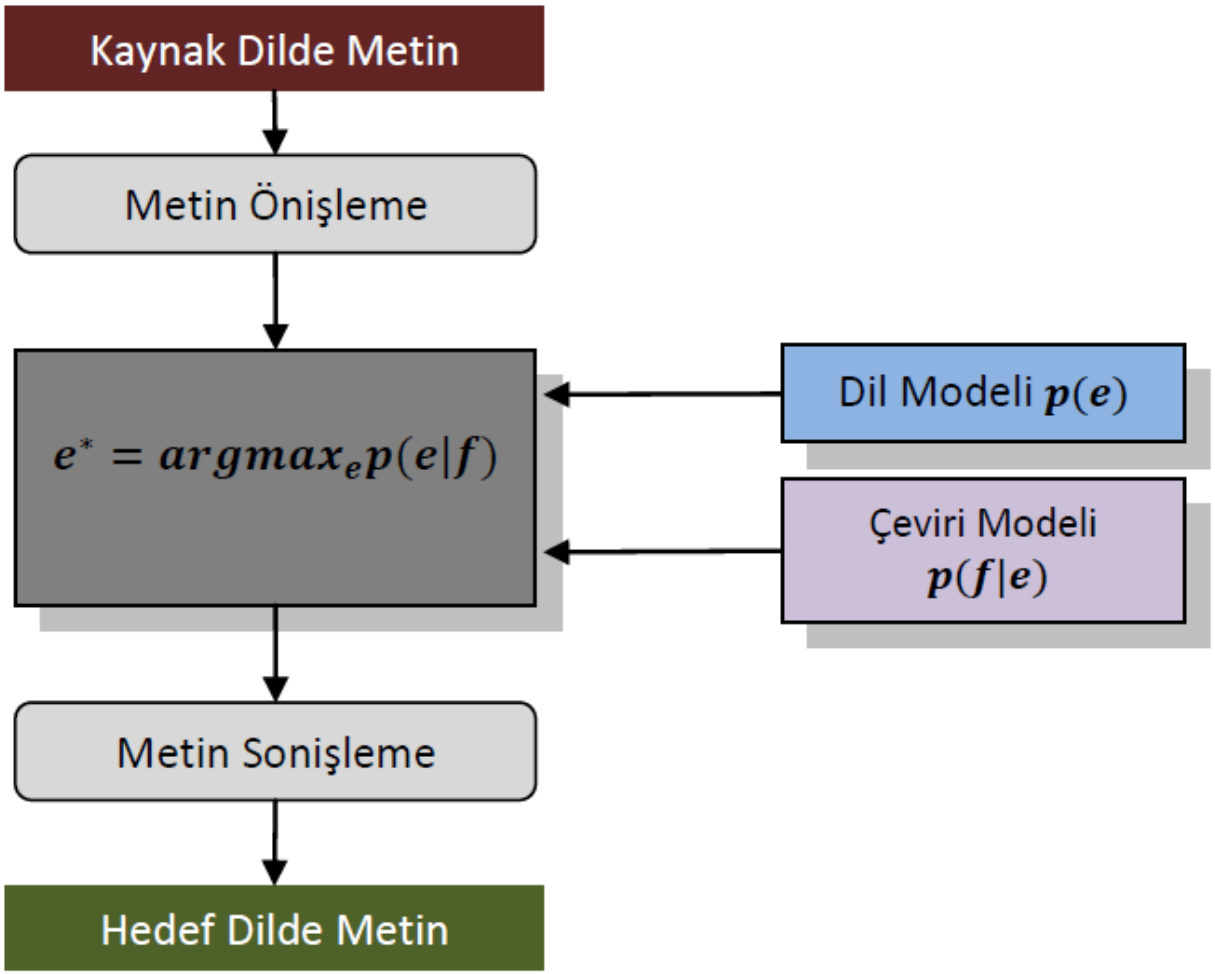
$$p(e|f)$$

$$\hat{e} = \arg\max_{e} \; p(e|f)$$

$$p(e|f) = \frac{p(e)p(f|e)}{p(f)}$$

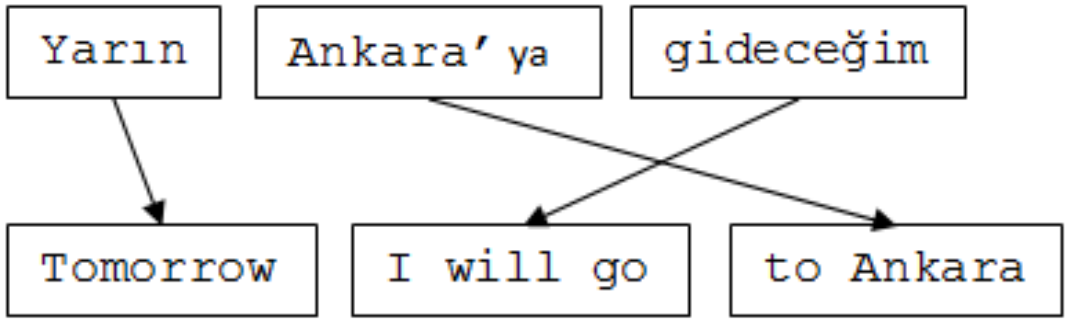
$$\hat{e} = \arg\max_{e} \; p(e)p(f|e)$$

İstatistiksel Makine Çevirisi - Genel Görünüm



Kelime Grubu Tabanlı Model

- Kaynak cümle kelime gruplarına ayrılması.
- Her kelime grubunun hedef dildeki kelime grubu ile eşleştirilmesi.
- Kelime gruplarının sıralarının düzenlenmesi.



- Çeviri modeli her bir öğeyi ayrı bir kelime biçimi olarak kabul eder. → Veri seyrekliği problemi!
  Çözüm: Biçimbilimsel Çözümleme

