İNGİLİZCE-TÜRKÇE İSTATİSTİKSEL MAKİNE ÇEVİRİSİNDE BİÇİMBİLİM KULLANIMI

USING MORPHOLOGY IN ENGLISH-TURKISH STATISTICAL MACHINE TRANSLATION

Onur Görgün

Bilgisayar Mühendisliği Bölümü
Işık Üniversitesi
onurg@isikun.edu.tr

Olcay Taner Yıldız

Bilgisayar Mühendisliği Bölümü
Işık Üniversitesi
olcaytaner@isikun.edu.tr

ÖZETÇE

Bu çalışmada, İngilizce-Türkçe dil ikilisi için biçimbilimsel çözümleme yardımı ile SIU dermecesi üzerinde istatistiksel makine çevirisi denemeleri yapılmıştır. Kelime biçimlerinin baz alındığı çeviri denemeleri İngilizce-Türkçe dil ikilisi gibi biçimbilimsel ve çekimsel olarak birbirinden uzak diller için düşük performans göstermektedir. Bu durumda, çeviri temel birimi olarak kelime formlarının yerine alt-sözcüksel temsiller kullanmak, makine çevirisi performansını önemli ölçüde arttırmaktadır.

ABSTRACT

In this study, statistical machine translation approaches for English-Turkish language couple are offered using morphological analysis in order to achieve different sub-lexical representations on SIU corpus. For language pairs having different inflectional morphologies like English-Turkish language pair, statistical machine translation efforts which use word forms as basic translation unit generally shows low performance. Hence, using different sub-lexical representations instead of word forms improves the machine translation performance.

1. GİRİŞ

Makine Çevirisi (MÇ) alanındaki çalışmaların istatistiksel yaklaşıma dönüşümü IBM' in CANDIDE sisteminin temel kural-tabanlı yaklaşıma olan üstünlüğünün kanıtlanması ile başlamaktadır. Hesaplama gücünün ve buna bağlı olarak paralel dil verisine ulaşımın kolaylaşması araştırmacıların bu alana olan eğilimlerine destek vermiştir. Ancak bu çeviri denemelerinin büyük bir çoğunluğunu kısıtlı sözcük dizimine ve kısıtlı çekimsel biçimbilime sahip dil ikilileri üzerine yapılan çalışmalar oluşturmaktadır. İngilizce-Türkçe dil ikilisi için yapılan çalışmaların azlığı, makine çevirisi probleminin biçimbilimsel olarak farklı özelliklere sahip diller için zorluğu ve yine aynı dil ikilisi için mevcut paralel metinlerin azlığı ile açıklanabilir.

Başarılı bir istatistiksel çeviri modeli oluşturmak için yeterli büyüklükte ve kalitede paralel metin kullanmak gerekmektedir. Çeviri modelinde kullanılacak olan paralel metin, kaynak dildeki cümleler ve bu cümlelere ait hedef dildeki çevirilerden oluşmaktadır. Birçok dil çifti için nitelikli ve büyük boyutlarda paralel metin bulmak mümkünken, İngilizce-Türkçe çifti için paralel metin eksikliği bilinen bir problemdir.

Hem istatistiksel veri seyrekliği problemine çözüm olarak, hem de biçimbilimsel ve çekimsel olarak farklı dil çiftlerinde bire-çok hizalama kalitesini arttırabilmek adına, yapısal olarak güçlü olan dil tarafında alt-sözcüksel gösterimlere ihtiyaç olduğuna dikkat çekilmektedir [1]. Bu gösterimlerin elde edilebilmesi için ise dil çiftinin biçimbilimsel analiz ve biçimbilimsel anlamsızlık giderme işlemlerinden geçirilmesi gerekmektedir. Özellikle Türkçe gibi sondan eklemeli diller için literatürde biçimbilimsel çözümleyiciler [2][3] ve anlamsızlık gidericilerin [4][5][6][7] varlığı hedeflenen gösterimlerin elde edilmesini de mümkün kılmaktadır.

