İSTANBUL TEKNİK ÜNİVERSİTESİ ★ FEN BİLİMLERİ ENSTİTÜSÜ

TÜRKÇE SÖZCÜK ANLAM BELİRSİZLİĞİ GİDERME

DOKTORA TEZİ

Bahar İLGEN

Bilgisayar Mühendisliği Anabilim Dalı

Bilgisayar Mühendisliği Programı

EKİM 2015

İSTANBUL TEKNİK ÜNİVERSİTESİ ★ FEN BİLİMLERİ ENSTİTÜSÜ

TÜRKÇE SÖZCÜK ANLAM BELİRSİZLİĞİ GİDERME

DOKTORA TEZİ

Bahar İLGEN
(504062506)

Bilgisayar Anabilim Dalı

Bilgisayar Mühendisliği Programı

Tez Danışmanı: Prof. Dr. Eşref ADALI
Eş Danışman: Yrd. Doç. Dr. Ahmet Cüneyd TANTUĞ

EKİM 2015

İTÜ, Fen Bilimleri Enstitüsü'nün 504062506 numaralı Doktora Öğrencisi Bahar İLGEN, ilgili yönetmeliklerin belirlediği gerekli tüm şartları yerine getirdikten sonra hazırladığı "TÜRKÇE SÖZCÜK ANLAM BELİRSİZLİĞİ GİDERME" başlıklı tezini aşağıda imzaları olan jüri önünde başarı ile sunmuştur.

Tez Danışmanı :	**Prof. Dr. Eşref ADALI** İstanbul Teknik Üniversitesi	.....
Eş Danışman :	**Yrd. Doç. Dr. Ahmet Cüneyd TANTUĞ** İstanbul Teknik Üniversitesi	.....
Jüri Üyeleri :	**Doç. Dr. Banu DİRİ** Yıldız Teknik Üniversitesi	.....
	Doç. Dr. Deniz YÜRET Koç Üniversitesi	.....
	Yrd. Doç. Dr. Gülşen C. ERYİĞİT İstanbul Teknik Üniversitesi	.....
	Prof. Dr. Tunga GÜNGÖR Boğaziçi Üniversitesi	.....
	Doç. Dr. Şule G. ÖĞÜDÜCÜ İstanbul Teknik Üniversitesi	.....

Teslim Tarihi : 18 Ağustos 2015
Savunma Tarihi : 21 Ekim 2015

Hocama,

ÖNSÖZ

Doktora öğrenimim ve tez çalışmam süresince bana göstermiş olduğu her türlü destek ve yardımlarından dolayı değerli hocam ve danışmanım Prof. Dr. Eşref ADALI'ya sonsuz minnet ve teşekkürlerimi sunarım. Aynı zamanda benden desteğini esirgemeyen, değerli görüş ve yönlendirmeleri ile bu çalışma süresince farklı bakış açısı kazanmamda yardımcı olan eş danışmanım Yrd. Doç. Dr. Ahmet Cüneyd TANTUĞ'a teşekkürlerimi sunarım.
Tez izleme komitemde yer alan değerli hocalarım Doç. Dr. Banu DİRİ, Yrd. Doç. Dr. Gülşen Cebiroğlu ERYİĞİT ve Doç. Dr. Deniz YÜRET'e bu çalışma süresince paylaştıkları görüşleri doğrultusunda çalışmanın her zaman daha iyiye yönlenmesindeki katkıları için teşekkürlerimi sunarım.
Çalışma boyunca her zaman yanımda olan, bana her konuda destek olan aileme ve tüm arkadaşlarıma sonsuz teşekkürlerimi sunarım.

Ağustos 2015 Bahar İlgen

İÇİNDEKİLER

Sayfa

KISALTMALAR

Aİ : Anlamsal İşaretleme
BÇ : Bilgisayarlı Çeviri
BD : Bilgisayarlı Dilbilim
BE : Bilgiye Erişim
BKA : Birlikteliklerin Kazandırdığı Anlamlar
BOW : Bag of Words
BY : Birliktelik Yöntemleri
ÇD : Çapraz Doğrulama
ÇG : Çekim Grubu
DDİ : Doğal Dil İşleme
DVM : Destek Vektör Makineleri
EBA : En Baskın Anlam
EFO : En Fazla Olabilirlik
EKKA : En Küçük Kapsayan Ağaç
GAİ : Gizli Anlamsal İndeksleme
HSD : Hedef Sözcük Derlemi
KA : Karar Ağaçları
KL : Karar Listeleri
KÖ : Konumsal Özellikler
NB : Naive Bayes
NKA : Niteliklerin Kazandırdığı Anlamlar
ÖA : Özellik Azaltımı
ÖÇ : Özet Çıkarma
ÖTÖ : Örnek Tabanlı Öğrenme
ÖY : Önyükleme Yöntemleri
SAA : Sözcük Anlam Ayrıştırma
SABG : Sözcük Anlam Belirsizliği Giderme
SC : Soru Cevaplama
SK : Sözcük Kesesi
SSY : Seçilmiş Sözcük Yaklaşımı
SÖ : Sözcüksel Örnekler
TAYS : Tek Anlamlı Yakın Sözcükler
TDA : Tekil Değer Ayrışımı
TDK : Türk Dil Kurumu
TS : Türetim Sınırı
TSY : Tüm Sözcükler Yaklaşımı
YAB : Yapısal Anlamsal Bağlantılar
YSA : Yapay Sinir Ağları

ÇİZELGE LİSTESİ

Sayfa

ŞEKİL LİSTESİ

Sayfa

TÜRKÇE SÖZCÜK ANLAM BELİRSİZLİĞİ GİDERME

ÖZET

Doğal dillerde yaygın olarak gözlenen "Anlam Belirsizliği" kavramı bir sözcüğün birden fazla anlama sahip olması durumudur. *Sözcük Anlam Belirsizliği Giderme* (SABG) işlemi, birden fazla anlama sahip sözcüğün kullanıldığı bağlamda en uygun anlamının belirlenmesi olarak tanımlanmaktadır.

İnsanlar arası iletişimde, önceki deneyimler ve karmaşık insan bilişsel süreçlerinin yardımıyla çözümlenen anlam belirsizliği, bilişim ve *Doğal Dil İşleme* (DDİ) alanlarında da ele alınmakta olan önemli ve güncel konular arasında yer almaktadır. Bir sözcüğün anlamının belirginleştirilmesi DDİ alanındaki uygulamaların tamamına yakınında başarıma katkı sağlayan ve gereksinim duyulan bir adımdır. Bu uygulamalar, *Bilgiye Erişim* (BE), *Bilgisayarlı Çeviri* (BÇ), *Anlamsal İşaretleme* (Aİ), *Soru Cevaplama* (SC) gibi pek çok alanı içine almaktadır. Günümüzde *Bilgisayarlı Dilbilim* (BD) çalışmalarına internet ve diğer alanlarda duyulan gereksinim büyük boyutlara ulaşmıştır. Bu gereksinim sonucunda, sözü edilen DDİ uygulamaları kapsamında çeşitli yöntem ve algoritmalar geliştirilmiştir. Bu çalışmalarda, dillerin yapısı, mevcut kaynak ve kısıtlar, uygulamanın gereklilikleri gibi unsurların önemli rolü olduğu ve yöntemlerin bu doğrultuda geliştirildiği bilinmektedir.

SABG alanındaki çalışmalar göz önünde bulundurulduğunda, bilgi, derlem tabanlı ve melez yöntemler olmak üzere üç yaklaşımın öne çıktığı görülmektedir. Bilgi tabanlı yöntem ailesinde sözlük, eş anlamlılar sözlüğü ve ontolojiler kullanılan temel kaynakları oluşturmaktadır. Derlem tabanlı yaklaşımlarda bilgi derlemlerden öğrenilmektedir. Derlem tabanlı yöntemler kendi içinde denetimli, denetimsiz ve yarı denetimli alt sınıflarına ayrılmaktadır. SABG alanında yapılan ilk çalışmalarda ağırlıklı olarak anlam işaretli derlemlere gereksinim duyulan denetimli yöntemler üzerinde durulmuştur. Denetimli yöntemlerle anlam belirsizliği yüksek doğrulukla giderilse de Türkçe gibi kaynak ve derlemlerin kısıtlı olduğu dillerde yarı denetimli ve denetimsiz yöntemler yakın zamanda yapılan çalışmalarda önem kazanmıştır. Denetimsiz ve yarı denetimli yöntemleri tercih edilir kılan bir diğer sebep ise derlem anlam işaretlemelerinin emek yoğun bir süreç olmasıdır. İşaretli derlemlerin yetersiz olduğu ya da kullanılmadığı çalışmalarda sözlük anlamlarına bağımlılık ortadan kalkmakta ve derlemler sözcüğün anlamlarını kendi bulmaktadır. Melez yöntemlerde ise bilgi ve derlem tabanlı yöntemlerden birlikte faydalanılmaktadır.

Kullanılan denetim seviyesinin yanında SABG alanı için yapılan bir diğer sınıflandırma, probleme olan yaklaşımın kapsamı ile ilgilidir. Sözü edilen sınıflandırma; 1^0 *Seçilmiş Sözcük Yaklaşımı* (SSY) ve 2^0 *Tüm Sözcükler Yaklaşımlarını* (TSY) içine almaktadır. İlk yaklaşımda, önceden saptanan bir hedef sözcüğe ilişkin örneklerdeki belirsizlik giderilmektedir. SSY yaklaşımında sözcük ve anlam kümeleri sınırlı olduğundan anlam belirsizliği giderilmesinde genellikle denetimli makine öğrenmesi yöntemlerinin kullanımı tercih edilmektedir. Anlam etiketli örnekler sınıflandırıcının eğitilmesinde kullanılmaktadır. Anlam belirsizliğine

sahip bir sınama örneğinin anlam ataması eğitilen sınıflandırıcı ile gerçekleştirilmektedir. TSY yaklaşımında ise belirli bir metin içerisindeki tüm sözcüklerin belirsizliğinin giderilmesi hedeflenmektedir. TSY yaklaşımı ile sözcük türü etiketleme arasında benzerlik bulunmakla birlikte, TSY yaklaşımında gereksinim duyulan etiket kümesi diğerine göre çok daha büyük olmaktadır. Etiket kümesinin büyüklüğü ise her sözcük için yeterli miktarda örnek bulmadaki zorluktan ötürü veri seyrekliği sorununa yol açmaktadır.

Sözlük ya da derlemlerden elde edilen bilgi, anlam belirsizliği gidermede en temel bileşendir. Bilgi kaynakları, *görünüm bilgisi* ya da *öğrenilmiş bilgi* sınıflarına ayrılmaktadır. İlk grup sözcük *anlam sıklıkları, kavram ağaçları, seçimsel öncelikler,* sözcük etiketleri gibi bilgi türlerini içine alırken, öğrenilmiş bilgi sınıfları ise *belirtici sözcükler, sözdizimsel özellikler, alana özgü bilgiler* ve *paralel derlemler* gibi alt sınıflardan oluşmaktadır. Yapılan çalışmalarda öğrenilmiş bilgi türlerinin daha çok denetimli yöntemlerde, görünüm bilgisinin ise denetimsiz yöntemler dahilinde kullanıldığı gözlenmiştir. Uygulamada ise bilgi kaynaklarının çeşitli kombinasyonları SABG çalışmalarında kullanılmaktadır.

Makine öğrenmesi yöntemleri derlem tabanlı SABG yöntemleri dahilinde anlam belirsizliği giderme bilgisinin otomatik olarak çıkartılmasında kullanılmaktadır. Bir SABG uygulamasında genellikle kullanılan kaynaklar; anlam işaretli derlemler, çevrimiçi sözlükler ve doğal dillere ilişkin geniş ölçekli kaynaklardan oluşmaktadır. Özellik kümesi ve öğrenme aşamasında kullanılan algoritma seçimi bir SABG uygulamasında gözetilen iki önemli unsurdur. Bir çok DDİ uygulamasında makine öğrenmesi yöntemleri ile elde edilen bilgiden faydalanılmaktadır. SABG alanında kullanılan denetimli yöntemler model ya da kuralların oluşturulma biçimine göre sınıflara ayrılmaktadır. Bu yaklaşımlar istatistiksel yöntemler (Naïve Bayes), benzerlik tabanlı yöntemler (k- En Yakın Komşu algoritması), konuya özgü özellikler (Bir söz öbeği/bağlam için bir anlam), ayrıştırıcı kural yöntemleri (karar listeleri, karar ağaçları, kural birleşimine dayalı yöntemler), doğrusal sınıflandırıcılar ve Kernel yöntemlerinden oluşmaktadır.

Sözcük etiketleme ve sözdizimsel analiz gibi DDİ alanındaki diğer çalışma konularına kıyasla SABG konusu bir takım zorlukları içermektedir. Her sözcük bir anlam ile eşleşeceğinden tam bir eğitim verisinin oluşturulabilmesi için çok büyük miktarda örnek gereksinimi ortaya çıkmaktadır. Dildeki veri seyrekliği problemini aşmanın bir yolu eğitim algoritmasında kullanılacak özelliklerin doğru seçilmesinden geçmektedir. Bu özellikler, yerel ya da geniş ölçekte bulunabilmektedir. Makine öğrenmesi yönteminin uygulanmasından önce tüm örneklerin öğrenme algoritması tarafından anlaşılacak şekilde kodlanması gerekmektedir.

Konumsal Özellikler (KÖ) ve *Sözcük Kesesi* (SK) özellikleri SABG çalışmalarında ele alınan hedef sözcüğün komşularından elde edilen iki önemli özellik grubudur. Yapılan çalışmaların tamamına yakınında belirsizliği giderilmek istenen sözcüğün merkezde olduğu bir "n" pencere aralığından faydalanılmaktadır. Konumsal özellikler ile hedef sözcüğün sol ve sağ komşularına ilişkin bilgiler kullanılmaktadır. Kullanılan bilgiler, sözcük gövde biçimleri ve sözcük türleri gibi bileşenlerden oluşmaktadır. İkinci grup olan SK özelliklerinde ise sözcükler herhangi bir sıra ya da konum gözetilmeksizin ele alınmaktadır. Benzerlik ölçütü olarak seçili penceredeki sözcüklerin konum gözetmeksizin bulunup bulunmama durumlarına ve sıklıklarına bakılmaktadır. Doğal dillerdeki kısıtlı kaynaklar göz önünde bulundurulduğunda,

faydalanılacak bilginin doğru seçilmesi ve etkin özelliklerin kullanılması derlemlerdeki doğru anlamların belirlenmesinde özellikle önemlidir.

Bu çalışma kapsamında yapılan özgün çalışmalar ve katkılar aşağıda açıklanmıştır:

- *Hedef Sözcük Derlemi* (HSD**):** Her bir paragrafı hedef sözcük içeren metinlerden oluşan bir derlem hazırlanmıştır. Derlemin hazırlanması sırasında önce Türkçede belirsizlik derecesi yüksek olan isim ve eylemler belirlenmiştir. Ardından seçilen sözcükler için dengeli olarak metinler toplanmıştır. Daha sonra bu sözcükler oylayıcılar tarafından Türk Dil Kurumu (TDK) sözlüğündeki anlamlar ile işaretlenmiştir.
- Etkin Özelliklerin Bulunması: Etkin özelliklerin bulunmasında iki yöntem denenmiştir: 1^0 *Konumsal Özellikler* , 2^0 *Sözcük Kesesi* Özellikleri.
- Bir sözcüğün belirsizliğinin giderilmesinde etkin özelliklerin ortaya çıkartılabilmesi için denetimli yöntemler üzerinde çalışılmıştır. Bu çalışmanın sonunda hedef sözcüğün öncesinde ve sonrasında yer alan sözcüklerin etkin özellikleri çıkartılmış ve bu özelliklerin belirsizlik gidermeye katkıları ortaya konulmuştur. Bu çalışmalar yapılırken pencere boyu sabit tutulmuştur. Çalışma kapsamında sözcük kök ve eklerinin oluşturduğu biçimbilimsel analiz çıktılarının hedef sözcük ve komşuları ile birlikte değerlendirmeye alındığı konumsal özellikler sınanmıştır.
- Etkin özellikleri belirlemede sözcük kesesi yönteminin katkısı da incelenmiştir. Etkin özelliklerin bulunmasında en uygun kese boyu belirlenmiştir. Bu çalışma hedef isim ve eylemler için gerçekleştirilmiştir. Kese içinde bulunan sözcüklerin hedef sözcüğün ne kadar yakınında araştırılması gerektiği ortaya konmuştur.
- Konumsal özellikler ile sözcük kesesi yöntemlerinin sonuçları karşılaştırılmış, konumsal özelliklerin anlam belirsizliği gidermedeki etkisinin daha yüksek olduğu gösterilmiştir. Her iki yöntem birlikte kullanıldığı takdirde daha iyi sonuç elde edilmiştir.
- Denetimli yöntemlerle belirsizlik giderme çalışmasını sürdürebilmemiz için Türkçeyi yetkin biçimde temsil eden ve sözcüklerin anlamları işaretlenmiş derlem gerekmektedir. Böyle bir derlemin Türkçe için olmadığı ve yapılmasının çok emek yoğun olacağı bilindiği için çalışmamızı denetimsiz yöntemlere yöneltmiş bulunmaktayız.
- Denetimsiz yöntem olarak DDİ alanında çok az kullanıldığına tanık olduğumuz çizge tabanlı yöntem Türkçe için geliştirilmiştir. İlk aşamada yöntemi gerçekleştirmek üzere gerekli program hazırlanmıştır. İkinci aşamada yöntemi etkin kılmak için çizge parametrelerinin etkisi değerlendirilmiştir. Denetimsiz yöntemin çizgesini oluşturmak ve başarımı ölçmek için HSD kullanılmıştır.
- Sonuç olarak geliştirdiğimiz denetimsiz yöntem ile Türkçe sözcüklerinin belirsizliğinin giderilebileceği ortaya konmuştur. Ancak yöntemimizin en son aşamasında merkez düğümlerin anlamları işaretlenmiş derlemden yararlanarak belirlenmiştir. Geliştirdiğimiz yöntemin başarımı denetimli yöntemlere yakın ölçülmüştür.

WORD SENSE DISAMBIGUATION FOR TURKISH

SUMMARY

As being one of the pervasive characteristics of the natural languages, the research on word sense ambiguity aims at resolving the problem of having more than one sense. A *Word Sense Disambiguation* (WSD) task is defined as automatic assignment of the most appropriate meaning to a polysemous word within a given context.

The problem of word sense ambiguity, which can be resolved during human communication by using previous experiences and complex cognitive processes, is also one of the active topics in computer science and *Natural Language Processing* (NLP) area. The identification of word meanings is required in almost all applications of the NLP area to provide them proper functioning. These applications include the areas such as *Information Retrieval* (IR), *Machine Translation* (MT), *Semantic Annotation* (SE), *Question Answering* (QA) and many others. From this aspect, WSD is an important intermediate step for all these applications that increases their performances. There is a huge need in NLP related fields and internet environment for the development of *Computational Linguistics* (CL) methods. As a result, several algorithms have been developed for the different fields of the NLP area. In the scope of these works, the properties such as nature of the languages, available resources and constraints, application requirements play important role to develop methods.

The WSD methods are classified under three broad categories: knowledge-based, corpus-based and hybrid methods. The family of knowledge-based methods primarily relies on dictionaries, thesauri, ontologies and lexical knowledge bases. Corpus-based methods are further classified into supervised, unsupervised and semi-supervised methods (or minimally supervised). Previous efforts on WSD have mainly focused on supervised approaches that require sense annotated corpora. There are also alternative approaches of unsupervised and semi-supervised methods that try to lower the sense-annotated portion of the texts. Although sense ambiguity can be resolved in supervised systems with high accuracy, usage of semi-supervised and unsupervised methods has gained attention recently since the sense annotation scheme is labor intensive and expensive. In some of the studies, word senses are extracted from corpus itself where sense-annotated corpora are insufficient or not used. Recently, the extraction of word senses from corpus is preferred by the researchers since the pre-defined sense definitions of dictionaries may be too limited. On the other hand, the adaptation of solutions and methods to new domains may be difficult because of the dynamic nature of word senses. In the scope of hybrid methods, knowledge-based and corpus-based methods are combined.

WSD can also be classified according to the scope of approach to the problem. The level of supervision is the first criterion to classify the methods. A secondary classification for generic WSD can be made by considering two variants. These options of WSD can be selected from: 1^0 *Lexical Sample* (LS) task and, 2^0 *All-words*

(AW) task. The former approach disambiguates the occurrences of a small sample of target word that has been determined previously. Since the words and the set of senses are limited, supervised *Machine Learning* (ML) methods are usually used to handle LS tasks. Hand-labeled examples are used to train the classifier. Then unlabeled test portion of the target words can be labeled by using trained classifier. In contrast, AW approach comprises the disambiguating all the words in a running text. All the entries in a given system are required to be disambiguated. There is a similarity between AW task and *Part of Speech* (POS) tagging. The only difference is that the former needs much larger set of tags. This larger set of tags resulted in data sparseness problem since it is hard to find adequate training data for each word.

Knowledge is the fundamental component for a WSD system which can be acquired from dictionaries or learned from a training corpus. The sources can be classified into "lexical knowledge" and "learned world knowledge" categories. The lexical knowledge category includes the knowledge sources such as "sense frequency", "concept trees", "selectional restrictions", "subject code" and the POS information. The latter category includes the usage of "Indicative words", "syntactic features", "domain specific knowledge" and "parallel corpora". It is usually observed that the unsupervised systems need lexical knowledge sources while supervised systems use world knowledge. But in practice the combinations of these sources have been used in WSD systems.

ML techniques are used to automatically acquire disambiguation knowledge in the scope of corpus-based WSD methods. A typical WSD system may utilize sense-tagged corpora, online dictionaries and large scale linguistic resources as components. The set of features to be used and the learning algorithm are two of the important decisions that have to be considered for the design of a WSD system. Many NLP systems rely on linguistic knowledge acquired from hand-labeled training text data and ML methods. The supervised methods of the WSD can be classified according to the induction principle they use to acquire model or rules. These methods consist of probabilistic models (e.g., Naïve Bayes), similarity based methods (e.g., k-Nearest Neighbor algorithm), methods based on discursive properties (e.g., one sense per discourse/collocation, attribute redundancy), methods of discriminative rules (e.g., decision lists, decision trees or methods based on rule combination), linear classifiers and Kernel-based methods.

Compared to the other subjects in NLP such as POS determination and syntax parsing, a WSD problem introduces extra difficulties. Since each word is associated with unique meaning, complete training set requires a huge number of examples. This language sparsity problem is dealt with by selecting features used in training algorithms. These features can be found in local or wider context. Before applying the ML algorithm, all the examples of a particular ambiguous word have to be encoded in a way the learning algorithm can handle.

Collocational and Bag-of-Words (BoW) features are two important classes of features that are generally extracted from neighboring contexts in WSD tasks. Almost all of these approaches are employed by defining a window of "n" content words around the word to be disambiguated in the corpus. Collocational features encode information about the lexical inhabitants of specific positions located to the left or right of the target word. The basic elements may consist of the word, its root form and the part of speech information. BoW is the second feature set in which the text is treated as an unordered bag of words. Within this approach, similarity

measures are calculated by looking at the semantic similarity between all the words in the window regardless of their positions. Considering the limited resources available for natural languages, it is especially important to select knowledge sources and the feature sets carefully to disambiguate senses.

Overall results of this study can be summarized as follows:

- Turkish Lexical Sample Corpus: In the scope of this study, a special corpus for Turkish has been prepared. For this task, the Turkish nouns and verbs have been determined by considering highly ambiguous ones among the dictionary of Turkish Language Association (TLA). Then samples have been collected for each ambiguous candidate word. Voters annotated the samples by using the sense definitions of the TLA dictionary.
- Extracting Effective Features: Two approaches have been tested to extract effective features: 1^0 *Collocational Features* (CF), 2^0 *Bag-of-Word*s.
- Supervised methods have been used to extract effective features on disambiguating word senses. The effective features of neighbor words around ambiguous headword, have been determined. The contribution of these features on disambiguating word senses has been investigated. A fixed window size has been used along the experiments. In the scope of the study, collocational features which comprise the morphological analysis outputs of the word roots and suffixes have been investigated.
- The contribution of using BoW features has also been investigated. The proper size for selected features has been determined. This work has been conducted for Turkish noun and verb sets. The optimal extent around headword to encode BoW features is determined.
- The results of collocational and BoW features have been compared. It is shown that the collocational features are more effective than BoW features on resolving sense ambiguities. Better results are achieved by combining two feature sets.
- Our research on supervised methods shows that a comprehensive and very large corpus that represents the language effectively is needed to be able to continue conducting research on WSD. There is no such a large corpus in Turkish. We focused our research on unsupervised methods since it is too labor-intensive to prepare such a corpus.
- A graph-based unsupervised method which previously used in a few NLP related studies have been developed for Turkish. At the initial phase of the study, a program has been developed to implement the algorithm. Then the effect of supervised method findings is investigated to enhance the results. The Turkish lexical sample corpus has been used to generate graph and evaluate the accuracy results.
- Our research show that sense ambiguities can be resolved by using unsupervised methods. We propose the gold standard evolution at the final stage and use annotated word senses of Turkish lexical sample corpus to map hub meanings. This method yielded nearly as reliable results with the supervised methods.