Dile Özgü Öğelerinin Sisteme Entegrasyonu

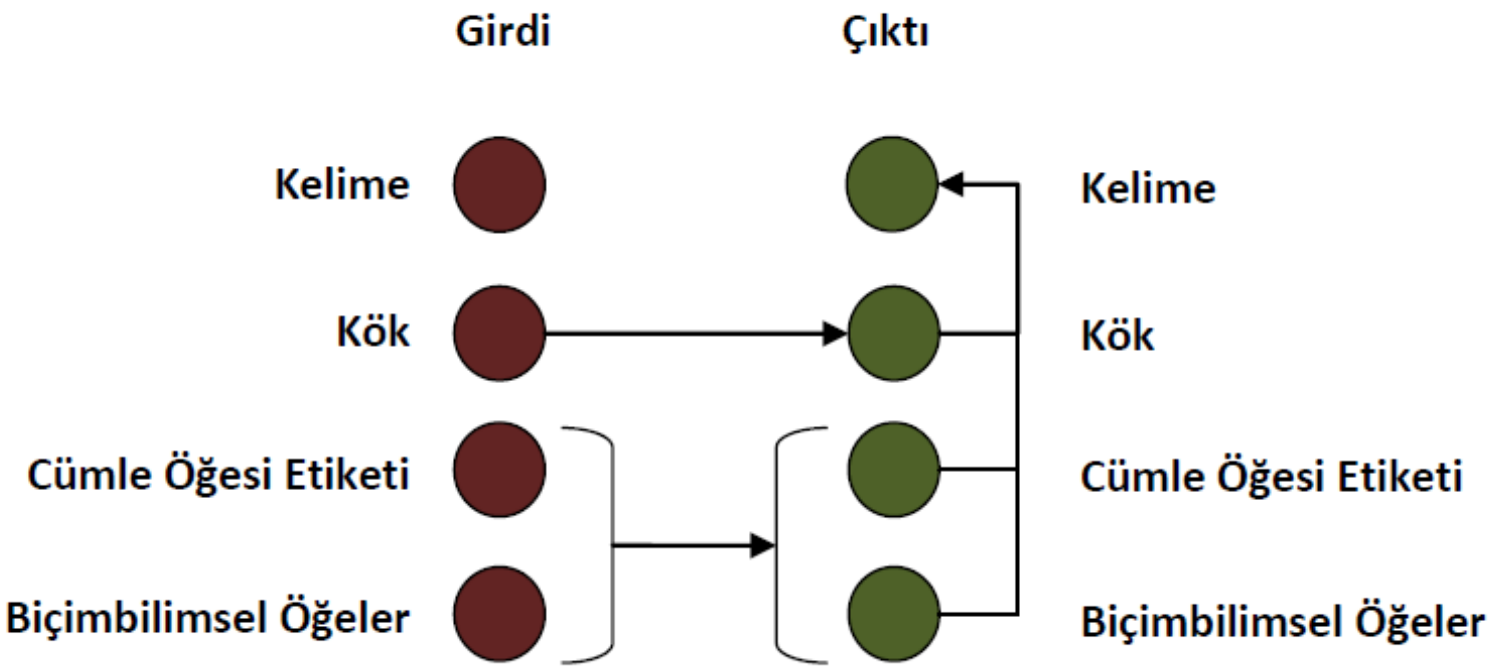


Figure : Faktörlü Çeviri Modeli. Kelime hizalama 3 safhaya ayrılmıştır: kök eşleme, cümle öğesi eşleme ve biçimbilimsel öğelerin eşlenmesi

Dile Özgü Öğelerinin Sisteme Entegrasyonu

Soru:

Kaynak-Hedef dil ikilisi biçimbilimsel olarak farklılık gösteriyorsa?

Table : Türkçe-İngilizce çeviri örneği. Türkçe'ye ait ekler İngilizce kelime veya eklerle eşleşebilir. Tekli çubuklar ek, çiftli çubuklar ise kelime sınırlarını temsil etmektedir.

sonuç	+lAr	+sH	+nA		daya+HnHl	+yarak
conclusion	+s		of	the	basis	on

bir	ortaklık	+sH	oluş	+dHr	+Hl	+yacak	+dHr
a	partnership		draw up	+ed		will	be

Dile Özgü Öğelerinin Sisteme Entegrasyonu

Soru:

Kaynak-Hedef dil ikilisi biçimbilimsel olarak farklılık gösteriyorsa?

Table : Türkçe-İngilizce çeviri örneği. Türkçe'ye ait ekler İngilizce kelime veya eklerle eşleşebilir. Tekli çubuklar ek, çiftli çubuklar ise kelime sınırlarını temsil etmektedir.

sonuç	+lAr	+sH	+nA		daya+HnHl	+yarak
conclusion	+s		of	the	basis	on

bir	ortaklık	+sH	oluş	+dHr	+Hl	+yacak	+dHr
a	partnership		draw up	+ed		will	be

İlgili Çalışmalar

- Faktörlü model ve kelime grubu bazlı modelin birleşimi. (El-Kahlout, 2009)
- Değişik gösterimler kullanılarak bu gösterimlerin çeviri performansına olan katkısı tartışılmaktadır.
- Türkçe
  - Kök + Biçimbilim $\rightarrow$ “bir+lA s+dIr+mA"
  - Kök| ek_1 | ek_2 ... $\rightarrow$ “bir | +lA s | +dIr | +mA"
  - Kök | Biçimbilim $\rightarrow$ “bir | +lA s+dIr+mA"
  - Seçimli Parçalı Model $\rightarrow$ Bazı Türkçe ekler için İngilizce eşleme mümkün olmamaktadır. (örn: “+sH")
- İngilizce:
  - Cümlenin öğeleri etiketleri (POS tags) ve kısıtlı biçimbilim kullanımı.
- BLUE metriği bazından performans artışı.