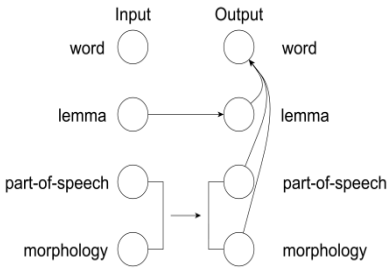
Bu bildiride yapılan çalışma, bu konuda literatürde başarılı bir yaklaşım olan ve Türkçe altsözcüksel ifadelerin çeviri modeline katılması ile başarım oranının yükseleceğini ifade eden çalışmayı [1] kendisine temel alarak, 2004-2010 tarihleri arasındaki SIU konferanslarına ait bildirilerden oluşan bir dil verisi üzerinde çeviri yapmayı hedeflemektedir. Deneyler için elde edilen paralel metin ile bu çalışmanın bir diğer amacı olan İngilizce-Türkçe çifti için bilimsel dile sahip nitelikli bir eğitim kümesi oluşturulmaktır.

Bildirinin genel akışı şu şekildedir. İkinci bölümde, istatistiksel makine çevirisi presipleri anlatılmakta bu alanda yapılmış olan ve İngilizce-Türkçe dil çifti için literatürde kendisine yer bulmuş çalışmalar sunulmaktadır. Üçüncü bölümde, benimsenen yaklaşım temel alınarak önerilen yöntem sunulmakta ve SIU verisi üzerinden örneklenerek açıklanmaktadır. Dördüncü bölümde, deney verisi için yapılan ön çalışmalara ve modellerin oluşturulmasına ilişkin detaylar sunulmuştur. Bildiri sonuçlar ve tartışma bölümü ile son bulmaktadır.

2. İLGİLİ ÇALIŞMALAR

İstatistiksel Makine Çevirisi alanında ilk çalışmalar kelime-tabanlı denemelerdir. Bu çalışmalar, çeviri temel birimi olarak kelimeleri kullanarak, her iki dile ait kelimeler arasında ki en olası eşleşmeleri bulmayı hedeflemektedir [8]. Bu işleme kelime eşleme adı verilmektedir. Ancak bu modeller, özellikle farklı biçimbilimsel özellikler gösteren (İngilizce-Türkçe) diller için bire-çok eşleşmelerde yetersiz kalmaktadır. Özellikle Türkçe' de tek bir kelimenin İngilizce bir kelime grubuna karşılık geldiği durumlarda bu açık bir şekilde görülmektedir. Bu yetersizliği aşmak adına, temel çeviri

Bu çalışma, Işık Üniversitesi **10A303** no' lu Bilimsel Araştırma Projesi kapsamında desteklenmektedir.
978-1-4673-0056-8/12/$26.00 ©2012 IEEE



Şekil 1. Faktörlü çeviri yaklaşımı.

biriminin değiştirilmesi gerekliliği duyulmuş ve araştırma çabaları kelime grubu tabanlı çeviriye yöneltilmiştir.

Gerek kelime gerekse kelime grubu tabanlı bir çalışma olsun, istatistiksel makine çevirisinde iki konuya özen gösterilmektedir: uygunluk ve akıcılık. Uygunluk ölçütü eşleştirme sonucunda sağlanırken, akıcılık için hedef dile ait *n*-gram temelli bir dil modeline ihtiyaç duyulur. Çeviri temel biriminden bağımsız olarak sistem aşağıdaki gibi modellenebilir:

$$t_{best} = argmax_t p(t|s) \quad (1)$$

Bayes kuramı yardımı ile en olası çeviri,

$$t_{best} = argmax_t p(s|t).P_{LM}(t) \quad (2)$$

şeklinde ifade edilebilir. Denklem 1 ve 2' de t_{best} en olası çeviri, 's' cümlenin kaynak dildeki, 't' ise hedef dildeki karşılığını ifade etmektedir. P_{LM} ise kaynak dile ait dil modelini ifade etmektedir.

İngilizce-Türkçe dil ikilisi için yapılan çalışmalar 1981 tarihine dayanmaktadır [9]. Yine aynı dil çifti için ilk yapısal eşleştirme denemesi [10] ve kural-tabanlı yaklaşım da [11] bu çalışmayı takiben önerilmiştir. Önerilen bu sistemlerin ortak noktası yoğun bir kural oluşturma yöntemini benimsemeleri ve kısıtlı bir alana yönelik olmalarıdır.