1. GİRİŞ

Dillerin temel öğelerinin sözcükler olduğu bilinmektedir. Her bir sözcüğün karşılık geldiği anlam o dile ilişkin sözlüklerde açıklanmaktadır. Bir başka deyişle sözlükler bir sözcüğün hangi anlamlarda kullanıldığını açıklar. Bilindiği gibi bir çok sözcüğün birden fazla anlamı bulunmaktadır. Türk dili değerlendirildiğinde, birden fazla anlamı olan bir sözcüğün ortalama 3,53 anlamı olduğu görülmüştür. Örneğin ekmek sözcüğünün Türk Dil Kurumu (TDK) büyük sözlüğünde 10 anlamı görülmektedir.

- Tahıl unundan yapılmış hamurun fırında, sacda veya tandırda pişirilmesiyle yapılan yiyecek
- İnsanı geçindirecek iş, kazanç
- Yemek, aş
- Bir bitkiyi üretmek için toprağa tohum atmak veya gömmek
- Toprağı ekip biçmek için kullanmak
- Serpmek
- Bir şeyin başlamasına yol açacak sebepleri hazırlamak
- Birini uydurma bir sebeple bırakıp gitmek, savuşmak, atlatmak
- Parayı boşuna harcamak, ziyan etmek
- Yarışta geçmek

İnsanlar dinledikleri ya da okudukları bir sözcüğün anlamını, daha önce dinlediği ya da okuduğu kısımlardan edindiği bilgilerin ışığında kesinleştirir. Aşağıdaki iki örnek sözcüklerin nasıl farklı anlamlandırıldığını göstermektedir:

Örnek 1. *Yorgun **gözleri** umutsuzca etrafı süzüyordu.*

*Çekmecenin **gözleri** ağzına kadar doluydu.*

Örnek 2. *Ağacın **kökleri** çok daha derinlere, alt katmanlara uzanıyordu.*

***Kökleri** eskilere, Osmanlıya dayanan bir gelenekti.*

Birinci örnekte yer alan ilk tümcede "yorgun" sözcüğü "gözleri" sözcüğünü nitelemekte ve sözcük anlamını belirgin hale getirmektedir. İkinci tümcede ise "gözleri" sözcüğü aldığı çekim ekleri ile farklı bir kullanımdadır, "çekmece" sözcüğü kapatıldığında anlamı doğrudan seçilemeyecek ve belirsizlik ortaya çıkacaktır.

İkinci örneğe ilişkin ilk ve ikinci tümcelerde ise "kök" sözcüğünün farklı kullanımları görülmektedir. İlk tümcede kök sözcüğü ile kastedilen sözcüğün bitki anlamıdır. İkinci tümcede ise "kök" sözcüğü bir geleneğin "köklü" olması durumunu ifade etmek için kullanılmaktadır. Bu örneklerde gözlemlenen belirsizlikler, insanlar tarafından geçmiş deneyimler ve önceki tümcelerden faydalanarak giderebilmektedir. Benzer işlemi bilgisayara yaptırmak *Doğal Dil İşleme* (DDİ) bilim dalında *Sözcük Anlam Belirsizliğinin Giderilmesi* (SABG) işlemi olarak adlandırılmaktadır.

Bilgisayarlı Dilbilimi (BD) çalışmalarında, diğer bir deyişle DDİ çalışmalarında bir sözcüğün hangi anlamda kullanıldığının bilinmesi önemlidir. Çünkü:

- Bir tümcenin anlamının çıkartılmasında,
- Bir makalenin anlamının çıkartılmasında,
- Özet çalışmalarında,
- Diller arası çevirilerde,

sözcüklerin kesin anlamlarının bilinmesi gerekmektedir. Diller arası çeviride ise son derece önem kazanmaktadır. Örneğin, İngilizce yazılmış *"the veels of wagon are broken"* tümcesi, sözcüklerin anlamlarına özen gösterilmeden Türkçeye çevrildiğinde; *"vagonun tekerlekleri kırık"* tümcesi elde edilir. Bu çeviriyi okuyan kişi tren vagonun tekerleklerinin kırık olduğu anlamını çıkarır. Ancak İngilizce tümcenin anlatmak istediği "posta arabasının tekerleklerinin kırık olduğudur. Dolayısıyla diller arası çeviri yaparken her iki dilde de sözcüklerin aynı anlamda kullanılması çok önemlidir. Yukardaki İngilizce tümcenin doğru çevirisi *"Posta arabasının tekerlekleri kırık"* olmalıdır.

Bir ilginç örnek aşağıda verilmektedir (Adalı, 2012).

Köprücüler İstanbul'da toplanıyor.

Çok satan ve saygın bir gazetemizde çıkan bu başlığı ilk okuyan bir okur, köprü inşaatı ile ilgilenen yetkililerin İstanbul'da bir toplantıda bir araya geleceklerini

düşünür. Yazıyı okumaya devam ettiğinde şaşıracaktır. Çünkü toplantı sonunda birinci geleceklere ödüllerin verileceğinden söz edilmektedir. Okuyucu biraz kafasını yorduğunda toplantıya katılanların köprü yapımcıları olmadığını, briç oyuncuları olduğunu anlayacaktır. Dış kaynaklı bu haberi dilimize çeviren kişi, İngilizcedeki (bridge) ile sesteş olan briç oyunu ve köprü sözcüklerini karıştırmıştır. Aslında İngilizcede "bridge" sözcüğünün başka anlamları da vardır.

Türkçe sondan eklemeli bir dil olması nedeniyle çok sayıda ek alabilmektedir. Özellikle yapım ekleri sözcüğe eklendiğinde sözcüğün anlamını da değiştirmektedir. Dolayısıyla Türkçe sözcüklerin belirsizliğinin giderilmesi sorunu, çekimli dillerin bu konudaki sorunlarına oranla çok karmaşıktır. Örneğin, Türkçede göz sözcüğünden sadece yapım ekleri kullanılarak farklı anlamdaki aşağıdaki sözcükler kolayca türetilebilmektedir.

Göz, gözlük, gözlükçü, gözlükçülük, gözcü, gözcülük, gözlem, gözleme, gözlemci, gözlemcilik, gözde…

Yukarıdaki her bir sözcük için çekimli bir dil olan İngilizcede ise ayrı ayrı karşılıklar kullanılmaktadır:

Eye, eyeglass, optician, opticians, watchman, …

1.1 Tezin Amacı

Bu tez çalışmasının amacı Türkçe sözcüklerin anlam belirsizliklerinin bilgisayarlı dilbilimi yöntemleriyle giderilmesidir. Bir sözcüğün anlamı içinde bulunduğu tümceye veya daha önceki tümcelere bağlı olarak kesinleştirilebilir. Dolayısıyla çalışmamızda bir metin içinde geçen ve seçilen bir sözcüğün hangi anlama geldiği bir başka deyişle kesin anlamının ne olduğu araştırılmıştır. Bu hedefe ulaşmak için şu ana kadar Türkçe için yapılmış olan benzer çalışmalardan daha yüksek başarımlı yöntem ve algoritmaların geliştirilmesi amaçlanmıştır.

1.2 Yakın Çalışmalar

Diller, kökenleri göz önünde bulundurulduğunda; Hint-Avrupa, Hami-Sami, Ural-Altay, Çin-Tibet ve Bantu dil ailesi olmak üzere beş sınıfa ayrılmaktadır. Temel yapılarına göre dil sınıfları incelendiğinde ise, eklemeli, çekimli ve tek heceli olmak üzere üç temel grup karşımıza çıkmaktadır. Tek heceli dillerde sözcükler tek heceden

oluşmakta, çekime girmeyerek her zaman kök durumunda kalmaktadır. Bu dillerde tümcelerin anlamı genellikle sözcüklerin diziliş sırasına göre ortaya çıkmaktadır. Biçim olarak birbirine benzeyen sözcüklerin anlam farkı genellikle dildeki zengin vurgularla belirginleşmektedir. Çekim eklerinin kullanılmadığı bu dillerde bir sözcük kullanıldığı yere göre pek çok farklı anlam kazanabilmektedir. Bazı Himalaya, Afrika dilleri ile Avrupa Bask dili bu gruba girmektedir. Bizim çalışmamıza yakın çalışmalar aşağıdaki sınıflandırmalar göz önünde bulundurularak değerlendirilmiştir:

- Türk dili için yapılmış olan çalışmalar
- Eklemeli diller için yapılmış çalışmalar
- Çekimli (bükümlü) diller için yapılmış çalışmalar
- Diğer çalışmalar

Bu bakış açısına göre yapılmış değerlendirmeler aşağıda verilmiştir.

1.2.1 Türk dili için yapılmış olan çalışmalar

Türkçe sözcük anlam belirsizliklerinin giderilmesi konusunda yayınlanmış doktora düzeyinde tek bir çalışma bulunmaktadır (Orhan, 2006). Orhan (2006) tarafından yapılmış olan bu çalışmada "Derleme Metin" tabanlı yaklaşımlar tercih edilmiştir. Orhan'ın (2006) iki tür derleme metin üzerinde çalıştığı tezde ilk derlem, dünya klasiklerinden seçilen yedi farklı hikâyeden (Gulliver, Candide, Ivan Nikiforovic, Tours Papazı, Mozart Prag Yolunda, Mektuplar ve Kır Atlı) oluşmaktadır. İkinci derleme metin ise ODTÜ ve Sabancı Üniversitesi işbirliği ile geliştirilmiş (ODTÜ-Sabancı Ağaç Yapılı Derlemi, tez içinde ODTÜ-Sabancı derlemi olarak anılacaktır) derlemdir.

İlk derleme metin, üzerinde dilbilimsel çalışma yapılmamış, tarayıcıdan ham veri olarak aktarılan bir kaynaktır. Derlemin bu özelliği nedeniyle kullanıma uygun hale getirilmesi için uzun bir ön işleme ve elle işaretleme süreci gerektirdiği kaydedilmiştir. İlk kaynakla ilgili karşılaşılan zorluklar sonucunda çalışmanın devamında ODTÜ-Sabancı derleme metninden faydalanılmıştır. Bu çalışma kapsamında seçilen algoritmalar, sözü edilen derlemlerden çıkarılan Konumsal Özellikler (KÖ) ve sözdizimi özellikleri gibi çeşitli özellikler kullanılarak sınanmıştır. Yapılan çalışmalar arasında; yapay sözcüklerin kullanılması ve Senseval (*www.senseval.org*) çalıştayları kapsamında yürütülen çalışmalarda kullanılan

Sözcüksel Örnek (SÖ) yapısındaki verinin ve benzer çalışmanın Türkçeye uyarlanması da yer almaktadır.

Bu çalışmada geliştirilen yöntem ve seçilen özellikler, sözü edilen doktora tezinde kullanılan ODTÜ-Sabancı derlemi üzerinde ayrıca sınanarak karşılaştırmalı sonuçlar elde edilmiştir. Elde edilen karşılaştırmalı sonuçlar ilerleyen bölümlerde ayrıntılı olarak anlatılmaktadır.

1.2.2 Eklemeli diller için yapılmış çalışmalar

Türkçe ile benzer özellik gösteren diller Japonca, Macarca, Moğolca, Fince, Korece vb. gibi sözcüklerin köklerinin değişmediği dillerdir. Bu dillerin kullanımında sözcüğe getirilen ekler sözcüklerin anlamlarını ve görevlerini belirler. İncelenen çalışmalardan değerlendirilenler aşağıda verilmektedir.

Moğolca için yapılan bir çalışmada Bataa ve Altangerel (2012) David Yarowsky'nin yaklaşımı izleyerek söz öbeklerini kullanmıştır. Yöntemdeki altı adım dört ana adıma indirgenmiştir. İlk adımda eğitim verisi toplama ve etiketleme işlemi gerçekleştirilmiştir. Anlam belirsizliği olan sözcükler için gazete, web, klasik romanlar ve hukuk yayınları gibi kaynaklar kullanılmıştır. Bir sonraki adımda söz öbek dağılımları incelenmiştir. Bu dağılımlar göz önünde bulundurularak anlam belirsizliği gidermede en faydalı durum araştırılmıştır. Sözcük anlamlarının sol-sağ komşu sözcükler ya da her ikisinin de ele alındığı bağlama bağlı olduğu düşünülmüştür. Anlam belirsizliğine sahip sözcüğün (hedef sözcük) sol ve sağ komşularının birlikte veya ayrı ayrı ele alındığı öbekler üzerinde çalışılmıştır. Söz öbekleri için elde edilen dağılımlar log-olabilirlik oranları (Logaritmik Olabilirlik Oranı) dikkate alınarak karar listelerine aktarılmıştır. Yarowsky (1993) çalışmasında her sözcük için bir karar listesi kullanmıştır. Moğol dili üzerinde yapılan çalışmada ise sözcüğün çekimleri de dikkate alınarak tüm sözcükler için bir karar listesi kullanılmış ve Moğol dili için bunun daha uygun olduğu belirtilmiştir. Son aşamada karar listesi kullanılarak anlam etiketlemeleri gerçekleştirilmiştir. Moğolca "cyp" sözcüğü için eğitim kümesinde yer almayan 137 tümce üzerinde yapılan sınama sonuçlarının doğruluğu %89,8 olarak bulunmuştur. Özetle Moğol dili için yapılan çalışmada "Bir Söz Öbeği İçin Bir Anlam" yaklaşımının uyarlaması kullanılmıştır.

Kore dili için yapılan çalışmada Yoon ve diğ. (2006), işlenmemiş metinlerden oluşan derlem ve bilgisayarla okunabilir sözlükleri kullanılmıştır. Sistem işaretsiz

derlemdeki sözcük çiftleri arasında bir benzerlik matrisini ve elektronik sözlükteki anlam tanımlarının vektör temsillerini kullanmaktadır. Çalışmada sözcüklerin anlam belirsizliğini gidermek için çevrimsiz, ağırlıklandırılmış ve yönlü bir çizge oluşturulmuştur. En uygun anlamın bulunması için Viterbi algoritması kullanılarak çizge yapısı üzerinden en iyi yol bulunmaktadır. Kore dili için yapılan bir diğer çalışmada Shannon'un bilgi kuramı kullanılmıştır (Lee ve diğ, 1997). Yapılan çalışmada sınıflandırma bilgisi "en olası sınıf" ve "belirginleştirme derecesi" adı verilen ölçütler kullanılarak elde edilmiştir. Anlam belirsizliği içeren hedef sözcüğü çevreleyen komşu sözcükler için en olası anlam ve gürültü dereceleri kullanılarak etiketli derlemde eğitim ve sınama yapılmıştır. Kore dili ve İngilizce üzerinde yapılan çalışmalarda doğruluk derecesi sırasıyla %84,6 ve %80,0 olarak bulunmuştur. Sondan eklemeli bir dil olan Japonca için de çeşitli çalışmalar yapılmıştır (Shinnou, 2001; Shinnou ve Sasaki, 2003; Atsushi ve diğ, 1996). Yapılan çalışmalarda kullanılan makine öğrenmesi yöntemleri ile başarılı sonuçlar elde edilmiştir (Li ve Takeuchi, 1997; Murata ve diğ, 2001).

Macarca için yapılan bir çalışmada ise bir bilgisayarlı çeviri sistemi içinde sözcük anlam belirsizliği giderimi yapılmıştır (Mihaltz, 2005). Çalışmada denetimli ve istatistiksel bir yöntem kullanılmıştır.

1.2.3 Çekimli diller için yapılmış çalışmalar

Çekimli diller sınıfına dahil olan Arapça, Farsça, İngilizce, Fransızca, Latince, Rusça vd. diller üzerinde yapılmış olan çalışmalardır. Özellikle İngilizce gibi Hint-Avrupa dilleri ise üzerinde en fazla çalışılmış ve ilerleme kaydedilmiş olan grubu oluşturmaktadır.

Yapılan çalışmalar, denetimli ve denetimsiz yöntemleri içine almaktadır. İngilizce gibi dillerde geniş kapsamlı olarak yapılmış çalışmaların sonucunda eklemeli dillerin aksine etiketli veriye ulaşma konusunda bir kısıt bulunmamaktadır. Bunun yanında WordNet (Miller ve diğ, 1990) gibi ontolojilerin kullanıma hazır olması, bilgisayarla okunabilir sözlükler ve tüm diğer kaynaklar bu diller için SABG alanında ilerleme sağlanmasına yardımcı olmuştur.

Bu gruptaki dillere ilişkin kaynakların kullanıma hazır olmasının sonucu olarak denetimli ve denetimsiz yöntemleri içine alan çok sayıda çalışma yapılmıştır. İlk zamanlarda yapılan çalışmalarda elle işaretlenmiş veri kullanılmıştır (Weiss, 1973;

Kelly ve Stone, 1975). Sözü edilen yaklaşımlarda kuralların elle oluşturulmasının pratikte sistemlere uygulanmasındaki zorluklar bildirilmiştir (Gale ve diğ, 1992a).

Yapılan çalışmalarda anlam belirsizliğinin giderilmesinde çeşitli kaynaklar kullanılmıştır. Bunların arasında bilgisayarla okunabilir sözlükler ve anlam işaretli derlemler bulunmaktadır. İlk grup arasında yer alan araştırmacılar; Lesk (1986), Walker (1987), Luk (1995) ve Ide (1990) anlam belirsizliğini ortadan kaldırmak için *Oxford's Advanced Learner's Dictionary of Current English* ve benzer kaynakları kullanmışlardır. Bu yöntemlerle, rastgele metinleri okuyan bir sistem geliştirilmekte ve metindeki her sözcük sözlükteki bir anlama işaret etmektedir. Bu yaklaşımların bir olumsuz yönü ise sözlüklerde yeterli miktarda ilgili bilgi bulunamadığı için verimli olmamasıdır.

İkinci tür kaynağın kullanıldığı çalışmalarda ise Miller (1994), Leacock (1993), Yarowsky (1992), Bruce (1994) ve Ng (1996) anlam belirsizliğini gidermek üzere kullanılan bilgiyi derlemlerden sağlamıştır. Bu yaklaşımlarda anlam belirsizliği içeren hedef sözcüğün komşuları sırasız olarak, hedef sözcük dilbilgisi etiketi, biçimbilimsel bilgiler ve sözdizimsel özellikler derlemlerden çıkartılmıştır. Derlemden çıkarılan bu bilgi istatistiksel sınıflandırıcılar, yapay sinir ağları, bilgiye-erişim tabanlı teknikler, ve örnek tabanlı öğrenme yöntemlerinde kullanılmıştır. Anlam işaretli derlemlerin kullanıldığı yöntemlerde insan müdahalesi daha az olmakta ve doğruluk değerleri daha yüksek elde edilmektedir.

Denetimli yöntemlerden sonra yapılan çalışmalarda "Bilgi Edinim Darboğazı" sorununu aşmaya yönelik yöntemler üzerinde durulmuştur. Yarowsky (1995) denetimsiz bir eğitim yöntemi geliştirmiş, Gale (1992a) ise bilgi edinim darboğazı sorunu için iki dilli bir derlem kullanmıştır.

Son zamanlarda yapılan çalışmalarda denetimsiz yöntemler kullanılarak *Sözcük Anlam Ayrıştırma* (SAA) çalışmaları üzerinde durulmuştur. SABG çalışmaları ile yakın ilişki içinde olan SAA yaklaşımlarında anlamlar derlemlerden çıkartılmakta ve anlam sınıfı atanması yerine var olan anlamların ayrıştırılması hedeflenmektedir. Bu yöntemlerde öncelikle bir anlam envanteri ortaya çıkarılmakta ve anlam ayrıştırma gerçekleştirilmektedir. Bu alanda yapılan bir çalışmada istatistiksel bir dil modelinden faydalanılarak anlamı belirginleştirilmek istenen hedef sözcük için temsil vektörü oluşturan bir sistem geliştirilmiştir (Başkaya ve diğ, 2013). Yöntemin

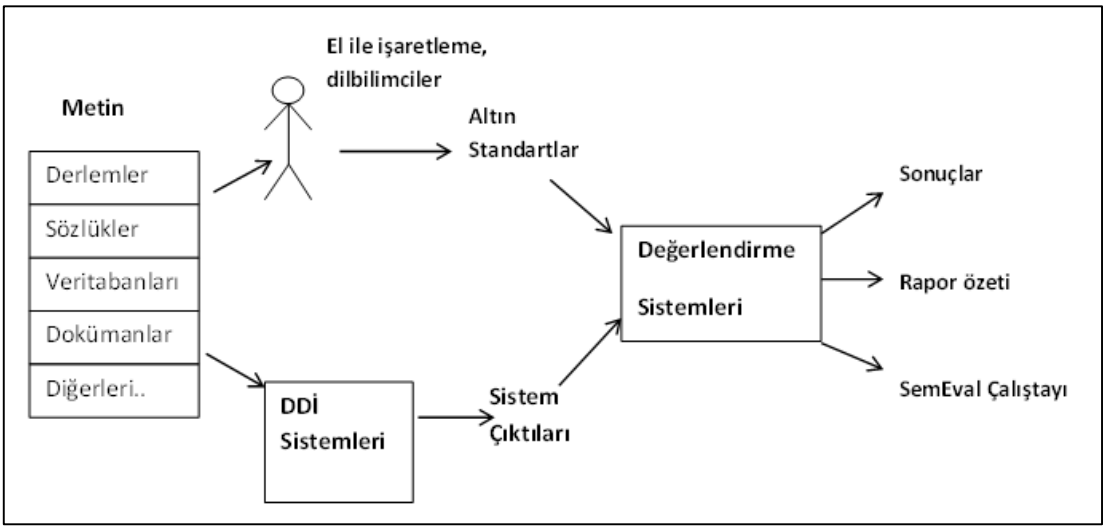
başarımı Semeval-2013 kapsamındaki benzer çalışma sonuçlarının başarımını geçmiştir. Yakın zamanda yapılan bir çalışmada *Deep Belief Networks* (DBN) (Hinton ve Salakhutdinov, 2006) algoritması ile elde edilen sonuçlar SABG alanındaki diğer algoritma sonuçları ile karşılaştırılmıştır (Wiriyathammabhum ve diğ, 2012). DBN yöntemleri veriden hiyerarşik bir temsil oluşturan grafiksel yöntemlerdir. DBN'ler ikili yapıdaki rastgele gizli değişkenlerden oluşan çoklu katman yapılarıdır. Gizli katmanlar aşamalı olarak öğrenilmekte ve aynı zamanda diğer katmanların öğrenilmesinde yinelemeli bir yapıda kullanılmaktadır. Çalışmada kullanılan algoritma, farklı özellik grupları ve bu özelliklerin birleşimi için sınanarak başarımı yüksek diğer SABG yöntemleri ile karşılaştırma sağlanmış ve daha yüksek doğrulukta sonuçlar elde edildiği kaydedilmiştir. Chen ve diğ. (2014) yaptıkları çalışmada SABG başarımını arttırmak için farklı bir sözcük anlam temsili kullanmıştır. Bunun arkasında yatan düşünce sözcüklerin anlam temsillerinin birbirinden bağımsız olmadığı fikridir. Son dönemde yapılan çalışmalarda çizge tabanlı yöntemlerle yüksek başarımlı sonuçlar elde edilmiştir (Moro ve diğ, 2014; Agirre ve diğ, 2014).

1.2.4 Diğer çalışmalar

İngilizce gibi çekimli dillerin dışında kalan ve kaynakların kısıtlı olduğu diller için yapılan çalışmalar da günümüzde ivme kazanmıştır. Kaynak kısıtı olan diller için yürütülen bazı ortak çalışmalar bulunmaktadır. BabelNet Roma Sapienza Üniversitesi dilbilim laboratuvarında geliştirilmiş olan çok dilli anlamsal bir ağ ve ontolojik bir yapıdır (Navigli ve Ponzetto, 2010, 2012). BabelNet yapısı, geniş bir web ansiklopedisi olarak tanımlanan Wikipedia ile sıklıkla kullanılan bir hesaplamalı sözlük olan WordNet arasında kurulan bağlantılar ile otomatik olarak oluşturulmuştur. İki kaynak arasındaki ilişkilendirme otomatik olarak eşleştirme yapılarak sağlanmıştır. Kaynak kısıtı olan dillere ilişkin bilgi ise, makine öğrenmesi yöntemlerinin kullanılması ile sisteme dahil edilmiştir. Çalışmanın sonucunda farklı diller için pek çok anlamsal ilişkiyi barındıran ve kavramları içeren ansiklopedik bir sözlük ortaya çıkmıştır.

SABG alanında yapılan önemli çalışmalardan bir tanesi Senseval/Semeval toplantıları kapsamında periyodik olarak gerçekleştirilmektedir. Senseval çok katılımcılı bir SABG değerlendirme çalışmasıdır. 1998 yılında İngiltere Sussex'te

Senseval-1 adı altında ilki düzenlenen toplantıda İngilizce, Fransızca ve İtalyanca dilleri için çalışılması hedeflenmiştir. 2001 yılında Toulouise'da gerçekleştirilen Senseval-2 çalıştayında ise 12 farklı dil yapılan çalışmalar kapsamına alınmıştır. Senseval-3 2004 yılında Barcelona'da gerçekleştirilmiş ve yapılan çalışmalar; SABG, anlamsal rollerin tanımlanması, çok dilli işaretleme, mantıksal biçimler ve alt sınıf çıkarımı gibi başlıkları içeren 14 bölüme ayrılmıştır. Senseval-3 çalıştayını izleyen toplantılar Semeval adını almıştır. Bunlardan ilki Semeval-2007 adı altında Prag'da gerçekleştirilmiş, sistemlerin değerlendirilmesi ve metinlerin anlamsal analizini kapsayan 18 bölüme ayrılmıştır. 2010 yılında Uppsala'da gerçekleştirilen çalıştayda anlamsal analiz çalışmalarını içine alan 18 bölüm yer almıştır. 2012 yılında Montreal'de yapılan konferans *SEM (StarSEM) adının kullanıldığı ve NAACL (Annual Conference of the North American Chapter of the Association for Computational Linguistics) konferansı ile birlikte yapılan ilk birleştirilmiş sözlüksel ve hesaplamalı anlambilim çalıştayı olmuştur. 2012'de yapılan çalıştayda SABG konusu yer almamış ancak konuyla ilgili çalışmaların Semeval-2013 kapsamında yapılması planlanmıştır. Georgia, ABD'de gerçekleştirilen Semeval-2013 ise NAACL 2013 ile birleştirilmiştir. Toplantı hesaplamalı anlambilim çalışmalarının yer aldığı 13 farklı bölüme ayrılmıştır. Semeval-2014 yirmi beşincisi düzenlenen Coling-2014 (International Conference on Computational Linguistics) konferansı ile birlikte gerçekleştirilmiştir. Aynı zamanda Dublin'de gerçekleşen bu toplantı *SEM 2014; ikinci sözlüksel ve hesaplamalı anlambilim çalıştayı adı altında yapılmıştır. Şekil 1.1'de SemEval çalıştayları için verilen taslak yapısı görülmektedir.