Hedefler

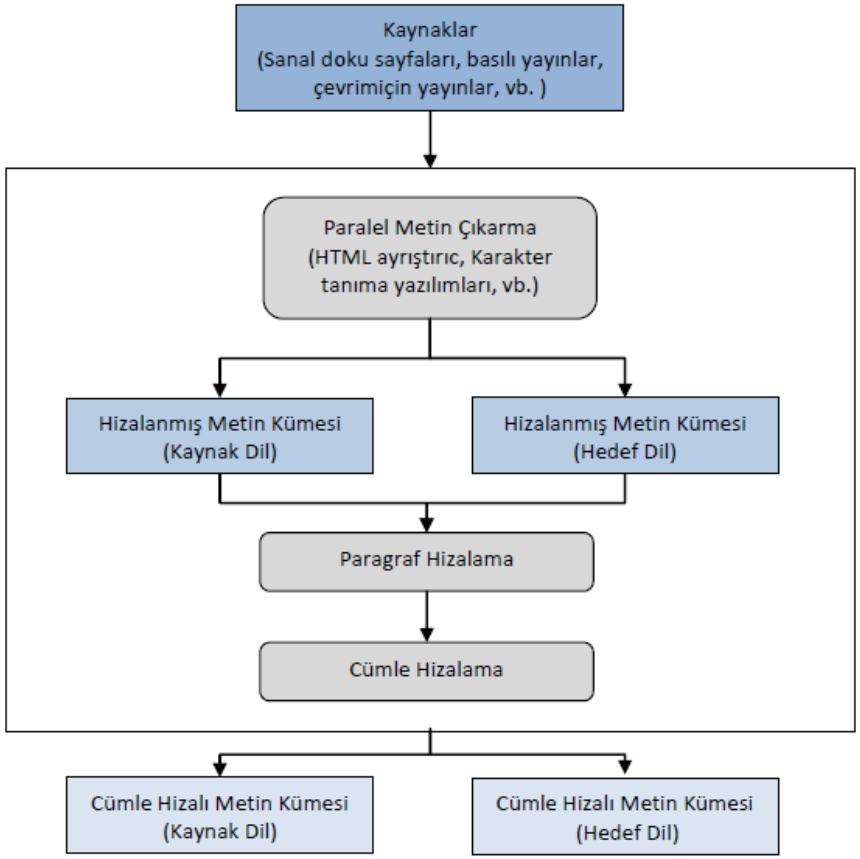
- İngilizce-Türkçe dil çifti için nitelikli paralel metin elde etme.
- Önerilen modellerin performanslarını, kısıtlı uzunlukta bir paralel metin üzerinde detaylı ve kapsamlı bir zengileştirme yapmadan karşılaştırma.
- Sözcük Dizilim Ağacı tabanlı bir çeviri modeli oluşturmak için temel oluşturma.

Biçimbilimsel Gösterimler

Original :	Sistem istemci/sunucu mimarisi üzerine
(Turkish)	kurulmuştur .
Original :	The system is built on client/server
(English)	architecture .
Model-1 :	sistem istemci/sunucu mimari+SH üzeri+Hn+NA
(Turkish)	kurul+YmHş+DHr .
Model-1 :	the+DT system be+VBZ build+VVN on+IN client/server