Kelime grubu tabanlı modeller makine çevirisi konusunda en popüler çeviri yaklaşımlarıdır ve İngilizce-Türkçe çifti için de başarılı denemeler bu alanda olmuştur [1][12]. Veri seyrekliği probleminin çözümü için, biçimbilimsel olarak zengin olan Türkçe tarafı için biçimbilimsel çözümleme teknikleri kullanılmış ve yine bu sebeple sesteş ekler sözlüksel ifadeler şeklinde ifade edilmiştir. İngilizce tarafında ise dilin yapısı gereği kısıtlı bir çözümleme yapılmıştır. Temel olarak bu modeller literatürde faktörlü yaklaşımlar olarak sınıflandırılmış (Şekil 1) olup, hem biçimbilimsel öğelerin hem de kelime köklerinin ayrı olarak eşleştirilmesini gerektirmektedir. Ancak araştırmacılar bu yöntemin bütün kurallarını takip etmemişlerdir.

Deneysel çalışmalar göstermektedir ki, yapısal olarak fakir-zengin dil çiftleri için faktörlü yaklaşımlar düşük başarım sergilemektedir. Bu nedenle, söz konusu çalışmalarda ekler veya ek grupları ayrı kelimeler olarak değerlendirmiş, kelime ve kelime grubu eşlemeleri bu şekilde yapılmıştır. Bu sayede Türkçe ek ve ek gruplarının İngilizce ek veya kelimelerle eşlenmesi sağlanmaktadır. Ancak kelime biçimi elde edilirken ek bazında çalışan bir başka dil modeline ihtiyaç duyulmaktadır. Ek-temelli dil modeli, kök ve ekler halinde ifade edilen çevirinin kelime biçimine çevirimi için gereklidir.

3. KULLANILAN YÖNTEM

Giriş bölümünde bahsedildiği üzere, bu çalışma da daha önce önerilmiş olan sözdizimsel yaklaşımla zenginleştirilmiş sözcük grubu tabanlı çalışma [1] temel alınmaktadır. Bu bölümde çalışmamızda kullandığımız bu temel modellere ait detaylar aktarılmakta, kendi işlem ve önişlem detayları ile birlikte sunulmaktadır.

Bu çalışmada temel olarak 4 farklı gösterim kullanılmakta ve bu gösterimlerle oluşturulan çeviri modellerinin performans değerlendirmeleri yapılmaktadır. Ön işlem olarak kelime biçimlerine, çevirinin Türkçe tarafı için biçimbilimsel çözümleme ve biçimbilimsel belirsizlik giderme, İngilizce tarafı içinse cümlenin öğelerinin bulunması işlemi uygulanmıştır. Ancak sözdizimsel olarak katkı sağlamayan etiketler (NN, isim etiketi) eğitim kümesine dâhil edilmemiştir. (Orjinal çalışma İngilizce tarafı için kısıtlı bir biçimbilimsel çözümleme de kullanmaktadır.)

- **Gösterim 1:** Kelime**,** çözümleme yapılmadan kelime biçimi olarak sunulmaktadır.
- **Gösterim 2:** Kelime kökü ve biçimbilimsel çözümleme elemanları ile birlikte kelime olarak sunulmaktadır. ("bir+lAş+DHr+mA")
- **Gösterim 3:** Kelime kökü ayrı, biçimbilimsel çözümleme elemaları eklenmiş bir şekilde sunulmaktadır. ("bir | +lAş+DHr+mA")
- **Gösterim 4:** Kelime kökü ve biçimbilimsel çözümleme elemanları ayrı kelimeler olarak sunulmaktadır. ("bir |+lAş | +DHr | +mA")

Bu gösterimlere ait cümle örnekleri, SIU verisi kullanılarak Tablo 1' de sunulmuştur.

Açıklanmış olan gösterimlerle 4 farklı eğitim kümesi oluşturulmuş ve bu eğitim kümeleri kullanılarak 4 farklı çeviri modeli elde edilmiştir. Yine bu gösterimlerle oluşturulmuş test kümeleri ile başarımlar hesaplanmıştır.

Tablo 1 Makine çevirisi için biçimbilimsel gösterimler: sadece kelime biçimi gösterimi, kök+biçimbilimsel çözümleme gösterimi, kök ve biçimbilimsel çözümleme gösterimi, kök ve biçimbilimsel çözümleme elemanları gösterimi

Gösterim 1: (Türkçe)	Sistem istemci/sunucu mimarisi üzerine kurulmuştur.
Gösterim 1: (İngilizce)	The system is built on client/server architecture .
Gösterim 2: (Türkçe)	sistem istemci/sunucu mimari+SH üzeri+Hn+NA kurul+YmHş+DHr .
Gösterim 2: (İngilizce)	the+DT system be+VBZ build+VVN on+IN client/server architecture .
Gösterim 3: (Türkçe)	sistem istemci/sunucu mimari +SH üzeri +Hn+NA kurul +YmHş+DHr .
Gösterim 3: (İngilizce)	the +DT system be +VBZ build +VVN on +IN client/server architecture.
Gösterim 4: (Türkçe)	sistem istemci/sunucu mimari +SH üzeri +Hn +NA kurul +YmHş +DHr .
Gösterim 4: (İngilizce)	the +DT system be +VBZ build +VVN on +IN client/server architecture.