Şekil 1.1 : SemEval çalıştay taslağı.

1.3 Tezin Katkısı

Kaynak araştırmaları sonucunda, Türkçe sözcüklerin belirsizliğinin giderilmesi konusunda bugüne kadar yapılmış olan çalışmaların yetersiz olduğunu görülmüştür. Bu tez çalışmasıyla başarım oranı kabul edilebilir derecede yüksek bir yöntem ve algoritma geliştirilmiştir.

Türkçe sözcüklerin anlam belirsizliklerinin giderilmesi amacıyla bir yöntem geliştirmek üzere yaptığımız çalışmalar aşağıda sırasıyla tanıtılmaktadır:

Anlam belirsizliğini gidermek üzere kullanılan yöntem sınıfları, bilgi tabanlı yöntemler, derlem tabanlı ve melez yöntemlerden oluşmaktadır. Bu yöntemlere ilişkin ayrıntılı açıklama ikinci bölümde yer almaktadır. Derlem tabanlı yöntemlerin ağırlıklı olarak kullanıldığı çalışmamızda, denetimli ve denetimsiz yöntemler olarak anılan yöntemler üzerinde de çalışmalar yapılmıştır. Denetimli ve denetimsiz yöntemlerin birbirine göre üstünlük ve eksikliklerini görebilmek ayrıca Türk dili için kullanılabilirliklerini değerlendirmek amacıyla bu yöntemler araştırılmış ve denenmiştir. Sözü edilen yöntemler kullanılarak Türkçe sözcüklerin belirsizliğini gidermek üzere deneyler yapılmıştır. Bu çalışmalarımız kapsamında yapılanlar sırası ile aşağıda verilmektedir.

1.3.1 Denetimli yöntemler

Bir sözcüğün anlamının belirlenmesinde, kendinden önce ve sonra gelen sözcüklerden faydalanılmaktadır. Anlam belirsizliği olan sözcüğün komşusu önceki ve sonraki sözcükleri kapsayan alana pencere adı verilmiştir. Pencere içindeki sözcükler ile seçilen sözcüğün anlamını belirlemek üzere iki yöntem geliştirilmiştir. Farklı özellik grupları ve yaklaşımları içine alan bu çalışmalara ilişkin kısa açıklamalar aşağıda verilmektedir:

1. Bir pencere içinde yer alan sözcüklerin niteliklerinin seçilen sözcüğün anlamına etkisi incelenmiştir ve etkin olduğu sonucuna varılmıştır. Çalışmada öncelikle farklı pencere genişlikleri sınandıktan sonra pencere boyunun uygun değeri -4 … +4 olarak saptanmış ve yapılan çalışmalarda kullanılmıştır. *Niteliklerin Kazandırdığı Anlamlar* (NKA) adını verdiğimiz bu çalışmanın sonucu INES2012 konferansında sunulmuştur (İlgen ve diğ, 2012).

2. İkinci yöntemde, içinde seçilmiş sözcüğün de geçtiği "n" tane metin ele alınmıştır. Bu işleme *Sözcük Kesesi* (SK) oluşturma adı verilmiştir. Seçilen sözcüğün öncesi ve sonrasını kapsayan alanda bu kesede yer alan sözcüklerin varlıklarına bakılarak seçilmiş sözcüğün anlamının çıkarılıp çıkarılamayacağı araştırılmıştır. Bu çalışma sırasında, sözcük kesesine dahil edilecek sözcüklerin sayısı için eşik değeri belirlenmeye çalışılmıştır. Sözcüklerin kullanım sıklıkları dikkate alınarak değişen özellik sayısının sonuçlara etkisi incelenmiştir. Sorunun yanıtı olumlu olmuş ve çalışmamız yayınlanmıştır (İlgen ve diğ, 2013). Bu yönteme *Birlikteliklerin Kazandırdığı Anlamlar* (BKA) adı verilmiştir.
3. İlk iki aşamadaki çalışmalar sabit pencere boyunda denenmiştir. Üçüncü aşamada, SK özelliklerinde pencere boyunun anlam belirlemedeki etkisi araştırılmış, isim ve eylem grupları için -5…+5 aralığının en uygun pencere boyu olduğu sonucuna ulaşılmıştır. Çalışmanın sonuçları yayınlanmıştır (İlgen ve diğ, 2013).

NKA ve BKA yöntemleri üzerindeki çalışmalarımızın sonuçları karşılaştırılmıştır. NKA yönteminin başarımı BKA yöntemine oranla daha başarılı bulunmuştur. Yapılan çalışmalara ek olarak iki yöntem birlikte kullanılmış ve denenmiştir. Bu durumda başarım NKA'nın başarımını da geçmiştir.

Türkçe için daha önce aynı konuda Orhan ve diğ. (2007) tarafından yapılmış olan çalışmanın sonuçları ile bizim çalışmamızın sonuçları ayrıca karşılaştırılmıştır. Aynı veri kümesi üzerinde yaptığımız karşılaştırmalarda, bizim geliştirdiğimiz yöntemler daha başarılı olmuştur.

1.3.2 Denetimsiz yöntemler

Denetimli yöntemler kullanılarak, seçilen bir sözcüğün anlamını belirleyebilmek için, tüm sözcüklerinin anlamları ve nitelikleri belirtilmiş bir derlem gerekmektedir. Böylesi bir derlemin Türkçe için var olduğu söylenemez. Deneylerimizde kullandığımız ODTÜ-Sabancı derlemi yeterli bir derlem sayılamaz.

Bu nedenle, anlam belirsizliğinin giderilmesi için denetimsiz yöntem arayışına geçilmiştir. Araştırmalarımızın sonucu olarak, farklı bir alanda kullanılan HyperLex algoritması bu amaçla uyarlanmıştır.

Bu yöntemde, içinde seçilmiş sözcük bulunan metinler üzerinde çalışılmış; tüm örnek metinlerin içindeki sözcüklerin birbirleri ile olan ilişkileri bir çizge biçiminde hazırlanmıştır. Her sözcüğün diğer sözcükler ile olan ilişkisine bir ağırlık değeri karşı düşürülmüştür. Çizge üzerinde belli sözcükler ağırlık kazanmaktadır. Ağırlıklı sözcüklerin gerçek anlamları sözlükten bulunmakta ve bu sözcüğe atanmaktadır. Bu ağırlıklı sözcüklerin anlamlarına bakılarak seçilmiş sözcüğün anlamına ulaşılmaktadır. Bu yöntemde, tezimizin katkısı, ağırlıklı sözcüklere, sözlükten anlam atama ve daha sonra bu bilgilerden yararlanarak, seçilen sözcüğün gerçek anlamının belirlenmesidir.

Geliştirdiğimiz bu denetimsiz yöntemin başarımı, denetimli yöntemlere yakın olmuştur.

1.4 Tezin Düzeni

Birinci bölüm tez çalışmasının tanıtımı, literatür araştırması ve tezin kendi alanındaki katkılarını anlatmak için ayrılmıştır.

İkinci bölümde sözcük anlam belirsizliği konusunda ayrıntılı bilgiler verilmiş, DDİ'nin değişik alanlarında sözcük anlam belirsizliğinin karşımıza nasıl çıktığı gösterilmiştir. Değişik bakış açılarından sözcük anlam belirsizliğinin tanımları verilmiştir. İlk bakış açısına göre yapılan sınıflandırmada SABG için faydalanılan denetim seviyesi göz önünde bulundurulmuştur. Diğer sınıflandırma ise SABG probleminin çözümüne olan yaklaşım ve kapsamla ilişkilidir. SABG sistem sınıfları olarak verilen sınıflandırmada kullanılan yaklaşımlar; önceden seçilmiş bir sözcüğün belirsizliğinin araştırılması ya da metin içindeki tüm sözcüklerin belirsizliğinin araştırılması seçeneklerini içine almaktadır.

Yine bu bölümde sözcük anlam belirsizliği gidermede genel olarak kullanılan yöntemler kısaca tanıtılmıştır. Bunun ardından yöntemlerin birbirlerine göre üstünlük ve eksiklikleri tartışılmıştır. Bu yöntemlerin başarımları da açıklanmıştır.

İkinci bölüm kapsamında sözcük anlam belirsizliği giderme için kullanılan denetimli ve denetimsiz yöntemler tanıtılmış ve bunların kullanım alanlarından örnekler verilerek, sonuçlar karşılaştırılmıştır.

Üçüncü bölümde bu çalışmada denetimli yöntemler kapsamında yapılmış olan çalışmalara yer verilmiştir. Bu bölümde çalışmamızda hem denetimli hem de

denetimsiz yöntemler dahilinde yapılan çalışmalarda kullanmakta olduğumuz özel Türkçe derlemin hazırlanma aşamalarının detaylı anlatımı yer almaktadır. Aynı zamanda konumsal özellikler ve sözcük kesesi özellikleri kullanılarak yaptığımız çalışmalara yer verilmektedir. Bu çalışmalar etkin özelliklerin bulunması, en uygun pencere boyunun ve özellik sayısının saptanması, SK ve KÖ özellik gruplarının etkinliğinin birlikte ve ayrı kullanımda sınanması ve diğer çalışmaları içine almaktadır.

Üçüncü bölümde bu tez kapsamında yapılmış olan çalışmaların sonuçları kendi aralarında ve benzer çalışmaların sonuçları ile karşılaştırılmıştır. Sözcük anlam belirsizliği giderme konusunda kullanılabilecek yöntem ve algoritmalar değişik bakış açılarından değerlendirilmiştir. Çalışmamızın, en yakın benzer bir çalışma karşılaştırılabilir olması için, Orhan (2006) tarafından yapılmış olan çalışmada kullanılmış olan ODTÜ-Sabancı derlemi üzerinde, kendi geliştirdiğimiz ilk yöntem çalıştırılmıştır. Ancak bu çalışmayı yapabilmek için ODTÜ-Sabancı derlemi, yöntemimiz ile uyumlu çalışabilir hale getirilmiştir. Bu çalışmanın sonunda, bizim yöntemimizin başarısı daha yüksek olmuş, dolayısıyla Orhan'ın (2006) başarımını geçmiştir.

Üçüncü bölümde tanıtılan çalışmalar denetimli yöntemler kapsamında gerçekleştirilmiş olan çalışmalardır. Denetimli olmalarının doğal sonucu olarak yoğun insan emeği gerektirmektedir. Bu nedenle, denetimsiz bir yöntem çalışmasına yönelinmiş ve çizge temelli yeni bir yöntem geliştirilmiştir. Bu yöntemin başarımı, özel derlem üzerinde ölçülmüş ve denetimli yöntemler seviyesinde başarılı bulunmuştur. Dördüncü bölüm bu tez çalışması kapsamında sözcük anlam belirsizliğini gidermek için geliştirilmiş olan denetimsiz algoritmanın tanıtımına ayrılmıştır. Geliştirilmiş olan algoritmamızı sınamak ve sonuçları değerlendirmek için hazırlanan Türkçe derlem kullanılmıştır. Beşinci bölümde ise çalışmamıza ilişkin sonuç ve değerlendirmeler yer almaktadır.

2. ANLAM BELİRSİZLİĞİ KAVRAMI VE GİDERME YÖNTEMLERİ

Doğal dillerin sık gözlenen bir özelliği olan anlam belirsizliği, DDİ alanında yapılan çalışmalar kapsamında ele alınan metinlerde, bir sözcüğün birden fazla anlamla eşleşebilmesi durumunda ortaya çıkan belirsizlik türüdür. Belirsizlik kavramı insanlar arası ve yüz yüze iletişimde, konuşmanın kapsamı ve akışı, önceki deneyimler ve edinilen diğer bilgilerin insan bilişsel süreçleri tarafından kullanılması ile ortadan kaldırılmaktadır. Kişi dinlediği bir cümle içinde geçen birden fazla anlama sahip sözcüğün doğru anlamını, bilgisi ve geçmiş deneyimleri yardımıyla anlayarak seçmekte ve diğer anlamları elemiş olmaktadır. Anlam belirsizliği giderme, insanlar tarafından doğal olarak gerçekleştirilen bu işlemin, bilgisayar yazılımlarının kullanımıyla sayısal ortamda gerçekleştirilmesidir. Doğal dillerdeki sözcüklerin birden fazla anlamlarının olması ve kullanıldıkları tümce içindeki gerçek anlamlarının çıkarılması *Bilgisayarlı Çeviri* (BÇ), *Özet Çıkarma* (ÖÇ) ve *Anlam Çıkarma* (AÇ) işlemlerinde önemli olmaktadır. Anlam belirsizliği konusuna gereksinim duyan alanlar aşağıda tanıtılmıştır:

- **Özet çıkarma**: Özet çıkarmanın amacı bilgisayarlar yardımıyla metindeki önemli noktaların göz önünde bulundurularak metin boyunun küçültülmesidir. Sayısal ortamdaki bilgi ve veri boyutunun çok büyük bir hızla artması, aynı zamanda konunun okuyuculara hızlı biçimde aktarılmasının amaçlanması konuya olan ilgiyi arttıran unsurlardır.

  Özet çıkarma konusunda genellikle iki yaklaşım tercih edilmektedir: Bunlardan ilki *çıkarma*, ikincisi ise *soyutlama* yaklaşımı olarak bilinmektedir *(extraction, abstraction)*. Çıkarma yönteminde metindeki sözcük, deyiş ve tümcelerin bir alt kümesi seçilerek özet oluşturulmaktadır. Soyutlama yaklaşımında ise öncelikle metinden bir anlamsal temsil oluşturulmakta ve doğal dil üretme teknikleri kullanılarak bir insanın çıkarımına yakın bir özet üretilmektedir. Dolayısıyla ikinci yöntemde anlamların belirlenmesi önemli olmakta ve özet çıkarma aşamasından önce çözülmesi gerekmektedir.

- **Metin Anlama:** Özellikle internetteki hızlı gelişmeler sonunda sanal ortamdaki belge miktarı üstel biçimde artmaktadır. Çok sayıdaki belge arasında gerçekten erişmek istediğimiz belgeyi bulabilmek sorun olmaktadır. Bu sorunu gidermek için belgelerin etiketlenmesine çalışılmaktadır. Etiket içinde belgenin konusu, konunun kahramanı veya kahramanları, olay yeri ve olay zamanı gibi bilgiler yer almaktadır. Metin anlamı çıkarma sürecinde etiket bilgileri çıkartılmaya çalışılmaktadır. Örneğin, bir belgeye ilişkin etiket şöyle olabilir:

  Konu: *Nutuk*
  Özne: *Mustafa Kemal Atatürk*
  Yer: *TBMM*
  Zaman: *10 Nisan 1920*

  Atatürk'ün Nutku bilgisayar tarafından metin anlama bağlamında değerlendirildiğinde yukardaki etiketin üretilmesi beklenmektedir. Metin anlamı çıkarmada metinlerin boyları bir paragraf olabileceği gibi yüzlerce sayfalık bir kitap da olabilmektedir.

- **Diller arası çeviri:** Diller arası bilgisayarlı çeviri dendiğinde ilk akla gelen, bir dilde yazılmış bir metni diğer bir dilde metne çevirmektir. Bilgisayarlı çeviri çalışmalarında sözcük anlam belirsizliği giderme en fazla gereksinim duyulan konulardan biri durumuna gelmiştir. Metnin hedef dile birebir çevrilebilmesi için kaynak dilin çok iyi çözümlenmesi ve birden fazla anlama sahip sözcüklerde belirsizliğin giderilmesi gerekmektedir. Kaynak dilde kullanılan sözcüğün referans ettiği anlam bilgisi hedef dilde karşılık gelen sözcüğü bulmak açısından gereklidir. Diller arası çeviride belirsizliğin önemini vurgulamak için aşağıdaki örneği verebiliriz:

  Kitabın arka <u>yüzünde</u> bir resim vardı.

  Yukardaki örnekte belirsizliğini gidermek istediğimiz sözcük "yüzünde" sözcüğüdür. Kaynak dil olan Türkçeden İngilizceye doğru çevirinin yapılabilmesi için "yüz" sözcüğünün bu cümlede geçerli olan anlamının belirlenmesi gerekmektedir. Türkçede "yüz" sözcüğünün olası anlamları şunlardır;

  1. Doksan dokuzdan sonra gelen sayının adı

2. Kere, kat vb. kelimeler ile birlikte kullanılarak yapılan işin çokluğunu abartılı bir biçimde anlatan söz
3. Başta, alın, göz, burun, ağız, yanak ve çenenin bulunduğu ön bölüm, sima, çehre, surat
4. Yüzey . "Suyun yüzünde."
5. Kesici araçlarda ağız . "Bıçağın keskin yüzü."
6. Bir kumaşın dikiş sırasında dışa getirilen gösterişli bölümü
7. Yorgana ve yastığa geçirilen kılıf
8. Bir şeyin görünen bölümünde kullanılan kumaş . "Yorgan yüzü. Kanepenin yüzü."
9. Birinin görülegelen veya umulan hoşgörürlüğüne güvenilerek gösterilen cüret . "Ne yüzle? Yüzü olmamak."
10. Nedeniyle, sebebiyle . "Bu yüzden Fuat Köprülü ile çatışmaya başlamışlardı gazetelerde." - Y. Z. Ortaç
11. Yan, taraf
12. Bir yapının dışa bakan düşey yüzeylerinin her biri . "Ön yüz. Yan yüz. Arka yüz."
13. Utanma. "Adamda yüz yok ki!"

Verilen örnekte görüldüğü üzere cümlede geçen "yüz" sözcüğü, TDK sözlüğünde verilen karşılıklarından on ikincisi ile eşleşmektedir. Yani, bir nesnenin dışı veya dışa bakan yüzeyi, diğer bir deyişle kitabın dış yüzeyi olarak değerlendirmektedir.

- **Anlam Çıkarma:** Bir cümle veya bir paragrafın ne demek istediğini anlamaya, anlam çıkarma denilmektedir. Örneğin cümlemiz şöyle olsun:

  "Şubenizdeki 12345 sayılı hesabımdan, Ayşe Güzel'in ABC Bankasındaki 54321 sayılı hesabına 125,50 TL havale etmek istiyorum. Gereğini dilerim. Ahmet Şık"

  Bu yazıyı bilgisayarın anlayıp eyleme dönüştürmesine anlam çıkarma adı verilmektedir. Belirsizliğini gidermek istediğimiz sözcük "havale" sözcüğü

olsun. TDK sözlüğüne bakıldığında, bu sözcüğün anlamları aşağıdaki gibi sıralanmaktadır:

1. isim Bir işi bir başkasının sorumluluğuna bırakma, ısmarlama, devretme: "Bütün belgelerin bakanlığa havalesi gerekiyor."

2. Banka, postane vb. aracılığıyla gönderilen para : "Ay başında havaleyi postaneye yatırdım."

3. Postane, banka vb. aracılığıyla para gönderildiğinde gönderenle alacak olanın adları ve para miktarı yazılı kâğıt, havale kâğıdı, havalename

4. Genellikle çocuklarda görülen, ateşli veya ateşsiz olan çırpınma nöbetleri: "Yeşil kadifeden dikilmiş yarım baklava şeklinde muska çok ufakken üzerine gelen havaleden Fikret'i kurtarırmış." - R. Enis

5. Bir arsayı çevirmek, kapamak için çekilen perde veya duvar : "Bu ufacık binayı bahçe ve bostan, ahır ve selamlık gibi müştemilatından birtakım duvarlar, bölmeler, havalelerle öyle bir ayırtmış..." - Y. K. Karaosmanoğlu

6. Yüksek ve büyük bir görünüşü olma

Anlamı çıkarılacak olan cümlenin içinde yer aldığı dünya belli ise, sözcük anlam belirsizliğini gidermek kolaylaşmaktadır. Örnek cümlede bankaya verilen bir emir olduğu göz önünde bulundurulduğunda *havale* sözcüğünün doğru karşılığının üçüncü anlam olduğu kolayca söylenebilmektedir.

2.1 Anlam Belirsizliği Giderme Yöntemleri

Sözcük anlam belirsizliğinin giderilmesindeki yöntem sınıfları; bilgi tabanlı, derlem tabanlı ve melez yöntem sınıflarından oluşmaktadır. Bilgi tabanlı yöntemlerde sözlükler, ontolojiler gibi harici kaynaklar anlam belirsizliği gidermede kullanılırken, derlem tabanlı yöntemlerde gerekli olan bilgi derlemden makine öğrenmesi yöntemleri ve algoritmalar ile öğrenilmektedir.

SABG alanında kullanılan yöntemler; 1^0 Bilgi tabanlı yöntemler, ve 2^0 Derlem tabanlı yöntemler, ve 3^0 Melez yöntemler olmak üzere üç genel başlık altında ele alınmaktadır. Derlem tabanlı yöntemler ise faydalanılan denetim seviyesine göre aşağıdaki sınıflara ayrılmaktadır.

- Denetimli Yöntemler
- Yarı Denetimli Yöntemler
- Denetimsiz Yöntemler

Bu yöntemlerle ilgili açıklamalar aşağıda verilmektedir.

2.1.1 Bilgi tabanlı yöntemler

Bilgi tabanlı yöntemler, anlam belirsizliği gidermede, bilgisayarla okunabilir sözlükler, anlam envanterleri ve eş anlamlılar sözlüğü gibi farklı kaynakların kullanımını esas alan yöntemlerdir. Sözü edilen yöntemler dahilinde WordNet (Miller ve diğ, 1990) en sık kullanılan kaynaklardan biri olarak bilinmektedir. Bilgi tabanlı yöntemler aşağıda verilen dört sınıfa ayrılmaktadır.

2.1.1.1 Bağlam ve sözlük anlam örtüşmesi yöntemleri

Lesk Algoritması (LA), bilgisayarla okunabilir sözlük kullanarak SABG için geliştirilen ilk yöntemdir (Lesk, 1986). Algoritma ile temelde gerçekleştirilen, sözcük kesişimlerinin anlam belirsizliği gidermede kullanılmasıdır. İlerleyen bölümde anlatıldığı gibi kesişimleri değerlendirmeye alınan sözcükler, metinde geçen sözcüklerin anlam tanımlamaları olabileceği gibi, metinde geçen sözcükler ve sözlük tanımları arasındaki örtüşme de olabilmektedir. En fazla örtüşmenin ya da sözcük kesişiminin sağlandığı anlam belirsiz sözcüğe ilişkin doğru anlam olarak seçilmektedir.

Yöntem ilk bilgi tabanlı yöntemlerden olmakla birlikte derlem tabanlı yöntemlere temel oluşturduğu düşünülmektedir. Lesk yöntemine ilişkin farklı yaklaşımlar bulunmaktadır. Algoritmanın özgün versiyonunda sözcük anlamları arasındaki kesişim göz önünde bulundurulmuştur. W_1 ve W_2 sözcüklerinin sahip olduğu N_{W1} ve N_{W2} anlamları sözlükte tanımlı olmak üzere, sözcüklerin eşleşebileceği anlamlar W_1^i ve W_2^j (i=1,2, …… N_{W1} ve j=1,2, …… N_{W2}) olarak verilmektedir. Verilen iki sözcük için en fazla örtüşmenin sağlandığı anlam çifti seçilmektedir. Lesk algoritması ile anlam belirsizliği içeren önceden seçilmiş ve anlamları işaretlenmiş sözcük çiftleri üzerinde *Oxford Advanced Learner's Dictionary* kullanılarak yapılan çalışmada %50-70 arasında tutturma değeri elde edilmiştir (Lesk, 1986).

Algoritma ilk olarak geliştirildikten sonra çeşitli varyasyonları ortaya konmuştur. Yapılan bu çalışmalarda; i) Anlam belirginleştirmede iki sözcükten farklı sayıda

sözcük değerlendirmeye alınmıştır, ii) Bir bağlamda verilen her sözcük bağımsız ele alınarak bu sözcüklerin sözlük örtüşmelerine bakılmıştır, iii) Sözcük anlamları için anlamsal uzaydan faydalanılarak anlamsal olarak ilişkili sözcükler de çalışmaya dahil edilmiştir.

Bir metinde iki sözcükten daha fazla sayıda sözcüğün anlamı belirginleştirilmek istendiğinde ve Lesk algoritmasının özgün versiyonu uygulandığında, sınanması gereken olası anlam eşleşmelerinin çok fazla sayıda olduğu görülmüştür. Bu sorunu aşmak üzere farklı yaklaşımlar önerilmiştir. Sorunun üstesinden gelmek için önerilen çözümlerden bir tanesi *Benzetilmiş Tavlama* yönteminin kullanılmasıdır (Cowie ve diğ, 1992). Aşırı anlam kombinasyonu sorununun üstesinden gelmenin bir diğer yolu ise *Basitleştirilmiş Lesk* algoritmasının kullanılmasıdır. Bu yaklaşımda tüm sözcüklerin eş zamanlı anlam örtüşmelerini değerlendirmeye almak yerine her sözcük bağımsız kabul edilmekte ve sözlük eşleşmelerine bakılmaktadır. Lesk algoritmasının bir diğer versiyonu ise *Uyarlanmış Lesk* yöntemidir (Banerjee ve Pedersen, 2002). Bu yaklaşımda doğru anlama karar verilirken sadece hedef sözcük değil aynı zamanda sözcükle ilintili diğer sözcük tanımlamaları da dikkate alınmaktadır.