(English)	architecture .
Model-2 :	sistem istemci/sunucu mimari +SH üzeri +Hn+NA
(Turkish)	kurul +YmHş+DHr .
Model-2 :	the +DT system be +VBZ build +VVN on +IN client/server
(English)	architecture .
Model-3 :	sistem istemci/sunucu mimari +SH üzeri +Hn +NA
(Turkish)	kurul +YmHş +DHr .
Model-3 :	the +DT system be +VBZ build +VVN on +IN client/server
(English)	architecture .

Deney Verisi

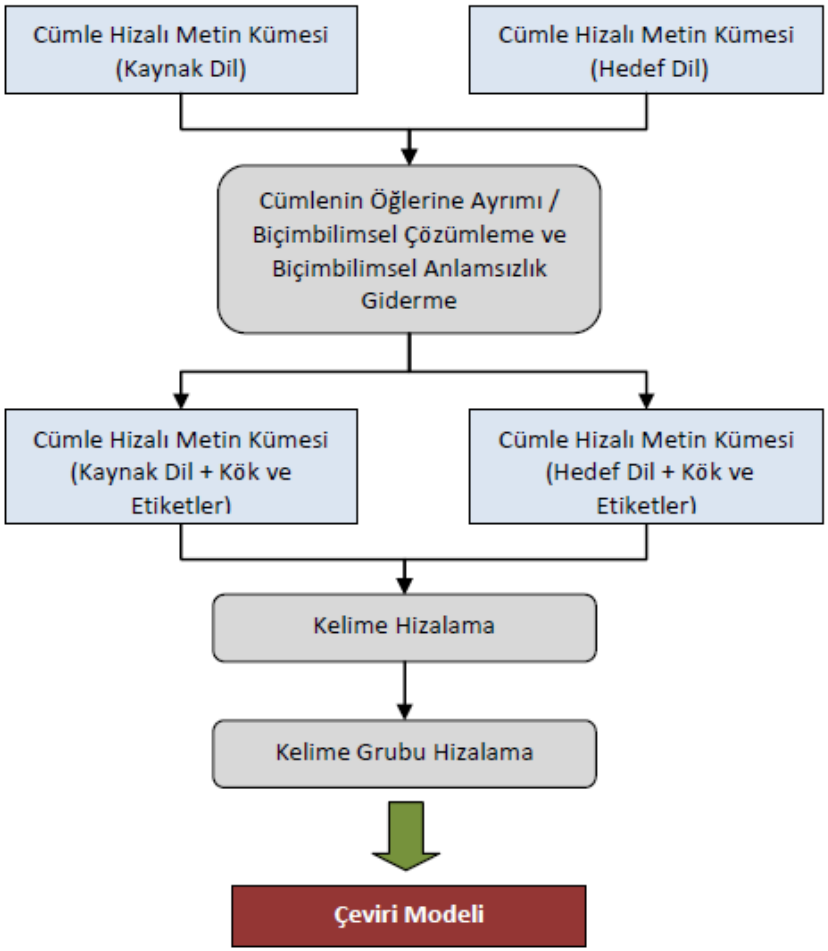
- 2004-2011 tarihleri arasında SIU Özetçe-Abstract çevirileri.
- Cümle Hizalama → Gale&Church algoritması.
- Eğitim kümesi: 3074 cümle ve 80000 kelime. Test Kümesi: 64 cümle ve 1700 kelime.