4. DENEYLER

4.1. Deney Verisi ve Düzeneği

Bu çalışmada, IEEE' de yayınlanan PDF formatındaki bildiriler toplanmış ve PDF dokümanlarından metin çıkarımı işlemi uygulanmıştır. Dil kodlama problemi olan ve her iki dilde özetçesi olmayan dokümanlar elendiğinde elde kalan doküman sayısı 634 olmaktadır. Bu dokümanlar üzerinde Microsoft Proofing Tools ile yazım denetimi işlemi uygulanmıştır. Dokümanların makine çevirisinde kullanılabilmesi için paragraf ve cümle bazında hizalanmış olması gerekmektedir. Bildirilere ait özetçeler genel olarak tek paragraftan oluştuğundan paragraf hizalama işlemi yapılmamıştır. Cümle hizalama işlemi için cümle karakter uzunluğunu temel alan Church&Gale [13] 'e ait cümle hizalama algoritması kullanılmıştır. Algoritmanın hata yaptığı hizalamalar el ile düzeltilerek, hizalama doğruluğu arttırılmıştır. Hizalama işlemleri şematik olarak Şekil 2' de sunulmuştur.

Hizalanmış cümleler, üçüncü bölümde belirtilen gösterime ulaşmak adına dil işleme için gerekli önişlemlere tabi tutulmuştur. Türkçe cümleler için Oflazer' in iki-seviyeli modeli [2] üzerine kurulmuş olan biçimbilimsel çözümleyici, biçimbilimsel belirsizlik giderme işlemi için de kural tabanlı bir anlamsızlık giderici [5] kullanılmıştır. İngilizce tarafı için sadece TreeTagger [14] yazılımı kullanılmış ve biçimbilimsel olarak anlam ifade etmeyen etiketler çıkarılmıştır. Bir önceki bölümde ifade edilen gösterimler kullanılarak, bu gösterimleri karşılayan ve 3075 cümleden oluşan 4 farklı veri kümesi elde edilmiştir.

Eğitim kümesi oluşturulduktan sonra, kelime hizalama ve kelime grubu hizalama işlemleri gerçekleştirilmiştir. Kelime hizalama için GIZA++ [15] ve MKCLS yazılımları [16], kelime grubu temelli çeviri modeli oluşturmak için Moses [17] makine çevirisi yazılımı kullanılmıştır. Türkçe' ye ait dil modeli oluşturmak için El-Kahlout tarafından sunulmakta olan Türkçe dil modeli eğitim kümesi kullanılmıştır. Uygulanan ön işlemler serisi şematik olarak Şekil 3' te sunulmuştur.

Test kümesi olarak, 2011 yılı SIU bildirilerinin bir altkümesi kullanılmıştır. Test kümesine ait cümleler de aynı eğitim kümesinde olduğu gibi biçimbilimsel çözümleme, biçimbilimsel anlamsızlık giderme işlemlerinden geçirilmiştir. Her bir gösterim tarzı için farklı bir test kümesi oluşturulmuş olup, bu test kümeleri ile deneyler gerçekleştirilmiştir.