Lesk algoritması, SABG için sahip olunan tek kaynak sözlük tanımları olduğunda tercih edilirliği yüksek olan bir yöntemdir. Yöntemin tüm varyasyonları değerlendirildiğinde özgün algoritmaya kıyasla etkinlik ve doğruluk anlamında *Basitleştirilmiş Lesk* yönteminin en fazla artışı sağladığı görülmüştür. Anlamsal uzay bilgileri ve anlamsal olarak ilişkili sözcüklerin kullanılması da başarımı önemli oranda arttırmıştır. Aynı zamanda anlamları işaretli derlemlerin kullanımı mümkün olduğunda, derlemden öğrenilen bilgi yönteme uyarlanabilmekte ve algoritma başarımını arttırmaktadır.

2.1.1.2 Anlamsal ağlar üzerinde benzerlik ölçütlerini kullanan yöntemler

Anlamsal Benzerlik (AB) yöntemleri, sözcüklerin anlamsal ilgililik derecelerinin dikkate alındığı yaklaşımlardır. Doğal dillerin yapısı gereği bir bağlamda yer alan sözcüğün aynı bağlamdaki diğer sözcüklerle anlamsal olarak bağlantılı olduğu bilinmektedir. Bağlam içerisindeki sözcüklerin anlamsal yakınlığını ve ilgililik derecesini hesaplamak için benzerlik ölçütlerinden faydalanılmaktadır. Bu ölçütler bölgesel ya da evrensel kapsamda ele alınabilmektedir. Lesk algoritmasında olduğu

gibi, bu yaklaşımda da iki sözcükten daha fazla sözcüğün değerlendirmeye alınması durumunda hesaplama karmaşıklığı artmaktadır. Sorunun çözümü de Lesk yöntemi ile benzerlik göstermekte ve her sözcük ayrı olarak ele alınmaktadır (Agirre ve Rigau, 1996).

İki farklı sözcüğün anlamsal benzerliğinin ölçülmesinde kullanılmak üzere çeşitli ölçütler tanımlanmıştır. Budanitski ve Hirst (2001) bu konuda çok kapsamlı bir çalışma hazırlamıştır. WordNet üzerinde benzerlik ölçütlerini hesaplayan bir yazılım aracı da geliştirilmiştir (Patwardhan ve diğ, 2003).

WordNet yapısı üzerinde özellikle etkin olduğu kanıtlanmış ölçütler tanımlanmıştır. Bu ölçütlerde kavramlar genellikle sisteme girdi oluşturmakta ve çıktı olarak anlamsal benzerlik değeri elde edilmektedir. Bu ölçütler arasından öne çıkanlar aşağıda verilmiştir:

- Sistem girdilerini oluşturan sözcükleri içeren eş anlamlı sözcük kümeleri arasındaki en kısa mesafe hesaplanmıştır. Bu değer taksonomiye ilişkin derinliğe göre normalize edilmiştir (Leacock ve diğ, 1998).
- Hirst ve St-Onge (1998), tanımladıkları benzerlik ölçütüne bağlantı yönelim bilgisini eklemiştir. Kavramlar arasındaki mesafe önemli olmakla birlikte yön değişiminin sık olması tercih edilmemektedir.
- Resnik (1995), değerlendirilen kavramın kapsamlı bir derlem içerisindeki gözlenme olasılığını esas alarak ne derece özellikli olduğu konusunda bir bağıntı ortaya koymuştur.
- Mihalcea ve Moldovan (1999), farklı sözcük türlerini kapsayan hiyerarşiler arasındaki benzerliği ifade eden bir bağıntı ortaya koymuştur.
- Agirre ve Rigau (1996), kavramsal derinlik kavramını tanımlayarak, anlamsal bir hiyerarşik yapının başında bulunan kavram ile bu kavramla aynı bağlamda yer alan sözcüklerin örtüşmelerini değerlendirmiştir.

2.1.1.3 Seçimsel önceliklerin kullanıldığı yöntemler

Seçimsel Öncelikler (SÖ) yaklaşımında sözcük tipleri arasındaki ilişkiler bulunarak ortak anlam belirlenmeye çalışılmaktadır. Birbiri ile uyum içinde olmayan anlamlar elenmektedir. Bu yaklaşımın ardındaki temel düşünce belirli bir sözdizimsel yapı içinde birlikte gözlenen sözcük çiftleri üzerinden anlamların ayrıştırılmaya çalışılmasıdır. Örneğin yüzmek sözcüğü derisini yüzmek ya da denizde yüzmek öbeklerinde olduğu gibi öznesine göre farklı anlamlara gelebilmektedir. *Yemek-*

Yiyecek ve *İçmek-İçecek* gibi anlamsal kısıtların tanımlanmasıyla, verilen yapıya uymayan anlamlar ayrılmaktadır. Örneğin "*Bu adam benim yüz bin liramı yedi*" cümlesindeki "*yüz bin lira"*, yemek sözcüğünün nesnesi yiyecek olduğu için verilen yapıya uymamaktadır. Diğer bir deyişle seçimsel öncelikler sözcük sınıfları arasındaki ilişkilerden faydalanmaktadır. Seçimsel önceliklerin öğrenilmesi oluşturulan taksonomiler yardımıyla, sözcük gözlenme sıklıkları, bilgi kuramı ölçütleri ve sınıflar arası ilişkiler esas alınarak gerçekleştirilmektedir. Brockmann ve Lapata (2003), çalışmalarında bu yaklaşımlara ilişkin detaylı çalışma sonuçlarını sunmuştur.

2.1.1.4 Sezgisel yöntemler

Sezgisel Yöntemler (SY), doğal dile ilişkin farklı özellikleri kullanarak hesaplamaların yapıldığı yaklaşımları içine alan gruptur. Kullanılan sezgisel özelliklerden bazıları: *En sık gözlenen sözcük anlamı*, *bağlamda tek anlam* ve *söz öbeği için tek anlam* kabulü yaklaşımlarından oluşmaktadır. En sık kullanılan sözcük anlamı yaklaşımında diğer anlamlara göre daha baskın kullanılan sözcük anlamı esas alınmaktadır. Sözcük anlam sıklıkları incelendiğinde, bir ya da çok az sayıda sözcük anlamının derlemlerde baskın olduğu ve geri kalan anlam sıklıklarında düşüş gözlendiği bilinmektedir. Bu doğrultuda, sözcük anlamlarının Zipf (1949) anlamlar yasası ile ortaya konan bağıntıya uygun bir dağılım göstermesi, sözcük anlam sıklıkları bilindiğinde uygun anlam atamasının gerçekleştirilmesinde kullanılabilmektedir. Bağlamda tek anlam yaklaşımında ise verilen bağlamda konu bütünlüğü içerisinde sözcüğe ilişkin aynı anlamın kullanıldığı kabul edilmektedir. Söz öbeği için tek anlam kabulü de benzer varsayımı söz öbekleri için ele almaktadır.

2.1.2 Derlem tabanlı yöntemler

2.1.2.1 Denetimli yöntemler

SABG yöntemi kapsamında kullanılacak olan özellikler ve metinden elde edilen bilgilerin yanında kullanılacak yöntemin saptanması da önemli aşamalardan birini oluşturmaktadır. Kullanılan genel yaklaşımlar Bayes olasılık algoritmaları (Gale ve diğ, 1992c; Yarowsky, 1992; Leacock ve diğ, 1993; Bruce ve Wiebe, 1994; Pedersen ve Bruce, 1997a), sinir ağları, bellek/örnek tabanlı öğrenme (Escudero ve diğ, 2000a;

Fujii ve diğ, 1998; Ng ve Lee 1996; Hoste ve diğ, 2002; Decadt ve diğ, 2004), karar ağaçları gibi yöntemleri içine almaktadır.

Denetimli yöntemler SABG alanında yaygın olarak kullanılan yaklaşımları sınıflayan genel bir gruptur. Bu yaklaşımlarda makine öğrenmesi yöntemleri anlamları etiketlenmiş bir eğitim kümesi üzerinde kullanılarak sınıflandırıcı oluşturulmaktadır. Etiketli veri kümesi belirli sayıda özelliğin bu kümedeki veriler üzerinden kodlandığı ve her örneğe ilişkin uygun anlam (sınıf) etiketini de içeren bir kaynaktır. SABG alanında denetimsiz yöntemlere oranla daha iyi sonuçlar verdiği bilinen denetimli yöntemlerden bazıları aşağıda verilmektedir:

- **Karar Listeleri (KL):** Karar listeleri sınama örneklerinin sınıflandırılmasında kullanılan sıralı kurallar kümesi olarak tanımlanmaktadır (Rivest, 1987). Karar listelerinin ağırlıklandırılmış *"if-then-else"* kuralları gibi değerlendirmek de mümkündür. Öncelikle özelliklerin çıkartılması için bir eğitim kümesi kullanılmaktadır. Eğitim aşamasının sonucunda tür (özellik değeri, anlam, derece) bilgilerine ulaşılmaktadır. Bu kuralların azalan derece değerine göre sıralanması ile karar listesi oluşmaktadır. Bir *w* örneği ve örneğin temsilinde kullanılan özelliklere bakılarak karar listesi kontrol edilmekte ve en fazla uyan özellik seçilerek anlam ataması gerçekleştirilmektedir.
- **Karar Ağaçları (KA):** Bir karar ağacı, eğitim verisini öz yinelemeli olarak bölen sınıflandırma yapısının ağaç ile temsil edildiği bir kestirim yöntemidir. Bu yöntemler ile büyük boyuttaki veri örneklerinin hiyerarşik bir yapıya dönüştürülmesi sağlanmaktadır. Ağaç yapısındaki her ara düğüm özellik değeri üzerinde bir sınamaya karşılık gelmekte, bir dal ise ilgili sınamaya ilişkin sonucu göstermektedir. Tahmin yapılması ise bir uç (yaprak) düğüme ulaşıldığında gerçekleşmektedir. ID3 algoritmasının (Quinlan, 1986) geliştirilmiş bir uyarlaması olan C4.5 (Quinlan, 1993) bilinen karar ağacı algoritmalarındandır.
- **Naive Bayes (NB):** Naive Bayes sık kullanılan olasılık yöntemlerinden bir tanesidir. Verilen bir örnek için en yüksek olasılığa sahip olanı seçen, sınıflara ilişkin koşullu olasılıkların hesaplandığı bir tümevarım algoritmasıdır (Domingos ve Pazzani, 1997). Denklem 2.1'de *s*, sözcük anlamı olmak üzere, V_w vektörü w sözcüğünün özellikleri ile kodlanmıştır. Bu özellikler; anlamsal, sözdizimsel, POS özellikleri KÖ ya da SK özelliklerinden oluşabilmektedir. Denklem 2.1'deki bağıntı, özelliklerin birbirinden bağımsız olduğu kabulü ile formül 2.2'deki gibi gösterilmektedir.

$$\hat{s} = argmax_{s \in senses} Pr(s|V_w) \quad (2.1)$$

$$\hat{s} = argmax_{s \in senses} \Pr(s) . \prod_{i=1}^{n} (V_w^i | s) \quad (2.2)$$

- **Örnek Tabanlı Öğrenme (ÖTÖ):** Örnek tabanlı sınıflandırmada eğitim kümesinden genelleme yapılması yerine veri kümesi bellekte uygun şekilde saklanarak, yeni örneklerin sınıflarının belirlenmesi sağlanır. Diğer deyişle sınıflandırma modeli örneklerden öğrenilir. Aynı zamanda sınıflandırılan örnekler de modele eklenerek güncelleme sağlanır. k-NN (k-En Yakın Komşu) yüksek başarımlı olduğu bilinen algoritmalardan bir tanesidir (Ng, 1997; Daelemans ve diğ, 1999). k-NN sınıflandırmasında yeni bir x = (x_1, x_2, ……,x_m) örneğinin temsili *m* özellik vektörüne göre yapılır ve en benzer *k* örneğe ilişkin sözcük anlamı değerlendirmeye alınır. X ve önceden saklanan her örnek için Hamming uzaklığı kullanılarak hesaplama yapılır:

$$\Delta(x, x_i) = \sum_{j=1}^{m} w_j \delta(x_j, x_{ij}) \quad (2.3)$$

  w_j, j numaralı özelliğin ağırlığı olup $x_i = x_{ij}$ olduğu durumda $\delta(x_j, x_{ij})$ değeri 0, aksi durumda ise 1 olacaktır. Örneğe en yakın k örnek değerlendirilerek anlam sınıfı belirlenecektir.
- **Birliktelik Yöntemleri (BY):** Başarımın arttırılması amacıyla farklı sınıflandırıcıların bir araya getirildiği yaklaşımlardır. Bir araya getirme yöntemleri tamamen farklı yapıdaki öğrenme algoritmalarını birleştiren yaklaşımlardır. Diğer bir deyişle dilbilgisi özellikleri, anlamsal özellikler gibi eğitim kümesine ilişkin birbirinden tamamen bağımsız ve farklı özellikler seçilebilmektedir. Birliktelik yöntemleri denetimli yöntemlerin zayıf yönlerinin olumsuzluklarını gidermek açısından tercih edilen yöntemlerdir. Bu yöntemlerle denetimli (Klein ve diğ, 2002; Florian ve diğ, 2002) ve denetimsiz yöntemleri (Brody ve diğ, 2006) içine alan çeşitli çalışmalar yapılmıştır. Tekli sınıflandırıcılar çoğunluk oylaması, olasılık karışımı, sıralama tabanlı birleştirme, ağırlıklı oylama, en fazla düzensizlik birleşimi ve AdaBoost gibi çeşitli yaklaşımlarla bir araya getirilmektedir (Navigli, 2009; Klein ve diğ, 2002).
- **Yapay Sinir Ağları (YSA):** Yapay sinir ağları yeni bilgilerin türetilebilmesi ve keşfedilmesi gibi insan beynine özgü işlevlerin otomatik olarak gerçekleştirilmesi amacıyla geliştirilen sistemlerdir. Bu sistemlerde insan beyninin öğrenme

sürecinden esinlenerek ortaya koyulan matematiksel modeller kullanılmaktadır. YSA'lar birbirine bağlı ve değerler atanmış olan birimlerden oluşmaktadır. Bu birimler ele alınan problem doğrultusunda uygun topoloji ile birbirine bağlanmaktadır. Problemlerin ifadesinde gizli birimler kullanılmakta ve öğrenme aşaması geriye doğru yayılma gibi yöntemlerin kullanımı ile gerçekleştirilmektedir. Bağlantı ağırlıklarının güncellenmesi ile başarım arttırılabilmektedir.

- **Destek Vektör Makineleri (DVM):** *Destek Vektör Makineleri* yöntemi eğitim verisinden öğrenilen bir hiper düzlem ile pozitif ve negatif örneklerin ayrıştırılması fikrine dayanmaktadır (Boser ve diğ, 1992). Hiper düzlem, destek vektörleri adı verilen ve birbirine en yakın pozisyonda bulunan pozitif ve negatif örnekler arasındaki mesafeyi maksimize edecek şekilde konumlanmaktadır. Diğer deyişle DVM'ler deneysel sınıflandırma hatasını en aza indirgerken, pozitif ve negatif örnekler arasındaki geometrik marjini maksimize etmektedir. DVM'ler ikili sınıflandırıcılar olduğu için SABG problemlerinde çoklu sınıf (hedef sözcük anlamları) adaptasyonu yapılması gerekmektedir. DVM yöntemi ile DDİ kapsamındaki metin sınıflandırma (Joachims, 1998), ayrıştırma (Collins, 2004) ve SABG (Escudero ve diğ, 2000b; Murata ve diğ, 2001; Keok ve Ng, 2002) gibi çeşitli alanlarda çalışmalar yapılmaktadır. Aynı zamanda SABG alanında diğer yöntemlere kıyasla daha iyi sonuçlar verdiği gösterilmiştir (Keok ve Ng, 2002).

2.1.2.2 Yarı denetimli yöntemler

Denetimli ve denetimsiz yöntemler arasındaki ayırım her zaman çok belirgin değildir. Yarı denetimli yöntemler olarak adlandırılan yöntemler ile en aza indirgenmiş ya da kısmi bir denetimden söz edilmektedir. Bu kapsamdaki yöntemlerde sınıflandırıcı oluşturmada kullanılan veri kümesinin sadece bir kısmı işaretlidir. Bu kapsamdaki genel yaklaşımlar iki başlık altında ele alınmaktadır (Navigli, 2009). Bu yaklaşımlar az miktarda elle işaretlenmiş verinin kullanıldığı önyükleme (bootstrapping) yöntemleri ve "Tek Anlamlı Yakın Sözcükler" yöntemlerini içine almaktadır.

- **Önyükleme yöntemleri (ÖY):** Önyükleme yönteminin amacı oldukça kısıtlı miktarda eğitim verisi kullanarak işaretli veri azlığı ve veri seyrekliği gibi zorlukların üstesinden gelmektir. Bu yöntemde, başlangıçta A gibi az miktarda

işaretli, çok büyük bir kısmı ise işaretsiz U veri kümesi ile birlikte bir veya daha fazla sayıda sınıflandırıcı bulunmaktadır. Tekrarlamalı algoritmaların sonucu olarak başlangıçtaki işaretli A kümesi artmakta, işaretsiz verinin olduğu U kümesi ise bu kümedeki örnekler için belli bir eşik değerine ulaşılıncaya kadar azalmaktadır. Başlangıçtaki işaretli küçük veri kümesi elle işaretleme veya sezgisel yöntemler ile oluşturulmaktadır (Yarowsky, 1995).

Önyükleme yöntemleri de iki bölüme ayrılmaktadır; bunlar Birlikte Eğitme (co-training) ve Kendiliğinden Eğitme (self-training) yaklaşımlarıdır. Her iki yöntemde de U etiketsiz kümesinin U′ gibi bir alt kümesi rastgele şekilde oluşturulmaktadır. Her sınıflandırıcı A etiketli eğitim verisi üzerinde eğitilmekte ve U′ kümesindeki etkiketsiz örnekleri etiketlemek üzere kullanılmaktadır. Etiketleme sonucuna göre ve bir takım kıstaslar gözetilerek en güvenilir örnekler seçilmekte ve A kümesine eklenmektedir. Bu işlem belli sayıda tekrarlanmaktadır (her tekrarda U′ kümesi U'dan belli sayıda rastgele örnek içermektedir). Bu belirlemeler ışığında birlikte eğitme ve kendiliğinden eğitme arasındaki temel fark; ilk yaklaşımda iki sınıflandırıcı kullanılıyorken, ikinci yaklaşımın kendi çıktısı üzerinde tekrar-eğitim yapılmasıdır. Bu yönteme ilişkin, birlikte eğitme için bölgesel ve konumsal özelliklerin, kendiliğinden eğitme için iki bilgi kaynağının birlikte kullanıldığı bir yöntem tanıtılmıştır (Mihalcea, 2004). Yarowsky'nin (1995) önyükleme yöntemi de kendiliğinden eğitme yaklaşımıdır ve iki sezgisel yönteme dayanmaktadır:

- Söz öbekleri için aynı anlam: Komşu sözcükler uzaklık, sıralama ve sözdizim özellikleri ışığında sözcük anlamının belirlenmesinde etkilidir (Yarowsky, 1993).
- Tüm metin için aynı anlam: Bir sözcük geçtiği konuşma ya da metnin tümünde aynı anlama sahip olmaktadır (Gale ve diğ, 1992b).

- **Tek Anlamlı Yakın Sözcükler (TAYS):** Web yapısı gereği sınırsız bir kaynak olduğundan çok büyük boyutta metni içeriğinde barındırmaktadır. Web'in etiketli veri kümelerinin oluşturulmasında bir derlem olarak değerlendirilebilmesi, veri seyrekliği probleminin çözümüne sağlayacağı katkı da göz önüne alındığında günümüzde ilgi çeken araştırma konularından biri durumuna gelmiştir. Bu ölçekte büyük bir derlemin tek anlamlı yakın sözcüklerin (tek anlama sahip eşanlamlı sözcükler) ve az sayıda çekirdek verinin yardımıyla Yarowsky'nin

(1995) yaklaşımına benzer şekilde işaretlenebileceği üzerinde durulmuştur. Bu doğrultuda SABG sınıflandırıcılarının eğitilmesinde kullanılacak verinin otomatik olarak işaretlenmesi mümkün olacaktır.

2.1.2.3 Denetimsiz yöntemler

Denetimli yöntemlerde karşılaşılan en büyük zorluk elle işaretlenmiş büyük ölçekli verinin azlığından kaynaklanmaktadır. Birinci bölümde değinilen bu sorun Bilgi Edinim Darboğazı olarak anılmaktadır (Gale ve diğ, 1992a). Anlam belirsizliği gidermede bilgi kaynaklarının kısıtlı olması ve çok büyük ölçekteki kaynakların elle işaretlenmesindeki zorluklar gibi problemler göz önünde bulundurulduğunda, denetimsiz yöntemlerin kullanımı son dönemlerde tercih edilir olmuştur. Denetimsiz yöntemlerdeki temel yaklaşım bir sözcüğün aynı anlamlarının benzer komşuluklara sahip olacağı fikridir. İşaretsiz derlemlerde sözcüklerin kümelenmesi ile anlamlar çıkartılabilir, yeni sözcükler ise bulunan kümeler doğrultusunda sınıflandırılabilmektedir. Denetimsiz yöntemler farklı uygulamalarının yanında, en yalın haliyle etiketli veri kümesine, sözlük, ontolojiler ve eş anlamlılar sözlüğü gibi bilgisayarla okunabilir kaynaklara gereksinimin bulunmadığı yaklaşımlardır. Tam denetimsiz bir SABG sisteminin ise başlıca olumsuz yönü, sözlük kullanımı olmamasından dolayı bulunan anlamların sözlük anlam envanteri ile eşleşmesinin sağlanamamasıdır.

SABG sistemleri temel tanımıyla hedef sözcüğe belirli bir anlam etiketinin atandığı sözcük etiketleme yöntemleri olarak bilinse de, denetimsiz yöntemlerle sözcük ile aynı anlamdaki kullanımların ortak kümelere dahil edilmesi sağlanarak sözcük anlam ayrıştırması gerçekleştirilmektedir. Bu yöntemler ile elde edilen anlam güncel sözlüklerdeki klasik anlamlardan farklı bir sonuç ortaya koymaktadır. Bununla birlikte elde edilen anlam kümelerinin kalitelerinin ölçülmesi ve değerlendirilmesi genellikle diğer sistemlere göre daha zordur; kullanılan yaklaşımlardan bir tanesi tarafsız kişilerce üretilen veya geliştirilen kümelerin (örneğin, anketler yardımıyla) değerlendirilmesidir. Bir diğer değerlendirme yöntemi ise elde edilen kümelerin uçtan uca bir uygulamada kullanılarak, uygulama başarımının değerlendirilmesidir.

Denetimsiz yöntemlerin amacı, denetimli ve bilgiye dayalı yaklaşımlarda olduğu gibi anlam etiketi bulmak yerine anlamlar arasında ayrıştırma yapmak olduğundan farklılık göstermektedir. Bununla birlikte hem denetimli hem de denetimsiz

yöntemler, SABG konusuna ilişkin problemler olup birbirleriyle sıkı bir ilişki içerisindedir.

Denetimsiz SABG konusundaki temel yaklaşımlar 3 başlık altında sunulmaktadır: Bunlar bağlam kümeleme, sözcük kümeleme, ve birliktelik çizge yapılarıdır.

- **Bağlam Kümeleme (BK):** Denetimsiz yöntemlerle ele alınan ilk yaklaşım bağlam kümelemedir. Hedef sözcüğün derlem içerisindeki her örneği bir bağlam vektörü ile temsil edilmektedir. Bu vektörler daha sonra her birinin bir anlama karşılık geldiği gruplar halinde kümelenmektedir.

  Bu yöntemin temelini sözcük alanı olarak anılan, boyutların sözcüklerden oluştuğu düşüncesi oluşturmaktadır (Schütze, 1992). Bir derlem içerisindeki *w* sözcüğü, *j* numaralı bileşeni w_j'nin belirli bir bağlam içinde *w* ile kaç defa birlikte gözlendiğini gösteren bir vektör ile temsil edilmektedir. Bunun altında yatan varsayım sözcüklerin dağılım profillerinin sözcük anlamları hakkında bilgi verdiğidir. İki sözcük *v* ve *w* arasındaki benzerlik sözcüklere ilişkin vektörlerin kosinüs değerinin hesaplanmasıyla geometrik olarak belirlenmektedir:

$$sim(v,w) = \frac{v.w}{|v||w|} = \frac{\sum_{i=1}^{m} v_i w_i}{\sqrt{\sum_{i=1}^{m} v_i^2 \sum_{i=1}^{m} w_i^2}} \tag{2.4}$$

  Formül 2.4'te "*m*" vektör dahilindeki özellik sayısını göstermektedir. Derlemdeki her bir sözcük için vektör oluşturulmaktadır. Bu temsil biçimi sözcük anlamlarını birleştirmektedir: bir vektör sözcüğün temsil ettiği tüm anlamları içermektedir.

