ÇİFT DİLLİ KELİME TEMSİLLERİ İLE SÖZLÜK EŞLENMESİ

EVALUATING BILINGUAL EMBEDDINGS IN BILINGUAL DICTIONARY ALIGNMENT

Yiğit Sever

Dr. Gönenç Ercan

Tez Danışmanı

Hacettepe Üniversitesi

Lisansüstü Egitim-Öğretim ve Sınav Yönetmeliğinin

Bilgisayar Mühendisliği Anabilim Dalı için Öngördüğü

YÜKSEK LİSANS TEZİ olarak hazırlanmıştır.

2019

This work titled "**Evaluating Bilingual Embeddings in Bilingual Dictionary Alignment**" by **Yiğit Sever** has been approved as a thesis for the Degree of Master of Science in Computer Engineering by the Examining Committee Members mentioned below.

Prof. Dr. İlyas Çiçekli

Head ..……

Dr. Gönenç Ercan

Advisor ..……

Dr. Tayfun Küçükyılmaz

Member ....……..…..

Dr. Mehmet Köseoğlu

Member ……..

Dr. Burcu Can

Member ……..……

This thesis has been approved as a thesis for the Degree of Master of Science in Computer Engineering by Board of Directors of the Institute of Graduate School of Science and Engineering on ..... / ..... /.....

Prof. Dr. Menemşe GÜMÜŞDERELİOĞLU

Director of the Institute of

Graduate School of Science and Engineering

Aileme...

ETİK

Hacettepe Üniversitesi Fen Bilimleri Enstitüsü, tez yazım kurallarına uygun olarak hazırladığım bu tez çalışmasında,

- tez içindeki bütün bilgi ve belgeleri akademik kurallar çerçevesinde elde ettiğimi,
- görsel, işitsel ve yazılı tüm bilgi ve sonuçları bilimsel ahlak kurallarına uygun olarak sunduğumu,
- başkalarının eserlerinden yararlanılması durumunda ilgili eserlere bilimsel normlara uygun olarak atıfta bulunduğumu,
- atıfta bulunduğum eserlerin tümünü kaynak olarak gösterdiğimi,
- kullanılan verilerde herhangi bir değişiklik yapmadığımı,
- ve bu tezin herhangi bir bölümünü bu üniversite veya başka bir üniversitede başka bir tez çalışması olarak sunmadığımı

beyan ederim.

02 / 09 / 2019

YİĞİT SEVER

YAYINLANMA FİKRİ MÜLKİYET HAKLARI BEYANI

Enstitü tarafından onaylanan lisansüstü tezimin/raporumun tamamını veya herhangi bir kısmını, basılı (kağıt) ve elektronik formatta arşivleme ve aşağıda verilen koşullarla kullanıma açma iznini Hacettepe üniversitesine verdiğimi bildiririm. Bu izinle Üniversiteye verilen kullanım hakları dışındaki tüm fikri mülkiyet haklarım bende kalacak, tezimin tamamının ya da bir bölümünün gelecekteki çalışmalarda (makale, kitap, lisans ve patent vb.) kullanım hakları bana ait olacaktır.

Tezin kendi orijinal çalışmam olduğunu, başkalarının haklarını ihlal etmediğimi ve tezimin tek yetkili sahibi olduğumu beyan ve taahhüt ederim. Tezimde yer alan telif hakkı bulunan ve sahiplerinden yazılı izin alınarak kullanması zorunlu metinlerin yazılı izin alarak kullandığımı ve istenildiğinde suretlerini Üniversiteye teslim etmeyi taahhüt ederim.

Yükseköğretim Kurulu tarafından yayınlanan "***Lisansüstü Tezlerin Elektronik Ortamda Toplanması, Düzenlenmesi ve Erişime Açılmasına İlişkin Yönerge***" kapsamında tezim aşağıda belirtilen koşullar haricince YÖK Ulusal Tez Merkezi / H. Ü. Kütüphaneleri Açık Erişim Sisteminde erişime açılır.

□ Enstitü / Fakülte yönetim kurulu kararı ile tezimin erişime açılması mezuniyet tarihimden itibaren 2 yıl ertelenmiştir.

□ Enstitü / Fakülte yönetim kurulu gerekçeli kararı ile tezimin erişime açılması mezuniyet tarihimden itibaren .... ay ertelenmiştir.

□ Tezim ile ilgili gizlilik kararı verilmiştir.

..... / ..... /.....