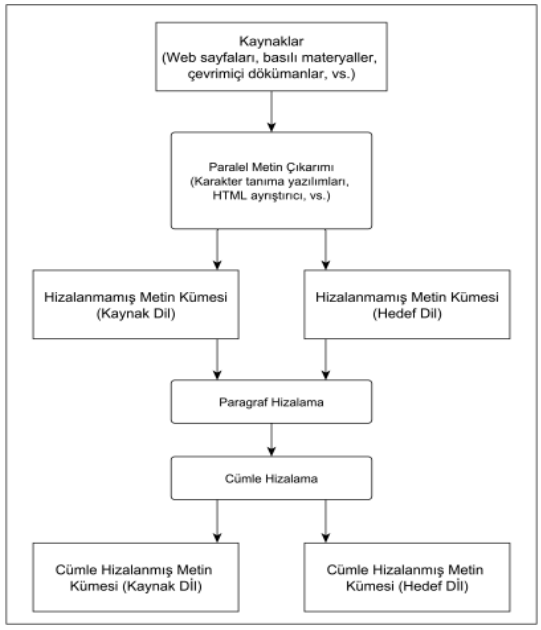
Başarım ölçütü olarak konum bağımsız bir kelime hata oranı (KHO) türevi olan BLEU metriği kullanılmaktadır. KHO' dan farklı olarak, temel ölçüm birimi olarak kelimeleri baz almayan BLEU, değişik uzunluklardaki *n*-gram öbeklerini kullanır. BLEU metriği:

$$BLEU_n = BP \times exp \sum_{i=1}^{n} \lambda_i \log precision_i \qquad (3)$$

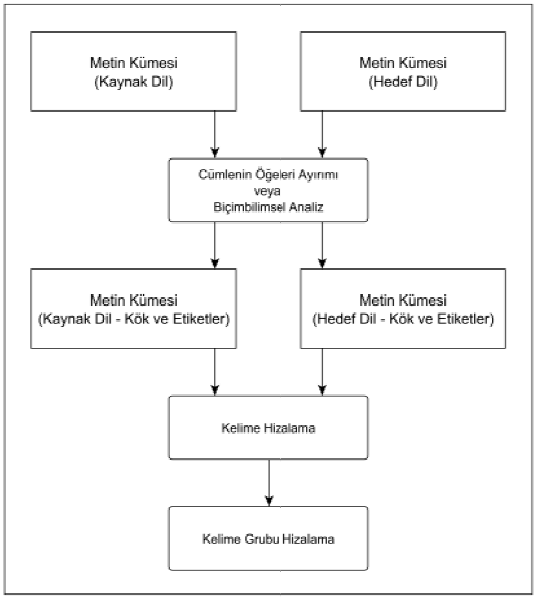
BP (Brevity Penalty) hedef çevirideki kelime düşmeleri nedeni ile oluşabilecek yanıltıcı yüksek skorları engellemek amacı ile belirlenmiş bir ceza katsayısıdır:

$$BP = min\left(1, \frac{output - length}{reference - length}\right) \qquad (4)$$

Söz konusu başarım ölçümleri için bir veya birden fazla referans çeviri kullanılmaktadır. Çeviri adayı bu referans



Şekil 2 Paralel metin çıkarımı ile başlayan ve cümle hizalama işlemi ile son bulan önişlemlerim şematik gösterimi.



Şekil 3 Kelime grubu hizalama ve çıkarımı için yapılan önişleme sürecinin şematik olarak gösterimi.

cümlelere olan *n*-gram yakınlıkları açısından değerlendirilerek ve 0-100 arasında derecelendirilir.

4.2. Deney Sonuçları

Deney sonuçları Tablo 2' de gösterilmiştir. Sonuçlar göstermektedir ki, sadece kelime biçimlerinin kullanıldığı gösterim en düşük performansı kaydetmiştir. Gösterim 3, kelime biçimlerine göre göreceli %21' lik bir performans artışı

sağlamıştır. Eğitim kümesinin küçüklüğü ve buna bağlı olarak, ek bazında hizalama sonuçlarının başarısızlığı nedeni ile Gösterim 4 performans artışına beklenilen katkıyı gösterememiştir. Bu yetersizliğin bir diğer nedeni de ek bazında çalışan bir dil modeli kullanılmamasıdır.

Aynı test kümeleri kullanılarak Google Translate ile çeviri yapılmıştır. Bu çeviri sonucunda İngilizce-Türkçe yönünde 14.71 BLEU puanı elde edilmiştir.

Tablo 2 Aday çevirilerin ve Google Translate çevirisinin BLEU metriği kullanılarak hesaplanan başarımları.