  Derlemdeki tüm sözcükler için oluşturulan vektör kümeleri bir araya getirildiğinde birliktelik matrisi elde edilmiş olmaktadır. Çok fazla boyut ortaya çıkması durumunda *Gizli Anlamsal İndeksleme* (GAİ) boyut azaltımında *Tekil Değer Ayrışımı* (TDA) ile uygulanmaktadır (Golub ve van Loan, 1989). Boyut indirgeme işlemi ile yüksek boyutlu düzlemdeki sözcükler kümesi düşük boyutlu bir alanda temsil edilmektedir; bunun sonucu olarak da benzer anlamların birleşmesi beklenmektedir.

Buradaki amaç bağlam olarak adlandırdığımız derlem ya da metin alt bölümlerine ilişkin vektörlerin kümelenmesidir. Bir bağlam vektörü metin içerisindeki kitle merkezi (vektörlerin normalize edilmiş ortalaması) olarak oluşturulmaktadır.

Son aşamadaki anlam ayrıştırması hedef sözcüğe ilişkin bağlam vektörlerinin kümeleme algoritmaları kullanılarak gruplandırılmasıyla yapılmaktadır. Schütze (1998) bu konuda *Context-group Discrimination* isimli belirsiz sözcüğe ilişkin örnekleri anlam kümelerine dahil eden bir algoritma geliştirmiştir. Bağlam benzerliği yukarıda açıklandığı şekliyle uygulanırken, kümeleme yinelemeli bir *En Fazla Olabilirlik* (EFO) modeli (maximum likelihood) olan beklenti maksimizasyonu (Expectation Maximization) algoritması kullanılarak gerçekleştirilmiştir. Farklı bir kümeleme yöntemi ise *Agglomerative* kümeleme olarak anılan yaklaşımı içermektedir (Pedersen ve Bruce, 1997b). Başlangıçta her örnek tekil bir kümedir. İlerleyen aşamalarda Agglomerative kümeleme ile birbirine en benzer kümeler birleştirilmekte, bu süreç belli bir eşik değerine kadar devam etmektedir.

Bağlam vektörlerinin oluşturulmasındaki bir problem, sözcük birlikteliklerinin dağılımının belirlenmesi için önemli miktarda etiketsiz eğitim verisine gereksinim duyuluyor olmasıdır. Bir diğer konu ise farklı bağlam kümelerinin farklı sözcük anlamları ile eşleşmeme olasılığıdır. Bu noktada denetimli bir sınıflandırıcının eğitilerek uygulanması sorunun üstesinden gelmek için önerilmiştir (Niu ve diğ, 2005).

- **Sözcük Kümeleme (SK):** Önceki bölümde sözcük anlamları birinci ve ikinci dereceden bağlam vektörleri olarak tanımlanmıştır. Farklı bir yaklaşım ise anlamların, anlam kümeleme yöntemleri kullanılarak bulunmasıdır. Bu yöntemler anlamsal olarak benzer ve belirli bir anlamı ifade eden sözcükleri kümeleyen yaklaşımlardır.

  Sözcük kümeleme konusunda bilinen bir yaklaşım w_0 hedef sözcüğü ile benzerlik gösteren sözcüklerin $W = (w_1, w_2, \dots\dots\dots\dots, w_k)$ tanımlanmasıdır (Lin, 1998). w_0 ve w_i arasındaki benzerlik derlemde gözlenen sözdizimsel bağımlılıklar (örn., özne-yüklem, yüklem-nesne, sıfat-isim) gibi bilgileri içeren özelliklerin değerlendirilerek gerçekleştirilmektedir. İki sözcüğe ilişkin bağlılık arttıkça, paylaşılan bilginin de arttığı düşünülmektedir. Bununla

birlikte, bağlam vektörlerinde olduğu gibi W içerisindeki sözcükler w_0'a ilişkin tüm anlamları kapsayacaktır. Bu anlamları ayrıştırmak için bir sözcük kümeleme algoritması uygulanmaktadır. W'nin w_0'a belli bir benzerlik derecesine göre sıralanmış benzer sözcükler listesi olduğu varsayalım. Bir benzerlik ağacı T başlangıçta w_0 tekil düğümünden meydana gelmek üzere oluşturulur. Sonraki adımda, her $i \in \{1,\ldots..,k\}$, $wi \in W$ T ağacına wj, wi'ye en benzer sözcük olacak şekilde eklenir. Ağacın budanması işleminden sonra w_0 altında yer alan her alt ağaç w_0'ın ayrı bir anlamı olarak değerlendirilir.

Bir sonraki yaklaşım *Clustering by Committee* (CBC) adındaki farklı bir sözcük kümeleme yöntemini kullanan algoritmadır (Lin ve Pantel, 2002). Her hedef sözcük için, benzer sözcükler kümesi yukarıda anlatıldığı şekilde gerçekleştirilmektedir. Benzerliği tekrar hesaplama için her sözcük her özelliğin, sözcüğün gözlendiği sözdizimsel bağlamın ifadesi olduğu bir özellik vektörü ile temsil edilmektedir. Hedef sözcüklerden oluşan bir küme verildiğinde (derlemdeki tüm hedef sözcükler) w_i ve w_j sözcükleri için karşılıklı benzerlik S_{ij} değerlerini içeren bir S benzerlik matrisi kurulur.

İkinci adımda, verilen bir E sözcük grubu için kümeleri belirlemek üzere özyinelemeli *Committees* adı verilen bir süreç uygulanır. Bu noktada standart bir kümeleme tekniği olan ortalama-bağ kümelemesi uygulanmaktadır. Her adımda herhangi bir komite tarafından içine alınmayan (her komitedeki kitle merkezine yeterince yakın olmayan) sözcükler devre dışı bırakılır. Yukarıda anlatılana benzer şekilde her sözcük tek bir komiteye bağlanabileceğinden bu yapı anlamları birleştirmektedir.

Son aşama olan anlam ayrıştırılması bölümünde, özellik vektörü olarak tanımlanmış her hedef sözcüğün $w \in E$ her komiteye ve kitle merkezine ilişkin benzerliği yinelemeli olarak değerlendirilerek en benzer kümeye atama gerçekleştirilir. Bir *w* sözcüğünün *c* komitesine atanmasından sonra w ve c içindeki diğer üyelerin kesişen özellikleri *w*'nin temsil edildiği özelliklerden çıkarılır.

CBC yöntemi WordNet sözcük anlamlarının tanımlanmasında sınandığında %61 tutturma ve %51 bulma değerleri elde edilmiştir. Önceki pek çok yaklaşımın aksine CBC kavramlar için düzlemsel bir kavram çıktısı sağlamaktadır (kümeler için hiyerarşik bir yapı oluşturmaz). Yakın zamanda tanıtılan yenilikçi bir yaklaşımla sözcük üçlüleri üzerinden anlam çıkarımı

yapılması üzerinde durulmuştur (Bordag, 2006). Bu yöntem her sözcük birlikteliği için tek anlam varsayımını kullanmakta ve birliktelik üçlülerini kesişimlerine bakarak kümelemektedir.

- **Birliktelik Çizgeleri (BÇ):** Anlam ayrıştırma konusunda farklı bir yaklaşım çizge-tabanlı yöntemlerin kullanımıyla son zamanlarda benimsenmiş ve başarılı sonuçlar elde edilmiştir. Bu yaklaşımlar birliktelik çizgelerinin kullanımına odaklanmaktadır; bu kullanımda $G = (V, E)$ sözcükler düğüm noktası olarak adlandırdığımız V'ye, kenarlar E ise bir paragraf veya daha geniş bağlamlarda bir ilişki içinde birlikte gözlenen sözcüklere karşılık gelmektedir.

  Bir bağlamdaki sözcüklerin birbirleri arasındaki ilişkilerden birliktelik çizgesinin oluşturulması tanımlanmıştır (Widdows ve Dorow, 2002). Bu yöntemler dahilinde öncelikle verilmiş olan bir w hedef sözcüğü için G_w çizgesi oluşturulur. G_w ile eşleştirilen komşuluk matrisinin normalize edilmesiyle, çizge bir Markov zinciri olarak ele alınabilmektedir. İzleyen adımda Markov kümeleme algoritması sözcük anlamlarının belirlenmesi amacı ile uygulanır (van Dongen, 2000). Bu aşamada daha uzak komşuları ve daha popüler düğümleri bulmak üzere genişleme adımları uygulanır.

  Bu alanda sözü edilen çalışmalardan sonra Hyperlex yaklaşımı sunulmuştur (Veronis, 2004). Bu yöntemle öncelikle, derlemi oluşturan paragraflarda hedef sözcüklerle birlikte gözlenen her sözcük çifti birliktelik matrisine eklenmektedir. İki sözcük aynı paragrafta yer alıyorsa bu sözcükler birlikte gözleniyor anlamına gelmekte ve bu düğüm noktaları bir kenarla birleştirilmektedir. Düğüm noktalarını birleştiren her kenara bu iki sözcüğün göreli sıklık değerlerine göre bir ağırlık atanmaktadır. Daha sonra sırasıyla merkez düğümler (hub düğüm) belirlenmekte, sözcük anlamlarını temsil eden bu düğümler esas alınarak, çizge yapısından ağaç temsiline dönüşüm gerçekleşmektedir. Sınama verisindeki sözcüklere ilişkin örnekler ağaç yapısındaki merkez düğümler altında aranmaktadır. Bir sözcük ağaçta yer alıyorsa merkez düğümden sözcüğü temsil eden düğüme kadar olan yola ilişkin ağırlık hesaplanmaktadır. Anlam ataması yapılmak istenen metin için her bir gözü bir merkez düğüme (anlama) karşılık gelen bir vektörde sözcüklerden elde edilen ağırlıklar saklanmakta, nihai anlam ataması ise

ağırlığı en büyük olan vektör gözünde elde edilen anlam olacak şekilde gerçekleştirilmektedir.

Anlam belirlemede çizge kullanımını esas alan bir diğer yöntem ise, PageRank yöntemidir (Brin ve Page, 1998). PageRank Google arama motorunun temel bileşeni olan ve web sayfalarının sıralamasını hesaplamak üzere geliştirilmiş, bilinen bir algoritmadır. Barındırdığı ilişkiler çizge yapısında tanımlanabilen çeşitli araştırma alanlarında da yapı içindeki birimlerin önem derecelerinin belirlenmesinde kullanılmaktadır. Ağırlıklandırılmış tanımlamada, $v_i \in V$ düğümüne ilişkin PageRank derecesi formül 2.5 ile verilmektedir:

$$P(v_i) = (1 - d) + d \sum_{v_j \to v_i} \frac{w_{ji}}{\sum_{v_j \to v_k} w_{jk}} P(v_j), \qquad (2.5)$$

Formül 2.5'de verilen $v_j \to v_i$ ifadesi v_j ile v_i arasında bir kenar bulunduğunu, w_{ji} ağırlık değerini, d ise genellikle 0,85 olarak alınan bir katsayıyı (damping factor) vermektedir. Bu katsayı bizi v_i'ye götüren bir bağlantıyı izlemek (eşitliğin ikinci bölümü) ile rastlantısal olarak v_i'ye ulaşma olasılıklarını modellemektedir. Formüldeki özyinelemeli yapıya dikkat edilmelidir, her düğüm noktasına ilişkin PageRank hesaplaması yinelemeli olarak belirli bir yakınsama noktasına kadar ya da çoğu durumda belirli bir tekrar sayısınca gerçekleştirilmektedir.

PageRank algoritması denetimsiz SABG'ye uyarlandığında, w_{ji} değeri Hyperlex algoritmasında olduğu gibi w_i ve w_j sözcüklerine ilişkin birlikte gözlenme olasılıklarıdır (Agirre ve diğ, 2006). PageRank algoritması uygulandığında, düğüm noktaları PageRank değerlerine göre sıralanmaktadır. En iyi değerleri alan düğüm noktaları merkez düğüm olarak seçilmektedir.

HyperLex yöntemi bir bilgiye erişim sistemi için sınanmış ve sınırlı sayıda sözcük için iyi sonuçlar verdiği gözlenmiştir (Veronis, 2004). HyperLex ve PageRank için yapılan daha sonraki denemelerle, merkez düğüm için komşu düğümlerin sayısı, kenarlara ilişkin en küçük sıklık değerleri, düğüm noktaları ve merkez düğüm gibi çeşitli parametrelerde ayarlamalar yapılmıştır (Agirre ve diğ, 2006). Denemeler Senseval-3 sözcüksel örneklerinin isim grubu üzerinde gerçekleştirilmiştir; her iki algoritma için denetimli yöntemlere oldukça yakın sonuçlar elde edilmiştir. Diğer sistemlerle

karşılaştırma yapmanın mümkün olabilmesi için, elde edilen merkez düğüm sözcükleri Senseval-3 için referans olarak alınan WordNet yapısındaki anlamlarla eşleştirilmiştir.

2.1.3 Melez yöntemler

Melez yöntemlerde sözlükler, ontolojiler ve eş anlamlılar sözlükleri gibi dış kaynakların yanı sıra derlemlerden öğrenilen bilgilerden de faydalanılmaktadır. Örneğin WordNet yapısındaki eş anlamlılık, üst anlam bilgilerinin ya da söz öbeklerinin, küçük bir bölümü işaretli olan veri kümesi ile birlikte kullanılması mümkündür. Bağlamdaki tek anlamlı sözcüklerin belirsizliği giderilmiş sözcükler için çekirdek veri olduğu kabulüyle, tekrarlamalı her adımda bir sözcüğün anlam belirginleştirme işlemi önceki sözcükler ile olan anlamsal uzaklığa bakılarak yapılabilmektedir. Aynı zamanda WordNet benzeri ontolojiler üzerinden pek çok farklı anlamsal ilişkinin ortaya konması da mümkündür. Melez yöntemler kapsamında yapılan bazı çalışmalar aşağıda tanıtılmaktadır:

- **SenseLearner:** Bu yöntemle eğitim verisinde gözlenen sözcükler için anlamsal bir dil modeli oluşturmak üzere kısmen etiketli olan veriden faydalanılmaktadır (Mihalcea ve Faruque, 2004). WordNet kullanımı ise derlemde gözlenmeyen sözcükler için genelleme yapılmasını sağlamaktadır. Anlamsal dil modeli oluşturmada izlenen adımlar şunlardır:
    - Derlem üzerinden her sözcük etiket türü için bir eğitim kümesi oluşturulmaktadır.
    - Her eğitim örneği bir özellik vektörü ve sınıf etiketi ile temsil edilmektedir.
    - Sınama aşamasında her örnek cümle için benzer vektör temsili oluşturulmaktadır.
    - Sınıflandırıcı ile sözcük ve anlam tahmin edilmektedir.
    - Tahmin edilen sözcük ile gözlenen sözcük aynı ise tahmin edilen anlam doğru anlam olarak seçilmektedir.

  Anlamların genelleştirilmesi ise, WordNet'teki anlamsal bağlılıkların kullanımı ve Lin'in algoritmasının (Lin, 1997) uyarlanması ile gerçekleştirilmektedir. Örneğin derlemde "su iç" cümlesi geçiyorsa, üst anlamlılık ağacı kullanılarak "içecek/sıvı al" yapısı çıkarsanmaktadır. "içecek al" bağlılığı ise alt – üst

anlamlılık ilişkilerinden faydalanılarak "çay al" yapısındaki çay sözcüğünün anlam belirginleştirmesinde kullanılabilmektedir.

- **Yapısal Anlamsal Bağlantılar (YAB):** Yapısal anlamsal bağlantılar yöntemi tekrarlamalı bir yaklaşımdır (Navigli ve Velardi, 2005). Yöntem dahilinde aşağıda listelenen anlam ilişkilerinden faydalanılmaktadır:
    - Üst anlamlılık (araba#1 bir araç#1 türüdür) (tür ilişkisi)
    - Alt anlamlılık (Üst anlamlılık özelliğinin tersi) (içerme)
    - Sahiplik ilişkisi (oda#1 da duvar#1 var) (sahip olma)
    - Parçası olma (Sahiplik ilişkisinin tersi) ile ifade edilir. (Duvar odanın bir parçası)
    - İlgililik, ait olma (dişçi#1 diş#)
    - Özellik ilişkisi (kuru#1 nemlilik#1 in bir özelliği)
    - Benzerlik (güzel#1 ile hoş#1)
    - Kullanılan diğer açıklamalar, bağlam ve alan bilgileri.

  Aynı zamanda tek anlamlı sözcükler belirsizlik giderme için çekirdek küme olarak kullanılmaktadır.

2.1.4 Anlam belirsizliği giderme yöntemlerinin karşılaştırılması

Anlam belirsizliğini giderme araştırmalarında kullanılan her yöntemin üstün ve eksik yanları olduğu, yaptığımız kaynak araştırmalarında vurgulanmaktadır. Araştırmalarımız sonucunda verilen genel yöntem sınıflarına ilişkin üstün ve eksik yönler Çizelge 2.1'de yansıtılmıştır.

Çizelge 2.1 : Yöntem sınıflarının karşılaştırılması.

Yöntem	Üstünlük	Eksiklik
Bilgi Tabanlı	Daha yüksek doğrulukta sonuçlar	Algoritmalar sözcük örtüşmesine dayalı, örtüşme seyrekliği sorunu gözlenebilir. Başarım sözlük tanımlamalarına bağımlı.
Denetimli	Gerçekleştirme açısından diğer yöntemlere üstün.	Kaynak kısıtı olan dillerde elde edilen sonuçlar yetersiz.
Denetimsiz	Anlam envanterleri ya da işaretli derlem gereksinimi bulunmuyor.	Gerçekleştirim zor ve diğer yöntemlere göre daha düşük başarım.

Yaptığımız araştırmalar sonucunda bilgi tabanlı yöntem sınıflarına dahil olan algoritmaların başarımı yüksek olduğu ancak kullanılan kaynaklar doğrultusunda bazı sıkıntılar gözlenebildiği anlaşılmıştır. Örneğin sözlük anlamlarının eşleşmelerinin kullanıldığı yöntemlerde örtüşmelerin seyrek olarak gözlenebilmesi bu yöntemlerin olumsuz yönlerindendir. Öte yandan denetimli yöntemlerle elde edilen sonuçlar başarım açısından oldukça tatmin edici olmakla birlikte özelikle kaynak kısıtı olan dillerde anlam işaretli derlemlerin azlığı önemli problemlerden bir tanesidir. Denetimsiz yöntemlerin ise diğer yöntemlere olan en önemli üstünlüğü anlam işaretli derlemlere olan bağımlılığın ortadan kalkmasıdır. Diğer yandan gerçekleştirmeye ilişkin sorunlar ve başarımın belirli ölçüde daha düşük olması da bu yöntemlerin olumsuz kabul edilebilecek yönlerindendir.

2.2 SABG Sistem Sınıfları

Belirsizliği giderilecek hedef sözcüğün ele alınma biçimi ve kullanılan derlem göz önünde bulundurularak yapılan sınıflandırmadır. İlk grupta önceden belirli olan sözcüklerin anlam belirsizliği giderilirken, ikinci grupta tüm sözcükler için belirsizlik giderilmesi hedeflenmekte ve kullanılan derlemler bu duruma uygun olmaktadır.

Bu bölümde belirsizlik gidermede izlenen yaklaşımlar, kullanılan derlem ve çalışmanın kapsamı değerlendirildiğinde yapılan sınıflandırmaya göre açıklanmaktadır. Bu yaklaşımlar *Seçilmiş Sözcük Yaklaşımı* (SSY) ve *Tüm Sözcükler Yaklaşımı* (TSY) olmak üzere iki başlık altında toplanmaktadır.

2.2.1 Seçilmiş sözcük yaklaşımı

Seçilmiş sözcük yaklaşımında anlam belirsizliği taşıyan sözcük önceden belirlenerek, bu sözcüğü içeren “n” tane paragraf ya da metin ele alınmaktadır. Her bir paragrafta seçilmiş sözcüğün anlamı işaretli olarak bulunmaktadır. Örnek paragraflarda hedef sözcüğün sol ve sağ komşuluklarında yer alan sözcüklere ilişkin çeşitli özelliklerden faydalanılarak anlam belirsizliğinin giderilmesi hedeflenmektedir. Eğitim ve sınama kümeleri örneklerin 10 katlı çapraz doğrulama (k- katlı ÇD) ya da farklı bir strateji ile ayrılması sonucu elde edilmektedir. Eğitim kümesinin ayrılması ve eğitim aşamasının ardından, içinde seçilmiş sözcüğün bulunduğu herhangi bir paragraf ele alındığında, eğitim kümesinde edinilmiş olan bilgilerden yararlanarak seçilmiş hedef sözcüğün anlamı çıkarılmaktadır. Kullanılan sözcük ve anlamlar sınırlı ve belirgin

olduğu için seçilmiş sözcük yaklaşımının kullanımı denetimli makine öğrenmesi yöntemlerinde sıklıkla tercih edilmektedir.

2.2.2 Tüm sözcükler yaklaşımı

Tüm sözcükler yaklaşımında amaç bir metinde karşılaşılan tüm sözcüklere ilişkin belirsizliğin giderilmesidir. Bu yaklaşımda bir metin değerlendirilirken metnin içindeki isim, sıfat, zarf ve eylemlerin belirsizlikleri giderilmeye çalışılır. Bu amaçla sözlük, sözlük ağacı, ontoloji vb. yardımcı araçlardan yararlanılmaktadır. Metnin içindeki her bir sözcük için belirsizlik giderme sırayla gerçekleştirilmektedir. Tüm sözcükler yaklaşımı bir yönü ile *Part of Speech* (POS) etiketleme ile benzerlik göstermekte ancak anlam belirsizliğinin giderilmesinde çok daha büyük bir etiket kümesine gereksinim duyulmaktadır. Büyük etiket kümesine gereksinim duyulması, uygun eğitim kümesi bulmada zorluğa ve veri seyrekliği problemine yol açmaktadır.

Seçilmiş sözcük ve tüm sözcükler yaklaşımları kapsamları bakımından farklılık göstermektedir. Yapılan çalışmalar kapsamında, seçilecek olan yönteme uygun veri kümeleri hazırlanmakta ve kullanılmaktadır. Bu tez çalışmasında Türkçe için anlam belirsizliği derecesi yüksek olan isim ve eylemler tespit edilmiş ve seçilmiş sözcük yaklaşımı benimsenmiştir.

2.3 Anlam Belirsizliği Gidermede Gerekli Bilgi Tipleri

Bu bölümde sözcük anlam belirsizliği gidermede kullanılan bilgi tipleri özetlenmektedir. Yapılan araştırmalar sonucunda gereksinim duyulan bilgiye ilişkin sınıflandırma aşağıdaki şekilde verilmektedir (Agirre ve Martinez, 2001; Hirst, 1987; McRoy, 1992).

- **Sözcük Öğeleri:** Bir sözcüğün anlamının belirginleştirilmesinde tümce içerisinde kullanıldığı öğe bilgisi ayırt edici bir özelliktir. Örneğin, yüz sözcüğü, yüz kilo, yüz (insan) örneklerinde olduğu gibi isim ya da *denizde yüz* kullanımındaki gibi eylem olarak karşımıza çıkabilmektedir. Sözcüklerin farklı türdeki kullanımlarından anlam belirginleştirilmesinde faydalanılmaktadır.
- **Biçimbilim**: Sözcük kökleri ve sözcüklerin türemiş biçimleri arasındaki ilişkiler de anlamların ayırt edilmesinde yardımcı olmaktadır.
- **Söz Öbekleri:** Anlam belirsizliği taşıyan sözcükler bir söz öbeği içinde kullanıldıklarında anlamları kolaylıkla netleştirilebilmektedir. Örneğin *pastırma*

yazı söz öbeğinde yazı sözcüğünün anlamı açıktır ve *yazı yazmak* kullanımındaki anlamdan kolaylıkla ayırt edilebilmektedir.

- **Anlamsal Sözcük Birlikleri:** Bu grup kendi içinde dört farklı sınıfa ayrılmaktadır.
    - **Cinse göre sınıflandırma**: İlk grup için örnek olarak *mobilya* ve *sandalye* arasındaki ilişki verilmektedir.
    - **Durum**: *Sandalye* ve *garson* arasındaki ilişki durum ilişkisine örnek olarak verilmektedir.
    - **Konu**: Örnek olarak *futbol* ve *top* arasındaki ilişki verilmektedir.
    - **Argüman-baş ilişkisi:** "*Köpek adamı ısırdı*" cümlesindeki köpek sözcüğü ve ısırmak eylemi bu ilişkiye örnek olarak verilmektedir.
- **Sözdizimi Bilgileri:** Sözcüklerin farklı sözdizimsel kullanımları, geçişli veya geçişsiz sözcük özellikleri göstermeleri hedef sözcüğün ilişkili olduğu anlam hakkında bilgi sağlamaktadır. Dolayısıyla anlamların ayrıştırılmasında sözdizimsel özellikler ayırt edici olabilmektedir.
- **Kullanım Alanı:** Kullanım alanı bilgisi ile bir sözcüğün hangi konuya göre anlamının tercih edileceği bilgisini verir. Örneğin, *havale* sözcüğü *sağlık* konusunda hastanın içinde bulunduğu durumken, *bankacılık* konusunda para aktarımı anlamında kullanılacaktır.
- **Anlam Sıklıkları:** Sözcük anlamlarını kullanım sıklıkları da belirsizlik gidermede yardımcı bir araç olarak kullanılabilmektedir. Birden fazla anlama sahip sözcüklerin genellikle ilk anlamı kullanılmaktadır.
- **Sebep Sonuç İlişkisi:** Bazı durumlarda anlam belirsizliğinin giderilmesi amacıyla çıkarsama içeren cümle ve yapılar faydalı olabilmektedir. Örneğin *"Tüm kitapçıları gezdi, sonunda aradığı baskıyı buldu"* cümlesinde kitap-baskı sözcükleri arasında benzer ilişki bulunmaktadır.
- **Anlamsal Roller:** *"Bu iş bütün günümü yedi."* Cümlesinde yemek sözcüğün nesnesi olan sözcük, yemek sözcüğü ile ilgili diğer anlamların değerlendirme dışında kalmasında yardımcı olmaktadır.
- **Seçimsel Öncelikler :** Anlam belirsizliğine sahip sözcüklerde, örneğin eylemlerde bir sözcüğün öznesinin sadece insanlar ya da *"sınıfta kalmak"* örneğinde olduğu gibi öğrenciler olduğu bilgisine sahip olmak belirsizlik gidermede kullanılan bilgi türlerindendir. Burada sözü edilen ilişki biçimi

anlamsal sınıflar düzeyinde olması ile sözcük düzeyinde ilişkilerin tanımlandığı argüman-baş ilişkisinden ayrılmaktadır.