YİĞİT SEVER

ÖZET

ÇİFT DİLLİ KELİME TEMSİLLERİ İLE SÖZLÜK EŞLENMESİ

Yiğit Sever

Yüksek Lisans, Bilgisayar Mühendisliği

Tez Danışmanı: Dr. Gönenç Ercan

September 2019, 97 sayfa

Sözlükler bir dilin kelime haznesini anlamlar açıdan anlatır ve dosyalar. WordNet, bunun üzerine anlamlar arası alt-üst ilişkilerini de tanımlar. Bilgisayar bilimi üzerine yapılan araştırmalarda elle derlenmiş kaynak WordNet özellikle metin özetleme ve makine çevirisi alanında kullanılmaktadır. Asıl WordNet İngilizce için hazırlanmış olup diğer dillerdeki karşılıkları kapsamlı ya da erişilebilir olmayabilir. İngilizce dışındaki dilleri esas alan çalışmaların WordNet'ten yararlanabilmesi adına makine yardımlı derleme ve değerlendirme yöntemleri esastır.

Kelime temsilleri bir dilin söz dağarcığını çok boyutlu bir uzaydaki noktalar, bununla birlikte vektörler olarak gösterir. Bu vektörleri kullanarak belgeleri matematiksel olarak tanımlamak ya da belgeler arası geometrik bağıntılar kurmak şimdinin çalışılan konularındandır. Bu çalışmaya bir sözcüğün sözlük tanımının onun bağlamsal yapısını temsil edebileceğini varsayarak başladık. Kelime temsilleri ile sözlük tanımlarını çok boyutlu bir uzayda gösterdik. Bu soyut uzaylar birden fazla dilin söz dağarcığına ev sahipliği yapmak adına eşlenebilir. Belirli anlamların diller arası erişimi ve eşlenmesi sorununa güdümlü ve güdümsüz öğrenme yöntemleri ile çözüm getirmeye çalışılmıştır. Var olan veri boyutunun önemini ve kimi yöntemlerin bu konuda zayıf başarı gösterdiğini keşfettik.

Anahtar Kelimeler: kelime temsilleri, sözlük eşleme, bağlamsal öğrenme, kısa metin benzerliği

ABSTRACT

EVALUATING BILINGUAL EMBEDDINGS IN BILINGUAL DICTIONARY ALIGNMENT

Yiğit Sever

Master of Science, Department of Computer Engineering

Supervisor: Dr. Gönenç Ercan

September 2019, 97 Pages

Dictionaries catalog and describe the semantic information of a lexicon. WordNet provides an edge by presenting distinct concepts with the hierarchy information among them. Research in computer science has been using this hand crafted tool in natural language applications such as text summarization and machine translation. Original WordNet has been compiled for English yet counterparts for other languages are not as readily available nor as comprehensive. In order for research on languages other than English to benefit from the power of a WordNet, machine assisted creation and evaluation methods are essential.

Word embeddings can provide a mapping between words and points in a real valued vector space. Using these vectors, representing documents as well as forming geometric relationships between them is a well studied area of research. In this thesis we start by hypothesizing that a dictionary definition captures the semantic basis of the described word. We used word embeddings as building blocks to map dictionary definitions into a multidimensional space. These spaces can be aligned to accommodate two languages, allowing the transfer of information from one language to another. We investigate the success of retrieving and matching discrete senses across languages by employing supervised and unsupervised methods. Our experiments show that dictionary alignment can be evaluated successfully by using both unsupervised and supervised methods but corpora sizes should be taken into consideration. We further argue that some methods are not viable considering their poor performance.

Keywords: dictionary alignment, word embeddings, semantic encoder, short text similarity

TEŞEKKÜR

Yüksek lisans sürecinde öğretileri ve deneyimleriyle bu tezin ortaya çıkmasındaki katkılarıyla bana kılavuz olan danışman hocam Dr. Gönenç Ercan'a,

Deneyimleri ve yorumları ile hep doğru yolda yürümeme yardımcı olan kıymetli hocam Dr. Tayfun Küçükyılmaz'a ve verimli bir çalışma ortamının hazırlanmasında her zaman desteğini duyumsadığım hocam Prof. Dr. Tolga Kurtuluş Çapın'a,

Bilgisayar mühendisliğini bana tanıtan değerli büyüğüm Hülya Küçükaras'a,

Bana her konuda yardımcı olan Ayça Deniz ve Hakan Kızılöz'e ve yaptığımız fikir alışverişleri ile bana farklı bakış açıları kazandıran Mehmet Taha Şahin'e,

Tez savunması sürecinde dönütleriyle tezimin bilimsel niteliğine katkı sağlayan juri başkanı Prof. Dr. İlyas Çiçekli'ye, Dr. Gönenç Ercan'a, Dr. Tayfun Küçükyılmaz'a, Dr. Mehmet Köseoğlu'na ve Dr. Burcu Can'a,

Beni bu günlere getiren, yol gösterici ve her zaman sevgi dolu olan canım aileme ve başından beri desteğini hiç esirgemeyen sevgili Asena Akkaya'ya,

İçtenlikle teşekkür ederim...

Yiğit Sever

Eylül 2019, Ankara