Gösterim				Google Translate
1	**2**	**3**	**4**	
4.36	4.92	**5.29**	4.41	**14.71**

5. SONUÇLAR VE TARTIŞMA

Deneyler sonunda karşılaşılan düşük başarım SIU dermecesinin yapısı ile ilintilidir. Seçilen eğitim kümesinde çeşitlilik bulunduğundan kelime hizalama performansı düşüktür. Bir diğer çıkarım ise biçimbilimsel çözümlemenin istenilen katkıyı yapamamasıdır. Biçimbilimsel çözümleme çözümlenecek kelimenin kökü tespit edilerek ve olası ek dizilimleri çıkartılarak yapılmaktadır. SIU dermecesinde yer alan kelimeler ise çözümleyici kök veritabanı için yeni kelimeler olup, biçimbilimsel çözümlemesi yapılamamaktadır. Bu durum kelime-ek hizalaması başarımını etkilediği gibi, kaliteli bir ek-tabanlı dil modelinin oluşturulmasını zorlaştırmaktadır. Bu kelimelere ait çözümlemelerin yapılabilmesi terimler sözlüğü oluşturulması ve biçimbilimsel çözümleme kuralları ile başarılabilir. Ayrıca, kelime ve kelime grubu hizalama işlemlerinde yüksek başarım yakalayabilmek için mevcut veri kümesinin kelime ve kelime grupları eşleşmeleri ile zenginleştirilmesi gerekmektedir. Devam eden çalışmaların zenginleştirme konusu üzerine olması planlanmaktadır.

KAYNAKÇA

[1] El-Kahlout, İ. D.: Exploring Different Representational Units in English-to-Turkish Statistical Machine Translation, In: Proceedings of the Second Workshop on Statistical Machine Translation, pp. 25-32 (2007).

[2] Oflazer, K.: Two-level Description of Turkish Morphology. Literary and Linguistic Computing 9, pp.137-148 (1994).

[3] Hakkani-Tür, D. Z., Oflazer, K.,Tür, G.: Statistical Morphological Disambiguation for Agglutinative Languages. In: Computers and the Humanities 36(4), pp.381-410 (2002).

[4] Sak, H., Güngör, T., Saraçlar, M.: Turkish Language Resources: Morphological Parser, Morphological Disambiguator and Web Corpus. In: GoTAL 2008, vol.5221 of LNCS, Springer, pp.417-427 (2008).

[5] Yüret, D., Türe, F.: Learning Morphological Disambiguation Rules for Turkish. In: Proceedings of HLT-NAACL, pp.328-334 (2006).

[6] Görgün, O., Yıldız, O. T.: A Novel Approach to Morphological Disambiguation for Turkish. In: Proceedings of International Symposium on Computer and Information Sciences (ISCIS), pp.77-83 (2011).

[7] Sak, H., Güngör, T., Saraçlar, M.: Morphological Disambiguation of Turkish Text with Perceptron Algorithm. In: Gelbukh, A. (ed.) CICLING 2007, LNCS 4394, pp.107-118 (2007).

[8] Hutchinson, J. : The Georgetown-IBM Demonstration. MT News International, no.8, pp.15-18 (1994).

[9] Sagay, Z.: A Computer Translation from English to Turkish: Masters Thesis, METU, Department of Computer Engineering (1981).

[10] Keyder Turhan, C.: An English to Turkish Machine Translation System Using Structural Mapping. In: Proceedings of the Applied Natural Language Processing, Washington, DC, p.320-323 (1997).

[11] Hakkani, D. Z., Tür, G., Oflazer, K., Mitamura, T., Nyberg, E.: An English-to-Turkish Interlingual MT System. In: AMTA, pp.83-94 (1998).

[12] Yeniterzi, R., Oflazer, K.: Syntax-to-Morphology Mapping in Factored Phrase-based Statistical Machine Translation from English to Turkish. In: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL), pp.454-464 (2010).

[13] Gale, W.A., Church, K. W.: A program for aligning sentences in bilingual corpora. Computational Linguistics, 19(1), pp.75-102 (1993).

[14] Schmid, H. : Probabilistic part-of-speech tagging using decision trees. In: Proceedings of International Conference on New Methods in Language Processing (1994).

[15] Och, F. J., Ney, H.: A systematic comparison of various statistical alignment models, *Computational Linguistics*, Vol. 29, No.1, pp. 19-51 (2003).

[16] Och, F. J.: An Efficient Method for Determining Bilingual Word Classes. In: Ninth Conf. of the Europ. Chapter of the Association for Computational Linguistics, pp. 71-76 (1999).

[17] Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Zens, R., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Dyer, C., Bojar, O., Herbst, E., Moses: Open Source Toolkit for Statistical Machine Translation. In: Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics, Companion Volume, pp.177-180 (2007).