Deney Düzeneği

- Biçimbilimsel Çözümleme $\rightarrow$ Sak, 2007, Biçimbilimsel Anlamsızlık Giderici $\rightarrow$ Yuret, 2007
- İngilizce Cümlenin Öğelerine ayırma $\rightarrow$ TreeTagger.
- Kelime Hizalama $\rightarrow$ GIZA++, Dil Modelleme $\rightarrow$ SRILM ve IRSTLM.
- Kelime Grubu Hizalama $\rightarrow$ Moses.

Çeviri Modelinin Oluşturulması



Performans Değerlendirme

- **BLEU:** Doğru çeviri sıralamasını bulabilmek için n-gram tabanlı bir çözüm.
- Konum bağımsız bir kelime-hata oranı.

$$BLEU\text{-}n = KC \times exp \sum_{i=1}^{n} \lambda_i \log precision_i \tag{1}$$

$$KC = min\left(1, \frac{cikti\text{-}uzunlugu}{referans\text{-}uzunluk}\right)$$

$$BLEU - 4 = min\left(1, \frac{cikti - uzunlugu}{referans - uzunluk}\right) \prod_{i=1}^{4} precision_i \tag{2}$$

BLEU: Örnek

R1: It is a guide to action that ensures that the military will forever heed Party commands.
R2: It is the Guiding Principle which guarantees the military forces always being under the command of the Party.
R3: It is the practical guide for the army always to heed the directions of the party.
C1: It is to insure the troops forever hearing the activity guidebook that party direct.
C2: It is a guide to action which ensures that the military always obeys the command of the party.

BLEU: Örnek

R1: It is a guide to action that ensures that the military will forever heed Party commands.
R2: It is the Guiding Principle which guarantees the military forces always being under the command of the Party.
R3: It is the practical guide for the army always to heed the directions of the party.
C1: It is to insure the troops forever hearing the activity guidebook that party direct.
C2: It is a guide to action which ensures that the military always obeys the command of the party.

Sonuçlar

	Model-0	Model-1	Model-2	Model-3
BLEU	4.36	4.92	**5.29**	4.41

- Zengileştirilmiş veri seti üzerinde yapılan çalışmalara oranla düşük bir performans (21 BLEU puanı).
- Model-1, Model-0' a göre göreceli olarak %13' lük bir performans artışı göstermiştir.
- Veri boyutunun ve cümle uzunlukluğunun bir sonucu olarak Kelime Hizalama problemi gözlemlenmiştir.
- Model-3, son işlemlerde ek bazında kulalnılan dil modelinin yetersizliğinden ötürü düşük performans göstermiştir.

Gelecek Çalışmalar

- Kısıtlı boyuttaki paralel metin üzerinde zengileştirme çalışmaları (kelime ve kelime grupları ekleme, (El-Kahlout, 2009)).
- Bilişim terimlerinden oluşan bir sözlük ile desteklenmiş ve bir biçimbilim çözümleyici.