2.4 Anlam Belirsizliği Gidermede Kullanılan Kaynaklar

SABG sistemleri kullandıkları algoritmalar kapsamındaki kaynaklar ile karakterize edilmektedir. Bu kaynaklar bilgisayarla okunabilir sözlükler, ontolojiler, derlemler ya da bu kaynakların kombinasyonlarını içine almaktadır (Agirre ve Martinez, 2001).

- **Bilgisayarla okunabilir sözlükler**: Bilgisayarla okunabilir sözlükler, önceki bölümde verilen bilgi tiplerinden; anlamsal sözcük birlikleri, sözdizimi bilgileri, seçimsel öncelikler ve anlam sıklıklarının sağlanmasında kullanılmaktadır. Örneğin sözlüklerde bir sözcük için ilk sırada verilen anlam en sık kullanılan anlam olarak karşımıza çıkmaktadır. Bazı çalışmalarda (Bruce ve diğ, 1992; Rigau ve diğ, 1997) anlamsal sözcük birlikleri farklı şekillerde kullanılmaktadır. Aynı zamanda bu çalışmalarda LDOCE'nin bilgisayarla okunan versiyonu ile konu kodları, sözdizimsel özellikler ve temel seçimsel öncelikler kullanılmaktadır. Diğer benzer sözlükler bu bilgileri içermemektedir.
- **Ontolojiler:** Özel ontolojilerin kullanıldığı az sayıdaki sistem dışında WordNet (Miller ve diğ, 1990) en sık kullanılan ontolojik yapıdır. WordNet'teki eş anlamlılık ve sınıflandırmalar anlamsal ilgililik ölçütlerinde kullanılan cinse göre sınıflandırma bilgisini sağlamaktadır (Resnik, 1993; Agirre, 1999).
- **Derlemler:** Elle işaretlenmiş derlemler makine öğrenmesi yöntemlerinin eğitilmesinde kullanılmaktadır. Eğitim kümesi hedef sözcük anlamının belirginleştirilmesinde ipuçları sağlayan özellikleri çıkarmak üzere eğitilmektedir. Yarowsky (1993) söz öbeklerinin bi-gram ve argüman-baş ilişkileri kullanılarak nasıl elde edildiğini göstermiştir. Yapılan çalışmalar dahilinde kolay çıkarılabilen özelliklerin, çok fazla işlem ve hesaplama gerektiren özelliklerin kullanımına tercih edildiği gözlenmektedir. Derlemlerle yapılan çalışmalarda bölgesel ve evrensel özelliklerden faydalanılmaktadır. İlk özellik kümesinde hedef sözcüğü çevreleyen komşu sözcükler ile olan ilişkiler ele alınmaktadır. Bu ilişkiler hedef ve komşu sözcüklere ilişkin temel etiketleme özellikleri, sözcük öbekleri ve basit sözdizimsel özellikler gibi bilgileri içine almaktadır. Evrensel özelliklerle ise hedef sözcükle birlikte 50 – 100 sözcüğü içine alan bir pencere değerlendirmeye alınmaktadır. Birbiri ile sıklıkla gözlenen

sözcükler arasında anlamsal bir bağ bulunduğu düşünülmektedir. Sözü edilen özelliklerin genellikle durum ve konu bilgilerinin çıkarılmasında faydalı olduğu düşünülmektedir.

Yapılan çalışmalarda sözlük, ontoloji ve derlemler haricinde bu kaynakların farklı kombinasyonlarından da faydalanılmaktadır. Bu kombinasyonlar, bilgisayarla okunabilir sözlükler ve ontolojiler, bilgisayarla okunabilir sözlükler ve derlemler, ontolojiler ve derlemler gibi ikili gruplardan oluşmaktadır.

2.5 Anlam Belirsizliği Gidermede Karşılaşılan Zorluklar

SABG konusu uzun yıllardır üzerinde çalışılan ve ilerleme kaydedilen bir alan olmakla birlikte zorlukların kaynaklandığı belli başlı noktalar; değerlendirmeye alınan çerçevenin belirlenmesi, anlam kümelerinin saptanması ve kullanılması, değerlendirme aşaması gibi noktalardan kaynaklanmaktadır.

SABG sistemlerinin tümünde gerekli bilgi sözcüğün yer aldığı çerçeveden çıkarılmaktadır. Bu nedenle sözcüğün kullanım yeri önemli verilerden bir tanesidir. Yapılan çalışmalarda hedef sözcük ve komşu sözcükler, bu sözcükler arasındaki ilişkiler, sözdizimsel özellikler gibi pek çok özellik ele alınmıştır. Kullanılan SABG sistemine göre bu özellikler oldukça farklılık göstermektedir. Yukarıda verilen özellikler ya da sözcüklerin kendi başına özellik olarak kullanılması tercih edilebilmektedir. Aynı zamanda özellikler bölgesel veya evrensel çerçevede incelenmektedir.

Karşılaşılan bir diğer zorluk ise sözcük anlam kümelerinin belirlenmesidir. Bu zorluk bir sözcüğe ilişkin kesin anlam sayısını belirlemenin zorluğundan kaynaklanmaktadır. Sözcük anlamları dinamik bir yapıda olduğu için geliştirilen yöntemlerin farklı bir alana taşınması veya uyarlanması sonucunda zorluklar ortaya çıkabilmektedir. Bununla birlikte sözlükler sözcüklerin tüm anlamlarını sağlamakta yetersiz kalabilmektedir. Eğitim verisinin olabildiğince büyük tutulması kaçınılmaz olmaktadır. Bu ve benzer zorluklar nedeniyle yakın zamanda yürütülen çalışmalarda önceden belirli sözcük kümelerinin kullanımı bırakılarak, anlamların derlemlerin içinden dinamik olarak çıkarıldığı yaklaşımlar benimsenmiştir.

SABG konusu dahilinde en önemli ve zor noktalardan biri de sistemlerin değerlendirme aşamasıdır. Farklı çalışmaların birbirleri ile karşılaştırılması, farklı

veri kümeleri üzerinden elde edilen sonuçların birlikte değerlendirilmesi önemli noktalardan bir tanesidir. SABG yaklaşımlarının standart bir biçimde değerlendirilmesi üzerine yapılan çalışmalar sürdürülmektedir.

2.6 Seçilen Yaklaşım ve Yöntemler

Bu tez çalışmasının yaklaşım ve yöntem belirleme aşamasında, Türkçe için var olan kaynaklar göz önüne alınarak, “Seçilmiş Sözcük Yaklaşımı” ve ”Denetimsiz Yöntem” kullanılması hedeflenmiştir. Tüm sözcükler yaklaşımını uygulayabilmek için gerekli olan sözlük ve ontolojilerin Türkçe için bulunmaması ya da yeterli olmaması bu yöntem üzerinde çalışmamızı engellemiştir. Sözcüklerin anlamlarının işaretlenmiş olduğu büyük ölçüde bir Türkçe derlemin bulunmayışı, çalışmamızı denetimsiz yöntemlere yönlendirmiştir.

Ancak, belirsizliği giderilmek istenen sözcüğün,

- Komşu sözcükler ile ilişkisini,
- Komşuluk ilişkisinin boyutunu,
- Birlikte yer alan sözcükler ile ilişkisini
- Belirsizliğin giderilmesinde etkin olan özelliklerini

ortaya çıkarabilmek amacıyla denetimli yöntemlerden yararlanılabileceği görüşüyle, bu yöntemler üzerinde de çalışılmasına karar verilmiştir.

3. DENETİMLİ YÖNTEMLER ÜZERİNDE YAPILAN ÇALIŞMALAR

Bu çalışma kapsamında denetimli ve denetimsiz yöntemler kullanılarak çeşitli çalışmalar yapılmıştır. İlk aşamada denetimli yöntemler üzerinde çalışılmış, anlam belirsizliği gidermede etkin özelliklerin, pencere genişliğinin ve algoritmaların araştırılması hedeflenmiştir. Elde edilen bulgular denetimsiz yöntemlerde de kullanılmış ve iki sınıfa ilişkin yöntemlerin başarımları karşılaştırılmıştır.

SABG alanındaki denetimli ve denetimsiz yöntemlerin araştırılması ve geliştirilmesinden önce bu çalışmaları yapabilmek için özel bir derlemin hazırlanması gerekmiştir. *Hedef Sözcük Derlemi* (HSD) olarak adlandırdığımız derlem bu tez çalışması kapsamında hazırlanmıştır.

Bu bölümde yapılan çalışmalar aşağıda sıralanmaktadır:

- Türkçe derlem hazırlanması
    - Türkçe kaynakların ve yapılan benzer çalışmaların araştırılması.
    - Türkçe için uygun veri kümesinin belirlenmesi.
    - Türkçe veri kümesi hazırlanması.
    - Veri kümesinin oylayıcılar tarafından etiketlenmesi.
- Denetimli yöntemler üzerinde yapılan geliştirmeler
    - Konumsal özelliklerin anlam belirsizliği gidermede etkisinin incelenmesi.
    - Konumsal özellikler için etkin özellik kümesinin isim ve eylem kümeleri için saptanması.
    - Veri kümesi üzerinde sözcük kesesi özelliklerinin değişen pencere boyutu için sınanması ve etkin pencere boyutunun belirlenmesi.
    - Konumsal özellikler ve sözcük kesesi özelliklerinin anlam belirsizliği gidermedeki etkilerinin ayrı ayrı ve birleştirilmiş özellikler olarak değerlendirilmesi.
    - Konumsal özellik alt gruplarının anlam belirginleştirmede etkisinin incelenmesi

    - Veri üzerinde sınanan denetimli yöntem sonuçlarının önceki Türkçe çalışmalarda elde edilen bulgularla karşılaştırılması.
- Sonuçların karşılaştırılması
    - Farklı özellik gruplarına ilişkin yöntem sonuçlarının karşılaştırılması.
    - Türkçe için yapılmış diğer çalışma sonuçları ile karşılaştırılması.

3.1 Türkçe Derlem Hazırlanması

Derlem hazırlanması sırasında iki konuya dikkat edilmiştir:

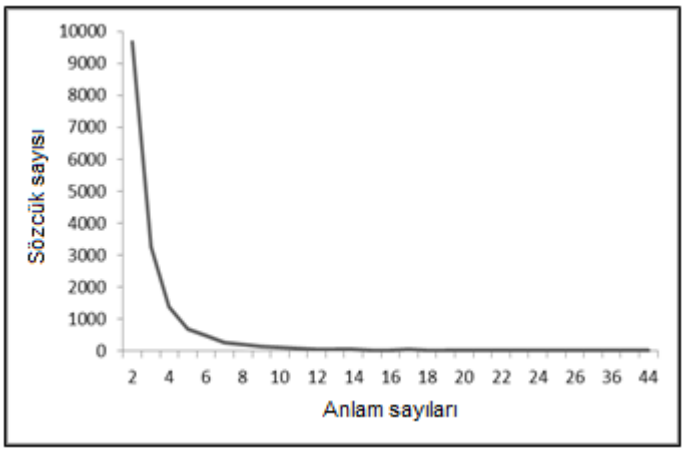
1. Hangi tür sözcüklerin belirsizliği üzerinde çalışılacağı,
2. Seçilen sözcük türü içinde belirsizliği yüksek olanların seçimi

Çalışmamızda isim ve eylemler üzerinde çalışılmasına karar verilmiş, ardından belirsizliği yüksek olan sözcükler seçilmiştir. Türkçedeki anlam belirsizliği konusunda Türk Dil Kurumu Sözlüğü üzerinde yaptığımız çalışmalar sonucunda aşağıdaki değerlere ulaşılmıştır:

- Türkçede 68.639 farklı sözcük bulunmakta ve sözcük başına ortalama 1,61 farklı anlam düşmektedir.
- Bununla birlikte, bu sözcüklerin 51.958 tanesi sadece 1 anlama sahip olarak SABG konusunun dışında kalmaktadır.
- Bu nedenle sözlükteki tek anlamlı sözcükler çıkarıldığında, belirsizliği olan toplam sözcük sayısı 16.681 olarak bulunmaktadır. Bu 16.681 sözcük için elde edilen ortalama anlam sayısı 3,53 bulunmuştur.
- Şekil 1, TDK sözlüğündeki sözcüklerin anlam sayıları ve karşılık düşen sözcük sayılarını göstermektedir.

Değerlendirdiğimiz ikinci bir kaynakta, Türkçe sözcükler anlam sayısına göre sıralanmış olarak verilmektedir (Göz, 2003). Çalışmamızda, Göz'ün (2003) hazırlamış olduğu sözcük listesinden yararlanılarak belirsizliği yüksek olan sözcükler seçilmiştir. Seçilen sözcüklerin ortalama anlam sayısı, isimler için 10,67, eylemler için 26,53 olarak hesaplanmıştır.

TDK sözlüğünde, bir sözcük için "n" tane anlam verilmişken, bizim oluşturduğumuz derlemde, aynı sözcük için "m" tane anlam bulunmaktadır. Aynı sözcük için derlemdeki anlam sayısı m<=n biçimindedir. Çizelge 3.1'de seçilmiş sözcükler, bu sözcüklerin TDK sözlüğündeki ve HSD derlemimizdeki anlam sayıları gösterilmiştir.



Şekil 3.1 : Anlam sayısı-sözcük sayısı dağılımı.

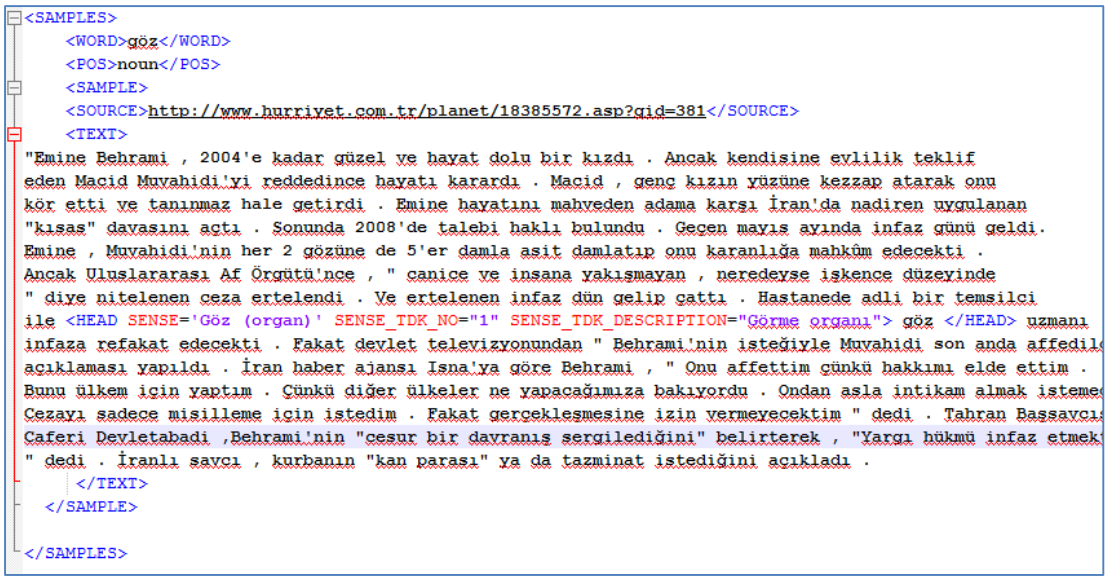
Derlem aşağıda verilen adımlar izlenerek oluşturulmuştur:

- 15 eylem ve 15 isim türünden sözcük seçilmiştir.
- İkinci aşamada, bu sözcükleri içeren yüzer adet örnek paragraf metin toplanmıştır.
- Paragrafların içindeki anlam belirsizliği olan sözcükler işaretlenerek XML formatında saklanmıştır.
- Örneklerdeki anlam belirsizliğine sahip sözcükler toplam beş kişi tarafından değerlendirilmiş ve oylatılmış ayrıca TDK sözlüğündeki en uygun karşılığı işaretlettirilmiştir. Temel olarak bu işlem sonucunda, anlam belirsizliği olan sözcükler farklı kişiler tarafından etiketlenmekte ve bu kişilerin oy çokluğuyla elde edilen en baskın anlam derlemin işaretlenmesinde kullanılmaktadır. Bu aşamada Değerlendiriciler Arası Uyum *(Inter Annotator Aggreement)* ölçütü dikkate alınmaktadır.

SABG alanında yürütülen benzer çalışmalarda belirsizliği giderilecek olan sözcük *head word* (target word) olarak adlandırılmaktadır. Çalışmamızda belirsizliği giderilmesi amaçlanan sözcük "hedef sözcük" olarak anılmaktadır. Derlemde "her paragraf için bir anlam" yaklaşımı benimsendiğinden, bir paragraf içindeki gözlenen hedef sözcüğün sadece bir anlamı olduğu varsayılmıştır. Derlemdeki 15 isim ve 15 eylemden meydana gelen sözcük grupları kapsamında 3100 adet örnek paragraf bulunmaktadır. Bu paragraflar XML formatında saklanmaktadır. Şekil 3.2, göz sözcüğünden alınmış bir örnek paragrafı göstermektedir.

Çizelge 3.1 : Derlemdeki sözcük grupları.

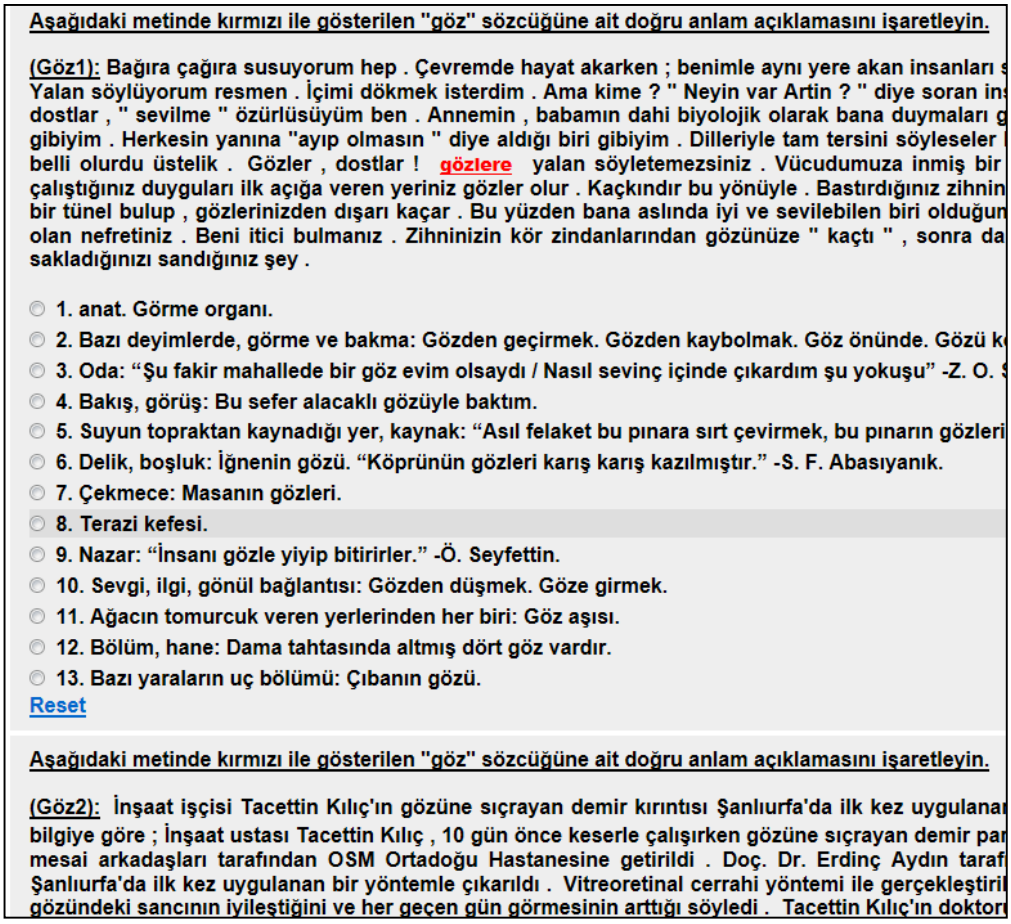
	Seçilmiş Sözcük	Anlam Sayısı			Seçilmiş Sözcük	Anlam Sayısı	
		TDK	HSD			TDK	HSD
İsim	Açık	16	12	Eylem	Aç	27	15
	Baskı	8	7		Al	33	15
	Baş	13	9		At	33	18
	Derece	7	6		Bak	17	14
	Dünya	7	7		Çevir	15	11
	El	10	5		Çık	56	19
	Göz	13	8		Geç	38	16
	Hat	9	9		Gel	36	18
	Hava	14	10		Gir	19	7
	Kaynak	8	7		Gör	20	14
	Kök	12	7		Kal	21	12
	Kör	7	6		Ol	25	17
	Ocak	11	6		Sür	16	11
	Yaş	9	7		Ver	22	14
	Yüz	15	12		Yap	20	11



Şekil 3.2 : Göz hedef sözcüğüne ilişkin bir örnek paragraf.

Bu bilgiler doğrultusunda derlemdeki her sözcüğe ilişkin örnek metinler Şekil 3.3'te görülen biçime dönüştürülüp işaretleme için değerlendiricilere sunulmuştur. Değerlendiricilere kolaylık sağlaması açısından paragraf boyları biraz

küçültülmüştür. Tüm derlem 15 farklı değerlendiriciye bir örnek için toplamda beş oy kullanılacak şekilde işaretlettirilmiştir. Değerlendiriciler sözcük anlamlarını Türk Dil Kurumu sözlüğündeki sözcük anlamlarını referans alarak seçmişlerdir. Şekil 3.3'te isim grubu içinde yer alan "göz" sözcüğüne ilişkin bir örnek görülmektedir:



Şekil 3.3 : "Göz" sözcüğü için anket örneği.

3.1.1 Değerlendiriciler arası uyum

SABG çalışmalarının üzerinde yürütüldüğü derlemlerde anlam işaretlemesi bir değerlendirici tarafından gerçekleştirildiğinde güvenilir kabul edilmemekte, farklı değerlendiricilerin oyuna sunulmaktadır. Bu doğrultuda derlemimizde daha önce birer defa işaretlenmiş olan paragraflar toplamda 15 kişinin katılımıyla, her örnek için 5 defa işaretlenmiştir. Bu çalışmada uyum derecesinin hesabı için Kappa istatistikleri kullanılmıştır (Cohen, 1960).

Derlemlerdeki anlam işaretlemeleri değerlendiriciler tarafından yapılırken verilen oylar arasındaki uyuşma göz önünde bulundurulmaktadır. Cohen'in Kappa ölçütünün, uyuşmanın şans eseri ile de ortaya çıkabilme durumunu göz önünde

bulundurması nedeniyle basit bir yüzde oranı ile bulunan uyuşmaya göre daha güvenilir sonuç verdiği kabul edilmektedir.

Cohen'in Kappa katsayısı iki değerlendirici arasındaki uyuşmanın güvenilirliğini ölçen bir istatistik yöntemidir. Bu ölçü her biri N tane maddeyi C tane birbirinden farklı sınıfa ayıran iki değerlendirici arasındaki uyumu ölçmektedir.

Bu ölçüye ilişkin bağıntı denklem 3.1'de verilmektedir. Denklem 3.1'de, Pr(a) iki değerlendirici arasında gözlemlenen uyuşmaların toplam oylamaya oranı, Pr(e) ise bu uyuşmanın şans eseri ortaya çıkma olasılığı olarak verilmektedir.

$$k = \frac{\Pr(a) - \Pr(e)}{1 - \Pr(e)} \tag{3.1}$$

Elde edilen 'k' değeri aşağıdaki gibi yorumlanmaktadır:

- k=1 ise iki değerlendirici birbiriyle tamamen uyuşmaktadır.
- k=0 ise iki değerlendirici için uyuşma sadece şans ile belirlenmekte ve diğer hallerde uyuşma olmamaktadır.

Cohen'in Kappa ölçüsü yalnızca iki değerlendiriciyi ele aldığından, çalışmamızda değerlendirici sayısının ikiden fazla olduğu durumlarda kullanılan Fleiss'in Kappa ölçüsü esas alınmıştır (Fleiss, 1971). Fleiss'in Kappa sayısı ikiden fazla sabit sayıda değerlendirici arasındaki karşılaştırmalı uyuşmanın güvenilirliğini ölçen bir istatistik yöntemidir. Bu ölçü ile sabit sayıda (n tane) değerlendiricinin her birinin, (N tane) maddeyi, (C farklı) sınıfa göre ayırmaları süreci sonunda ortaya çıkan, değerlendiriciler arasındaki uyum ölçülmektedir.

İşaretleme sırasında görev alan *n* sayıda değerlendirici N farklı durumu değerlendirmektedir. Her bir durum için ise *k* farklı sınıf bulunmaktadır. Eğer bir durum için iki değişik değerlendirici aynı anlamı veriyorsa değerlendiricilerin uyuştuğu kabul edilmekte, eğer farklı anlam veriyorlarsa uyuşmazlık durumu anlaşılmaktadır. Fleiss'in Kappa ölçüsü denklem 3.2'deki gibi tanımlanmaktadır:

$$K = \frac{\bar{P} - \overline{P_e}}{1 - \overline{P_e}} \tag{3.2}$$

Denklem 3.2'de $\overline{P}$ uyuşma olasılığını, $\overline{P_e}$ ise rastgele uyuşma olasılığını göstermektedir. Bu iki olasılık denklem 3.3'te verilen şekilde hesaplanmaktadır.

Denklem 3.3'te N durum sayısını, *n* değerlendirici sayısı ve k değerlemede kullanılacak sınıf sayını göstermektedir.

$$\boldsymbol{P_j} = \frac{\mathbf{1}}{\boldsymbol{Nn}} \sum_{\boldsymbol{i=1}}^{\boldsymbol{N}} \boldsymbol{n_{ij}} \quad ,$$

$$\mathbf{1} = \frac{\mathbf{1}}{\boldsymbol{n}} \sum_{\boldsymbol{j=1}}^{\boldsymbol{k}} \boldsymbol{n_{ij}} \tag{3.3}$$

Sonraki adımda ise i-inci durum için değerlendiricilerin ne derece uyuştuklarını gösteren P_i hesaplanmaktadır. Formül 3.4'te hesaplamaya ilişkin bağıntı verilmektedir.

$$P_i = \frac{1}{n(n-1)} \sum_{j=1}^{k} n_{ij}(n_{ij} - 1)$$

$$= \frac{1}{n(n-1)} \left(\sum_{j=1}^{k} n_{ij}^2 - n \right) \tag{3.4}$$

$\overline{P}$ değeri P_i'lerin ortalaması alınarak elde edilmektedir:

$$\bar{P} = \frac{1}{N} \sum_{1}^{N} P_i$$

$$= \frac{1}{Nn(n-1)} \left(\sum_{i=1}^{N} \sum_{j=1}^{k} n_{ij}^2 - Nn \right) \tag{3.5}$$

Toplam ortalama $\sum_{j=1}^{k} p_j^2$ olan $\overline{P_e}$ hesaplanmaktadır.

$$\overline{P_e} = \sum_{j=1}^{k} P_j^2 \tag{3.6}$$

Elde edilen değerler denklem 3.2'de yerine konularak Fleiss'in kappa katsayısı değerleri bulunmuştur. Çizelge 3.2'de sözü edilen değerler verilmektedir.

Çizelge 3.2 : Değerlendiriciler arası uyum.

	Sözcük	**Anlam sayısı**	***K* Kappa**		**Sözcük**	**Anlam sayısı**	***K* Kappa**
İsim	açık	16	0,56	**Eylem**	aç	27	0,54
	baş	13	0,76		al	33	0,58
	baskı	8	0,61		at	33	0,59
	derece	7	0,80		bak	17	0,39
	dünya	7	0,77		çevir	15	0,83
	el	10	0,74		çık	56	0,53
	göz	13	0,90		geç	38	0,59
	hat	9	0,62		kal	21	0,66
	hava	14	0,75		gel	36	0,29
	kaynak	8	0,85		gir	19	0,63
	kök	10	0,86		gör	20	0,54
	kör	7	0,86		ol	25	0,66
	ocak	11	0,83		sür	16	0,68
	yaş	9	0,65		ver	22	0,40
	yüz	15	0,76		yap	20	0,45

3.2 Denetimli Yöntemler Üzerinde Yapılan Geliştirmeler

SABG kapsamında kullanılan denetimli yöntemlerle yüksek doğrulukta sonuçlar elde edilmekte, yöntemlerin uygulanabilmesi için ise yeterli sayıda ve etiketli derlemlere gereksinim duyulmaktadır. Yapılan çalışma kapsamında ilk olarak bu yöntemlerin sınanması gerçekleştirilmiştir. Bu sınıfa giren yöntem ve algoritmaların bir bölümü; *Naive Bayes* (NB), *Karar Listeleri* , *Karar Ağaçları*, *İşlevsel Ağaç algoritması*, *Örnek tabanlı yöntemler* ve *DVM*'lerden oluşmaktadır.

3.2.1 NKA özelliklerinin kullanılması

Niteliklerin kazandırdığı anlamlar yöntemi ile, konumsal özellikler elde edilerek ±4 pencere aralığındaki komşu sözcükler ve hedef sözcüğün dilbilgisi özellikleri kullanılmıştır. Bu özelliklerin pencere aralığında bulunup bulunmama bilgisi ikili bir vektör yapısında tutulmuştur. Verinin Şekil 3.4'de görülen biçime gelmeden önce geçtiği aşamalardan birincisi biçimbilim aşamasıdır. Bu adımda hedef sözcük ve ele alınan komşu sözcüklere ilişkin tüm olası çözümlemeler elde edilmiştir. Bu aşamada

Oflazer'in (1994) biçimbilim çözümleyici aracından faydalanılmıştır. Bu çözümleyicinin ürettiği sonuçlar belirsizlik içerdiğinden çözümleme işleminin ardından belirsizlik giderici yazılımı kullanılmıştır (Yüret ve Türe, 2006). En son adımda, tüm sözcükler gövde biçimleri ve doğru biçimbirimlerinin olduğu şekle getirilmişlerdir. Türkçe sondan eklemeli dil yapısıyla oldukça zengin ve üretken bir dil olup, içeriğindeki sözcükler çok sayıda çekim ve yapım eki alabilmektedir. Şekil 3.5 "kuvvetlendirme" sözcüğüne ilişkin biçimbilimsel analiz çıktısını ve sözcüğün dönüşümlerini göstermektedir. Şekil 3.6'da görüldüğü gibi, çözümlemelerin içinde *Türetim Sınırlarının* (TS) olduğu durumlarda, en sağda yer alan türetim sınırından sonra gelen biçimbirim değerlendirmeye alınmıştır. Şekil 3.5'da görülen kuvvetlendirme sözcüğünün üç TS (Derivational Boundry - DB) içerdiği, dolayısıyla üç kez dönüşüm geçirdiği görülmektedir. Şekil 3.4'te ise yapılan çalışmanın çıktısı derlemimizden alınan bir örnek üzerinde gösterilmiştir.

```
kriz:(Noun)(A3pl)(Pnon)(Nom)
sonra:(Noun)(Zero)(A3sg)(P3sg)(Loc)
büyük:(Adj)
şirket:(Noun)(A3pl)(Pnon)(Gen)
<HEAD-SENSE='baş'-SENSE_TDK_NO="2"..
baş:(Noun)(A3sg)(P3sg)(Loc)
</HEAD>
bulun:(Adj)(PresPart)
yönetici:(Noun)(A3pl)(Pnon)(Gen)
görev:(Noun)(A3sg)(Pnon)(Nom)
değişim:(Noun)(A3pl)(P3sg)(Nom)
```

Şekil 3.4 : Seçili pencere aralığındaki örnek özellikler.

Kuvvet+len+dir+me
kuvvet+Noun+A3sg+Pnon+Nom^**DB**
+Verb+Acquire^**DB**
+Verb+Caus+Pos^**DB**
+Noun+Inf2+A3sg+Pnon+Nom

to make (something) become strong /
to strengthen (something)

Şekil 3.5 : Kuvvetlendirme sözcüğüne ilişkin biçimbilimsel çözümleme.

Bir sözcüğün nitelikleri olarak, isim, eylem, zamir vd. türler olmak üzere temel etiketleme (POS) özellikleri ve bu özelliklere bağlı alt özellikler kullanılmıştır. Her

bir komşu sözcük ve hedef sözcük için toplam 100'ün üzerinde özellik kategorisi ile sözcük kökünün kendisi kullanılmıştır. Böylece hedef sözcük ile birlikte toplamda 1081 elemanlık bir vektör yapısı elde edilmiştir. Ele alınan temel POS özellikleri Çizelge 3.3'te gösterilmektedir.

Çizelge 3.3 : Kullanılan temel POS özellikleri.

Tür	**Ele Alınan Alt Özellikler**
Verb (eylem)	Able, Fut, Past, Acc, Caus, Pres, Neces, Pos, Pass, Neg, Zero, Acquire, Become, Prog1, Prog2, Narr, Cond, Neces, Axxx, Pxxx.
Noun (isim)	Ness, Nom, Loc, Dat, Abl, Gen, Prop, Pnon, Ins, Dim, Zero, Acc, PastPart, FutPart, Agt, Axxx, Pxxx.
Postp (ilgeç)	PCNom, PCDat, PCAbl, PCIns.
Adverb (zarf)	AsLongAs, While, ByDoingSo, AfterDoingSo, When, Ly.
Pron (adıl)	Demons, Nom, Reflex, Pers, Dat, Pnon, Ques, Quant, Acc.
Adj (sıfat)	With, PresPart, Without, Rel, PastPart, FutPart, FitFor.
Ques (soru)	Pres, Past, Axxx.
Num (sayı)	Card, Ord, Range, Real, Percent.
Conj (bağlaç)	Alt özellik yok.
Det (belirteç)	Alt özellik yok.
Punc (noktalama)	Alt özellik yok.
Dup (tekrar)	Alt özellik yok.
Interj (ünlem)	Alt özellik yok.

Yukarıda verilen özellik kümesi içinde etkin olanlardan bir etkin özellikler kümesi oluşturulmuştur ve başarımdaki etkisi incelenmiştir. Çalışmamızda özellik azaltımı için CFS (*Correlation-based feature selection*) uygulanmıştır. Türkçe isim ve eylem grupları üzerinde yapılan çalışmada etkisiz özelliklerin elenmesinin ardından Çizelge 3.4'te verilen özelliklerin isim ve eylem gruplarının tamamında etkili olan özellikler olduğu gözlenmiştir. Etkin özellikler ve tüm özelliklerin kullanıldığı durumlara ilişkin doğruluk değerleri farkı dikkat çekici düzeyde bulunmuştur. Etkin özellik araştırması konumsal özelliklerin araştırıldığı ±4 pencere aralığında sınanmış ve etkin özelliklere konum bilgileri ile birlikte ulaşılmıştır. Özelliklere ilişkin konumlar hedef sözcük için 0, hedef sözcükten önce gözlenen komşuluklar için (-), hedef sözcüğü takip eden komşuluklar için ise (+) ile gösterilmiştir. Bulunan temel özellik

türleri sırasıyla adıl, isim, sıfat, eylem, zarf ve sözcük gövdesi (Pron, noun, adj, verb, adverb, root) gibi özellikler ve bu özelliklere bağlı alt gruplardan oluşmaktadır.

Çizelge 3.4 : İsim ve eylemlerde etkin konumsal özellikler.

Tür	**Özellik**	**Tür**	**Özellik**
İsim	Pron(-4), Noun+Acc(-2), Adj+PresPart(-2), Root(-1), Num(-1), Noun+Pnon(0), Noun+Abl(0), Noun+Ins(0), Root(+1), Adj+PastPart(+1), Verb+Zero(+2)	**Eylem**	Noun+Gen(-2), Adverb+AfterDoingSo(-2), Root(-1), Noun+Dat(-1), Noun+Abl(-1), Verb+Neg(0), Verb+Fut0), Pron(+1), Pron+Ques(+2)

Çizelge 3.5 ve Çizelge 3.6 konumsal özelliklerin Türkçe isim ve eylem grupları üzerinde özellik azaltımına gidilmeden elde edilen sınama sonuçlarını sırasıyla göstermektedir. Kullanılan yöntemler sırasıyla NB (Naive Bayes), örnek tabanlı bir yöntem olan Ibk (Instance Based Learning with parameter k; KNN algoritması) algoritması , J48 (C4.5 Karar ağacı algoritması), FT (Functional Tree – İşlevsel Ağaç), ve DVM olup, sonuçlar *En Baskın Anlam Oranı* (EBA) ile birlikte sunulmaktadır. Bu değer hedef sözcüğe ilişkin örneklerde en yaygın olarak gözlenen anlam oranını göstermekte ve başarım için alt sınır değeri olarak kabul edilmektedir. Çizelge 3.7'de ise Weka (Hall ve diğ, 2009) kullanılarak elde edilen yöntem sonuçlarının isim ve eylem grupları için ortalama değerleri görülmektedir. Çalışmamız kapsamındaki tüm deneylerde, DVM sınıflandırması için çok terimli (polynomial) Kernel kullanılmış, Weka'daki parametre değerleri korunarak C = 1 olarak alınmıştır. k-NN algoritması için değerlendirmeye alınacak en yakın komşuluk sayısını ifade eden *k* değeri 1 olarak alınmıştır.

Çizelge 3.5 : Türkçe isim grupları için NKA özellikleri doğruluk değerleri (%).

Hedef Sözcük	EBA	NB	Ibk	J48	FT	DVM
Açık	27,00	44,79	43,75	57,29	62,50	52,08
Baskı	30,00	52,04	32,65	29,59	45,92	48,98
Baş	30,00	57,43	56,44	52,48	67,33	69,31
Derece	38,00	75,76	67,68	83,84	84,85	84,85
Dünya	18,00	58,16	48,00	55,10	72,45	62,24
El	42,00	72,16	61,86	65,98	87,63	77,32
Göz	51,00	71,72	60,61	63,64	79,80	74,75
Hat	20,00	48,90	49,62	82,00	76,70	66,92
Hava	30,00	57,58	55,56	55,56	62,63	66,67
Kaynak	30,00	59,18	57,14	67,35	71,43	69,39
Kök	28,00	60,61	42,42	75,76	76,77	70,71
Kör	53,00	61,62	54,55	53,54	85,86	73,74
Ocak	26,00	55,45	54,55	48,18	68,18	65,45
Yaş	52,00	78,35	70,10	77,32	84,54	83,51
Yüz	27,00	55,10	53,10	47,96	75,51	68,37

Çizelge 3.6 : Türkçe eylem grupları için NKA özellikleri doğruluk değerleri (%).

Hedef Sözcük	EBA	NB	Ibk	J48	FT	DVM
Aç	15,00	27,96	26,88	53,76	45,16	43,01
Al	23,00	34,70	32,65	55,10	46,94	40,81
At	22,00	33,33	38,54	75,00	71,88	54,17
Bak	20,00	35,79	29,47	44,21	49,47	44,21
Çevir	27,00	57,00	59,00	74,00	77,00	66,00
Çık	16,00	39,80	46,94	57,14	69,39	63,27
Geç	23,00	62,00	52,00	75,00	73,00	69,00
Gel	24,00	50,00	50,00	53,70	64,81	65,74
Gir	39,00	59,00	39,00	71,00	68,00	57,00
Gör	26,00	40,54	37,84	56,08	65,54	55,41
Kal	15,00	59,00	52,00	74,00	72,00	68,00
Ol	18,00	48,50	51,52	73,74	73,74	60,61
Sür	41,00	66,00	62,00	65,00	80,00	78,00
Ver	17,00	36,00	27,00	70,00	67,00	40,00
Yap	28,00	48,00	42,00	91,00	85,00	73,00

Çizelge 3.7 : İsim ve eylem grupları için NKA özellikleri ortalama doğruluk değerleri (%).

Sözcük Tipi	EBA	Algoritma				
		Naive Bayes	IBk	J48	FT	DVM
İsim	33,47	60,59	53,87	61,04	73,47	68,95
Eylem	23,60	46,51	43,12	65,92	67,26	58,55

Çizelge 3.5'te özetlenen isim grubuna ilişkin sonuçlar, ağaç tabanlı algoritma sınıfındaki FT algoritması ve DVM yöntemi başarımının daha yüksek olduğunu göstermektedir. Çizelge 3.6'da gösterilen eylem grubuna ilişkin sonuçlar ise NKA özellikleri için ağaç tabanlı yöntem (J48 ve FT algoritması) sonuçlarının daha yüksek başarımda elde edildiğini göstermektedir. Çizelge 3.7'de ise Çizelge 3.5 ve Çizelge 3.6 sonuçlarının yöntem bazındaki ortalaması verilmektedir. NKA özellikleri için farklı yöntemlere ilişkin ortalama sonuçlardan DVM ve ağaç tabanlı algoritma sonuçlarının daha yüksek başarımda elde edildiği gözlenmektedir. Örnek tabanlı bir yöntem olan IBk ise her iki grup için en düşük değeri vermektedir.

3.2.2 BKA özelliklerinin kullanılması

Birlikteliklerinin Kazandırdığı Anlamlar (BKA) ile sözcüklerin belirli bir bağlamda birlikte gözlenme durumları değerlendirmeye alınmıştır. Diğer deyişle sözcük kesesi özelliklerinin (BoW özellikleri) anlam belirginleştirmedeki etkisi incelenmiştir. Sözcük kesesi özellikleri konumsal özelliklerin aksine belirsizlik giderimi yapılacak paragraf veya bağlamdaki sözcüklerin, herhangi bir ilişki, sıralama ya da sözdizimi gözetilmeksizin özellik olarak ele alınmasıdır. Öncelikle metinler içinde geçen *işlev sözcükler* (stopwords) çıkarılmaktadır. İşlev sözcükler, dilde sık gözlemlenen, kullanımları çok genel olup anlamsal olarak ayırt edicilik özelliği taşımayan (ve, ile, çok, bu, şu, vd. gibi) sözcüklerdir. Bu eleme aşamasının ardından, konu açısından anlam değeri taşıyan *içerik sözcükler* (content words) kullanım sıklıklarına göre sıralanarak belirli bir eşik değerinin üzerinde kullanılan sözcükler seçilmektedir. Çalışmada Türkçe isim ve eylem grupları için derlemden işlev sözcükler çıkarıldıktan sonra en sık gözlenen belirli sayıdaki içerik sözcük özellik olarak alınmış, eğitim ve sınama verisi bu özellikler kullanılarak kodlanmıştır. Daha sonraki sınamalarda bulunan etkin değer isim ve eylem gruplarında özellik sayısı olarak kullanılmıştır.

Çalışmamızda bu özellikler aşağıdaki sıraya uygun olarak sınanmıştır:

- Tüm veri üzerinde biçimbilimsel çözümleme ve belirsizlik giderimi yapılarak, sözcük gövde biçimlerinin elde edilmesi.
- İşlev sözcüklerin derlemden çıkarılması.
- Eğitim derlemi üzerinden en yüksek kullanım sıklığına sahip içerik sözcüğün elde edilmesi.
- En yüksek kullanıma sahip ilk N sözcüğün özellik olarak seçilmesi; eğitim ve sınama verilerinin bu özellikler kullanılarak ikili yapıda kodlanması.
- Konumsal özellikler için kullanılan algoritmaların sözcük içerik özellikleri ile tekrar sınanması.

BKA özellikleri sınanırken iki farklı çalışma yapılmıştır:

- **İsim ve eylem kümesi için etkin özellik sayılarının belirlenmesi:** En sık kullanılan N içerik sözcüğüne ilişkin etkin değer, isim ve eylem grubu için ayrı ayrı belirlenmiştir. Bu değer isim grubunda ilk 100 içerik sözcük, eylem grubunda ise ilk 75 içerik sözcük olarak elde edilmiştir. Daha sonraki sınamalarda bu iki değer özellik sayısı olarak kullanılmıştır.
- **Etkin pencere aralığının isim ve eylem kümeleri için bulunması:** Çalışmanın devamında içerik özellikleri kullanıldığında etkin pencere aralığı ±5, ±10, ±15, ±30 değerleri için araştırılmıştır. Küçük pencere aralığının her iki grupta daha etkin sonuçlar verdiği gözlenmiştir.

Yapılan sınamalar sonucunda isim ve eylem gruplarındaki sözcükler için Çizelge 3.8 ve Çizelge 3.9'daki değerler elde edilmiştir. Çizelge 3.10'da ise isim ve eylem gruplarına ilişkin ortalama doğruluk oranları verilmiştir.

Çizelge 3.8 : Türkçe isim grupları için BKA özellikleri doğruluk değerleri (%).

Hedef Sözcük	EBA	NB	Ibk	J48	FT	DVM
Açık	27,00	59,00	49,00	51,00	57,00	55,00
Baskı	30,00	64,00	50,00	67,00	63,00	63,00
Baş	30,00	51,00	39,00	56,00	48,00	53,00
Derece	38,00	64,00	60,00	70,00	67,00	66,00
Dünya	18,00	37,00	25,00	29,00	34,00	27,00
El	42,00	66,00	50,00	68,00	60,00	59,00
Göz	51,00	55,00	60,00	60,00	57,00	61,00
Hat	20,00	66,40	48,50	70,80	73,10	66,40
Hava	30,00	51,00	47,00	47,00	42,00	50,00
Kaynak	30,00	62,00	40,00	66,00	62,00	58,00
Kök	28,00	69,00	44,00	69,00	76,00	68,00
Kör	53,00	56,50	38,30	61,60	59,00	63,60
Ocak	26,00	62,50	50,00	58,90	64,30	66,90
Yaş	52,00	63,00	58,00	64,00	71,00	69,00
Yüz	27,00	45,00	36,00	53,00	50,00	41,00

Çizelge 3.9 : Türkçe eylem grupları için BKA özellikleri doğruluk değerleri (%).

Hedef Sözcük	EBA	NB	Ibk	J48	FT	DVM
Aç	15,00	37,00	27,00	31,00	29,00	36,00
Al	23,00	38,00	37,00	35,00	34,00	32,00
At	22,00	44,00	29,00	43,00	44,00	38,00
Bak	20,00	36,00	37,00	37,00	38,00	33,00
Çevir	27,00	52,00	39,00	56,00	46,00	48,00
Çık	16,00	42,00	36,70	41,80	42,90	41,80
Geç	23,00	50,00	41,80	50,00	58,20	51,02
Gel	24,00	32,32	31,31	33,33	33,33	39,39
Gir	39,00	61,00	42,00	58,00	54,00	57,00
Gör	26,00	29,53	24,83	34,22	29,50	31,54
Kal	15,00	42,00	35,00	54,00	50,00	41,00
Ol	18,00	53,00	41,00	51,00	45,00	44,00
Sür	41,00	66,00	63,00	71,00	67,00	66,00
Ver	17,00	35,00	35,00	40,00	42,00	42,00
Yap	28,00	54,00	38,00	65,00	58,00	53,00

Çizelge 3.10 : İsim ve eylem grupları için BKA özellikleri ortalama doğruluk değerleri (%).

Sözcük Tipi	EBA	Algoritma				
		Naive Bayes	IBk	J48	FT	DVM
İsim	33,47	58,09	46,32	59,42	58,89	57,79
Eylem	23,60	44,79	37,18	46,69	44,73	43,58

BKA özellikleri kullanılarak isim grubu üzerinde yapılan çalışmalar sonucunda ağaç tabanlı yöntem sonuçlarının daha yüksek başarımlı olduğu, bununla birlikte Naive Bayes ve DVM sonuçlarının da yakın düzeyde olduğu gözlenmektedir. Eylem grubunda elde edilen sonuçlar da benzer özellikte olup ağaç tabanlı algoritmalar ve Naive Bayes yöntem sonuçları birbirine yakın bulunmuştur. Her iki özellik grubunda örnek tabanlı yöntem sonuçları NKA özelliklerinde olduğu gibi diğer yöntemlere oranla daha düşük bulunmuştur.

3.2.3 Özellik kümelerinin birlikte kullanılması

Çalışmamızda sonuçları daha önce ayrı ayrı elde edilmiş olan BKA özellikleri ile NKA özellikleri birlikte kullanılarak daha yüksek başarımda sonuçlara ulaşılmıştır. BKA özelliklerinde önceki deneylere paralel olarak isim grubu için SK özellik sayısı 100 olarak alınmış, eylem grubunda ise 75 özellik değerlendirmeye alınmıştır. Deneylerde NB, IBk, J48, FT ve DVM yöntemleri sınanmıştır. IBk algoritması *k* en yakın komşu değeri parametresi 1 olarak alınmıştır. DVM yöntemindeki *C* parametresi de 1 olarak alınarak diğer tüm parametreler için Weka'daki standart parametre değerleri kullanılmıştır. NKA ve BKA özellik kümelerine ayrıca CFS özellik azaltımı uygulanması ile elde edilen sonuçlar Çizelge 3.11 ve Çizelge 3.12'de gösterilmiştir.

Çizelge 3.11 : Türkçe isim grupları için doğruluk değerleri - Tüm özellikler (%).

Hedef Sözcük	EBA	NB	Ibk	J48	FT	SVM
Açık	27,00	62,50	71,88	47,92	68,75	70,83
Baskı	30,00	70,41	68,37	66,33	74,49	70,41
Baş	30,00	74,49	77,55	74,49	77,55	79,59
Derece	38,00	86,87	89,90	84,85	91,92	91,92
Dünya	18,00	69,39	68,37	37,76	60,20	72,45
El	42,00	76,29	80,41	77,32	78,35	78,35
Göz	51,00	80,81	76,77	78,79	85,86	80,81
Hat	20,00	83,46	85,71	81,20	84,96	84,96
Hava	30,00	65,66	69,70	59,60	67,68	69,70
Kaynak	30,00	73,47	71,43	73,47	76,53	73,47
Kök	28,00	79,80	80,81	70,71	79,80	79,80
Kör	53,00	88,89	82,83	86,87	87,88	88,89
Ocak	26,00	74,55	75,45	47,27	78,18	76,36
Yaş	52,00	89,58	92,71	83,33	91,67	91,67
Yüz	27,00	74,49	76,53	68,37	72,45	74,49

Çizelge 3.12 : Türkçe eylem grupları için doğruluk değerleri - Tüm özellikler (%).

Hedef Sözcük	EBA	NB	Ibk	J48	FT	DVM
Aç	15,00	54,84	62,37	52,69	62,37	63,44
Al	23,00	50,53	42,11	61,05	56,84	58,95
At	22,00	45,83	59,38	79,17	76,04	78,13
Bak	20,00	49,47	54,74	53,68	50,53	51,58
Çevir	27,00	78,89	83,33	73,33	80,00	81,11
Çık	16,00	61,86	67,01	71,13	73,20	69,07
Geç	23,00	77,32	81,44	73,20	81,44	81,44
Gel	26,00	56,12	61,22	56,12	62,24	63,27
Gir	39,00	67,74	69,89	72,04	70,97	76,34
Gör	26,00	53,10	66,90	66,21	71,03	71,03
Kal	15,00	82,29	82,29	78,13	78,13	81,25
Ol	18,00	67,35	76,53	79,59	78,57	77,55
Sür	41,00	81,82	88,89	78,79	81,82	88,89
Ver	17,00	70,10	68,04	70,10	80,41	80,41
Yap	28,00	69,57	81,52	89,13	86,96	88,04

NKA ve BKA özellikleri birlikte kullanıldığında elde edilen sonuçlara ilişkin başarım FT ve DVM yöntemleri için daha yüksek olmuştur. Eylem grubu için ise FT yöntem sonuçları en yüksek başarımda elde edilmiş, J48 ve DVM yöntem sonuçları ise birbirine yakın bulunmuştur. Özellik azaltımı uygulanan deney sonuçlarında ise isim ve eylem grupları için FT ve DVM yöntem başarımları diğer algoritma sonuçlarına oranla daha yüksek bulunmuştur.

3.2.4 Denetimli yöntem sonuçlarının değerlendirilmesi

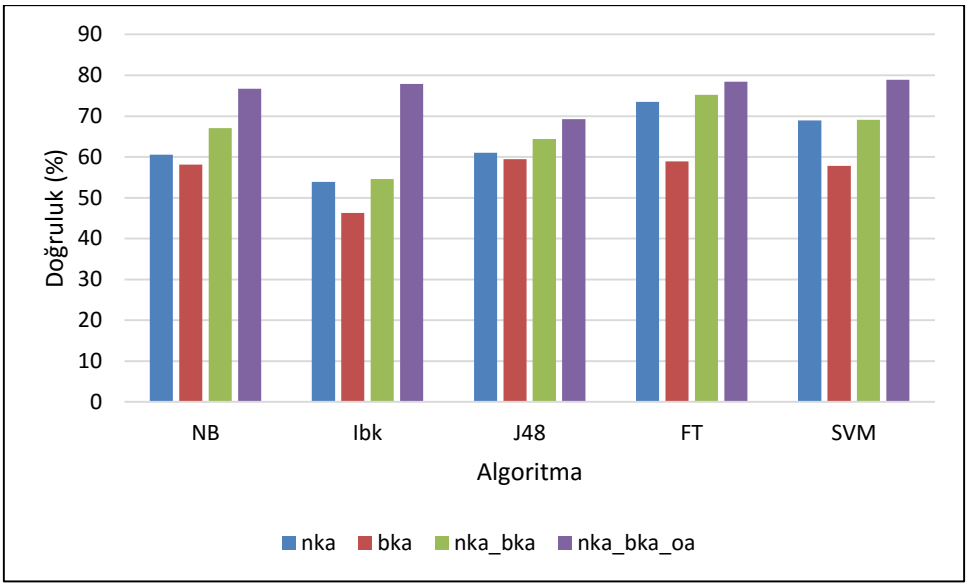
Bu bölümde denetimli yöntemlerde kullanılan iki farklı özellik kümesine ilişkin doğruluk yüzdesi değerlerinin karşılaştırılması yapılmıştır. Bu özellikler sırasıyla; sözcük kesesi ve konumsal özellik (BKA ve NKA özellikleri) gruplarından oluşmaktadır.

Özellik gruplarının ayrı olarak uygulanmasından sonra tüm özellik seçimlerinden yöntem bazında elde edilen ortalama doğruluk değerleri Çizelge 3.13 ve Çizelge 3.14'te sırasıyla Türkçe isim ve eylem grubu için verilmiştir. Şekil 3.6 ve Şekil 3.7 ise Çizelge 3.13 ve Çizelge 3.14'e ilişkin grafik gösterimi yansıtmaktadır. Çizelge 3.13 ve Çizelge 3.14'te farklı renklerle gösterilen özellikler sırası ile aşağıda verilmektedir.

- NKA özellikleri
- BKA özellikleri
- NKA + BKA özellikleri
- NKA + BKA + OA (CFS özellik azaltımı).

Çizelge 3.13 : İsim grubu için karşılaştırmalı ortalama doğruluk değerleri (%).

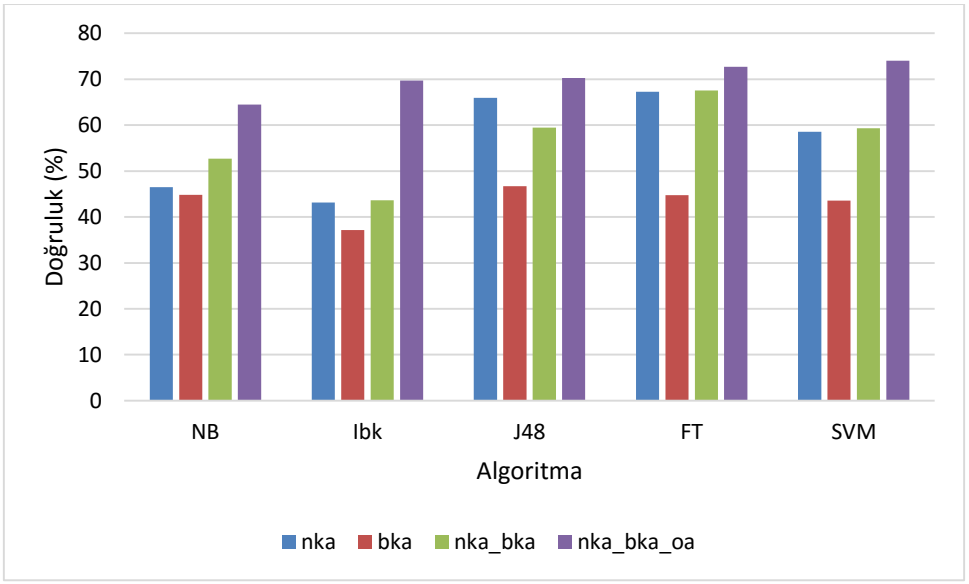
	NB	**Ibk**	**J48**	**FT**	**DVM**
nka	60,59	53,87	61,04	73,47	68,95
bka	58,09	46,32	59,42	58,89	57,79
nka_bka	67,05	54,62	64,37	75,18	69,14
nka_bka_oa	76,71	77,89	69,22	78,42	78,91



Şekil 3.6 : İsim grubu için 4 farklı özelliğe ilişkin doğruluk değerleri (%).

Çizelge 3.14 : Eylem grubu için karşılaştırmalı ortalama doğruluk değerleri (%).

	NB	**Ibk**	**J48**	**FT**	**DVM**
nka	46,51	43,12	65,92	67,26	58,55
bka	44,79	37,18	46,69	44,73	43,58
nka_bka	52,68	43,63	59,42	67,53	59,29
nka_bka_oa	64,46	69,71	70,29	72,70	74,03



Şekil 3.7 : Eylem grubu için 4 farklı özelliğe ilişkin doğruluk değerleri (%).

Yapılan bu testlerde hem isim hem de eylem grubu için kullanılmakta olan özellik seçimlerinin en yüksekten düşüğe başarı sıralaması aşağıda verilmektedir.

1. bka + nka (CFS özellik indirgemesi ile)
2. bka + nka
3. nka özellikleri
4. bka özellikleri

Elde edilen sonuçlar değerlendirildiğinde, her iki özellik kodlaması birlikte kullanıldığında en iyi sonuçların elde edildiğini, ayrı ayrı değerlendirme yaptığımızda ise NKA özelliklerinin, BKA özelliklerine oranla daha iyi sonuçlar verdiği görülmüştür.

3.2.5 Sonuçların diğer çalışma sonuçları ile karşılaştırılması

Yapılan çalışmalarda elde edilen değerler, Türkçe SABG konusunda önceden yapılmış bir çalışmanın sonuçları ile karşılaştırılmıştır (Orhan, 2006). Sözü edilen çalışmada ODTU-Sabancı derlemi kullanılarak işaretli olan özelliklerden faydalanılmıştır (Say ve diğ, 2002). Bu çalışmadaki özellikler; "kök", "tür", "hal eki", "iyelik", "ilişki", özelliklerinin önceki, sonraki ve hedef sözcük için değerlendirilmesi ile elde edilmiştir. Bu özellikler ile farklı kombinasyonlar oluşturulmuştur. Sözü edilen özelliklerin çıkarıldığı ODTÜ-Sabancı derlemine ilişkin örnek yapısındaki bir cümleye ilişkin birimler Şekil 3.8'de verilmektedir.

```
IG='[(1,"baş+Noun+A3sg+P3sg+Acc")]' REL="[2,1,(OBJECT)]"> Başını </W>
IG='[(1,"kaşı+Verb+Pos+Prog1+A3sg")]' REL="[3,1,(SENTENCE)]"> kaşıyor </W>
IG='[(1,",+Punc")]' REL="[4,2,(COORDINATION)]"> , </W>
IG='[(1,"karar+Noun+A3sg+Pnon+Nom")(2,"Adj+Without")]' REL="[5,1,(SENTENCE)]"> kararsız </W>
IG='[(1,".+Punc")]' REL="[,( )]"> . </W>
```

Şekil 3.8 : ODTÜ-Sabancı derlemi XML örneği.

İki çalışma arasındaki önemli farklardan biri, önceki çalışmada kullanılan özellikler arasında bu çalışmada kullanılmamış olan sözdizimsel özelliklerin de bulunuyor olmasıdır. Ağaç yapılı derlemden isim, eylem ve geri kalan sözcüklerden oluşan üç grup belirlenmiştir. Belirsizlik içeren bu sözcükler için ağaç yapılı derlemden örnekler alınmıştır. Seçilen belirsiz sözcükler 10 Türkçe isim ve 10 eylemden oluşmakta olup bu çalışmadaki sözcükler ile kısmen örtüşmektedir. Örnekler bu çalışma kapsamında daha önce uygulanan ön işleme adımlarından geçirilmiştir. ODTÜ-Sabancı ağaç yapılı derlem ve daha önce üzerinde çalıştığımız Hedef Sözcük Derlemine dahil olan sözcükler aşağıda verilmektedir:

ODTÜ-Sabancı ağaç yapılı derlem:

İsim: *ara*, ***baş***, ***el***, ***göz***, *kız on, sıra, üst, yan, yol*

Eylem: ***al, bak***, *çalış,* ***çık***, ***geç***, ***gel, gir***, *git*, ***gör***, *konuş*

Hedef Sözcük Derlemi:

İsim: *açık,* ***baş***, *baskı, derece, dünya*, ***el, göz***, *hat, hava, kaynak, kök, kör, ocak, yaş, yüz.*

Eylem: *aç,* ***al***, *at*, ***bak***, *çevir,* ***çık***, ***geç***, ***gel***, ***gir***, ***gör***, *kal, ol, sür, ver, yap*

Önceki çalışmalarla karşılaştırma yapılabilmesi için ODTÜ-Sabancı ağaç yapılı derlem kullanılarak hazırlanmış olan *Turkish Lexical Sample Task* (TLSD) verisi değerlendirmeye alınmıştır. Derlemin isim ve eylem grubuna dahil olan tüm sözcük örnekleri çalışmamızda kullanılan konumsal özellikler ile kodlanmış ve değerlendirmede ince anlamlı sözcükler göz önünde bulundurulmuştur. Türkçe isim grubu için önceki çalışmada verilen tutturma ve bulma değerleri sırası ile 0,15 ve 0,50'dir. Eylem grubu için benzer değerler sırasıyla 0,10 ve 0,38 olarak kaydedilmiştir. Çalışmamızda aynı veri kümesini kullandığımızda elde ettiğimiz sonuçlar Çizelge 3.15'de verilmektedir. Elde etiğimiz tutturma değerleri isim ve eylem grubu için sırasıyla 0,46 ve 0,54 oranında daha yüksektir. Bulma değerleri ise bu çalışmada bulunan değerlere göre 0,07 ve 0,28 oranında daha yüksektir. Sonuçlar değerlendirilirken karşılaştırılma yapılan çalışmanın aksine çalışmamızda sözdizim özelliklerinin kullanılmadığı da göz önünde bulundurulmalıdır. Elde edilen değerler birbirine yakın aralıklarda bulunmakla birlikte eylem grubuna ilişkin sonuçlar daha başarılıdır. Eylem grubunun belirsizliği isimlere oranla yüksek başarımda elde edilmiştir. Eylem grubu sonuçlarındaki sözü edilen durumda eğitim verisi boyutu ve örnek uzunluklarının eylem kümesinde daha büyük olmasının etkili olduğu düşünülmektedir.

Çizelge 3.15 : ODTÜ-Sabancı ağaç yapılı derlem üzerinde elde edilen ortalama tutturma – bulma değerleri.

	Tutturma	**Bulma**
İsim	0,61	0,57
Eylem	0,64	0,66
Ortalama	**0,63**	**0,62**

3.2.6 Denetimli yöntemler üzerinde yapılan diğer çalışmalar

3.2.6.1 Biçimbilimsel özellik gruplarının anlam belirginleştirme üzerinde etkisinin incelenmesi

Bu bölümde 3.2.1'de biçimbilimsel özellikler üzerinde yapılan çalışmalara ek olarak farklı biçimbilimsel özellik gruplarının anlam belirginleştirme üzerine etkileri incelenmiştir. Bunun için aşağıda verilmekte olan ve Türkçede yaygın olarak kullanılan üç çekimsel grup ele alınmıştır. Bu gruplar, iyelik, sahiplik ve durum ekleridir.

- Kişi/Sayı (Number/Person) uyumu: +A1sg, +A2sg, +A3sg, +A1pl, +A2pl, +A3pl.
- Sahiplik (Possessive) Ekleri: +P1sg, +P2sg, +P3sg, +P1pl, +P2pl, +P3pl.
- Durum (Case): +Nom, +Acc, +Dat, +Abl, +Loc, +Gen, +Ins.

Yukarda verilen çekim grupları kullanılarak farklı özellik kombinasyonları için denemeler yapılmıştır. Kullanılan özellik grupları ve isimleri aşağıda verilmektedir.

a. Tüm özellikler - TO

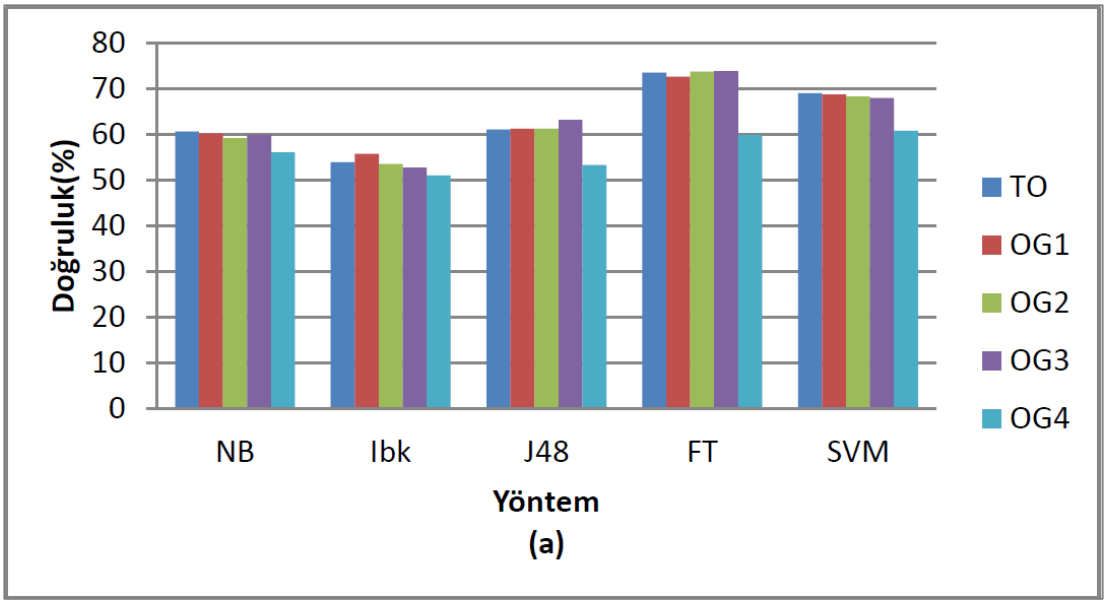
b. İyelik ekleri dışındaki tüm özellikler –OG1

c. Sahiplik ekleri dışındaki tüm özellikler – OG2

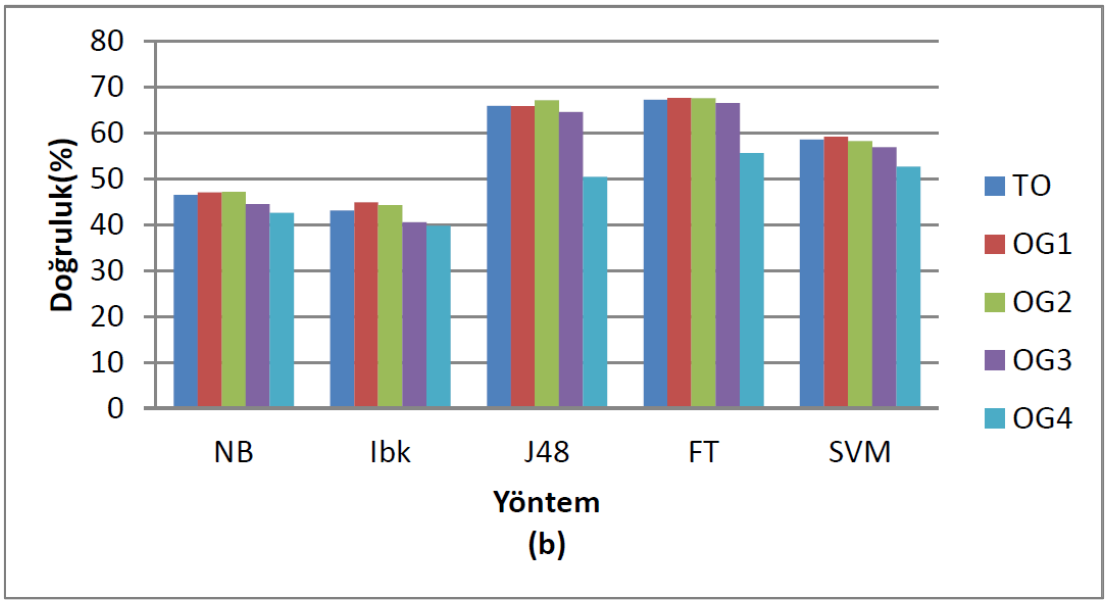
d. Durum ekleri dışındaki tüm özellikler – OG3

e. Sözcük gövdesi dışındaki tüm özellikler – OG4

"Tüm özellikler" dışındaki gruplarda belirtilen özellikler, tüm özelliklerin olduğu gruptan çıkarılarak sonuçlardaki başarım değişimi gözlenmiştir. Elde edilen bulgular sözcük gövdesinin özellik kümesinden çıkarıldığı durumun isim ve eylem gruplarında ve tüm algoritmalar için en düşük doğruluk değerini verdiğini göstermiştir. Bununla birlikte isim grubunda TO, OG1, OG2 ve OG3 özelliklerinin kullanımı ile elde edilen başarım ayırt edici bulunmamıştır. Eylem grubunda ise OG4 kümesinden sonra en düşük performans durum eklerinin kullanılmadığı OG3 grubu için gözlenmiştir. OG3 özellikleri ile elde edilen başarım tüm algoritma sınıflarında en düşük ikinci değeri vermiştir. Şekil 3.9 ve Şekil 3.10'da sözü edilen özellik kombinasyonları ile elde edilen sonuçlar sırasıyla isim ve eylem grupları için görülmektedir.



Şekil 3.9 : Özellik gruplarının Türkçe isimler için anlam belirsizliği gidermede etkisi.



Şekil 3.10 : Özellik gruplarının Türkçe eylemler için anlam belirsizliği gidermede etkisi.

3.3 Bölüm Sonucu

Denetimli yöntemler üzerinde yapmış olduğumuz çalışmanın sonucunda aşağıdaki değerlendirmeler yapılmıştır:

- Denetimli yöntemler kapsamında çalışmamızın ilk aşamasında anlam belirsizliği ortadan kaldırılmak istenen hedef sözcük örneklerinin yer aldığı paragrafları içeren özel bir derlem hazırlanmıştır. HSD adını verdiğimiz

derlemde 15 isim ve 15 eylem soylu sözcük ile her bir sözcük için de yaklaşık olarak 100 örnek metin bulunmaktadır. Bu sözcüklerin kesin anlamlarının belirlenmesinde 5 uzman kişi çalıştırılmıştır ve 5 kişinin önerdiği anlamlar oylama yöntemi ile keskinleştirilmiştir. Eğitim kümesi verilerini hazırlamanın yoğun insan emeği gerektirdiği anlaşılmıştır.

- Denetimli yöntemler üzerinde çalışmalarımızın sonucunda bu yöntemlerin belirsizlik gidermede başarılı olduğu gözlenmiştir. Bununla birlikte yöntemlerin anlamları etiketlenmiş derlemlere bağımlı olduğu bilinmektedir. Özellikle kaynak kısıtı bulunan diller için bu sorunu aşabilmenin bir yolu, bu bağımlılığı ortadan kaldıracak yöntemlere yönelinmesidir. Çalışmamızda eğitim kümesinin hazırlanmasının çok emek yoğun olduğu öğrenilmiştir. Ancak bu çalışmalar sonunda KÖ ve SK özellikleri gibi farklı özellik gruplarının anlam belirginleştirme etkisi incelenmiş, aynı zamanda daha etkin olan özellikler elde edilmiştir. Bu çalışmalar sırasında hedef sözcük anlamı belirginleştirmede etkili olduğu düşünülen pencere boyu gibi ölçütler de göz önünde bulundurulmuş ve en etkin genişlik elde edilmiştir.
- KÖ özelliklerinin bir alt grubu olan kişi ekleri, sahiplik ve durum eklerinin etkisi bu bölüm kapsamında incelenmiştir. Özelliklerin etkisini belirlemek için tanımlanan 4 özellik grubu tüm özelliklerden çıkarılarak, ilgili özelliğin başarım üzerindeki etkisi gözlenmiştir. Yapılan deneyler sonucunda, hem isim hem de eylem gruplarında sözcük gövdesini çıkardığımız durumda en fazla kayıp yaşandığı görülmüştür. KÖ özellikleri etkin özellikler olmasına karşın incelen alt grupların etkisi ayırt edici düzeyde bulunmamıştır.
- Çalışmanın ilerleyen bölümünde çizge tabanlı ve denetimsiz bir algoritma gerçekleştirilmiş, anlam belirginleştirmede etkin olduğu gözlemlenen parametrelerin iyileştirilmesi üzerinde çalışılmıştır. Denetimli yöntemlerde edinilen bilgilerin özellikle belirsizlik derecesi çok yüksek olan eylem grubu için ilerdeki çalışmalarda, denetimsiz yöntemler kapsamında kullanılması düşünülmüştür.

4. TÜRKÇE İÇİN DENETİMSİZ ÇİZGE TABANLI BİR YÖNTEM GELİŞTİRİLMESİ

Çizge tabanlı yöntemler özellikle bilgi tabanlı SABG (Mihalcea, 2005; Navigli ve Velardi, 2005), özet çıkarma (Erkan ve Radev, 2004; Mihalcea ve Tarau, 2004) gibi DDİ alanı kapsamındaki pek çok çalışmada kullanılan yöntemlerdir. Denetimli yöntemlerle yapılan çalışmalardan başarılı sonuçlar elde edilmekte ancak yöntemlerin uygulanabilmesi için anlam işaretli derlemlere gereksinim duyulmaktadır. Ayrıca denetimli yöntemlerde hedef sözcük için sonlu bir anlam listesi olduğu varsayımı kabul edilmektedir. Bu durum geliştirilen yöntemlerin ve sözcük anlam envanterlerinin ölçeklenebilirliği, yöntemlerin farklı alanlara taşınabilmesi gibi konularda zorluklara neden olmaktadır. Sözü edilen sebeplerden ötürü sonlu sayıda sözcük anlamının kullanıldığı SABG yöntemleri dilbilimcilerin ve DDİ alanındaki araştırmacıların son zamanlarda kaçındığı yaklaşımlardır. Denetimli yöntemlerin aksine denetimsiz yöntemlerde önceden belirlenmiş bir anlam listesine bağımlı kalınmamakta, derlemlerin kendi anlamlarını dinamik bir yapıda çıkarabileceği fikri benimsenmektedir. Denetimsiz çizge tabanlı yöntemler ile yapılan çalışmalar (Agirre ve Soroa, 2009; Mihalcea ve diğ, 2004; Navigli, 2006; Sinha ve Mihalcea, 2007; Tsatsaronis ve diğ, 2007), özellikle çizge yöntemlerinin denetimli yöntemlerle elde edilen başarım farkını kapatması nedeniyle ivme kazanmıştır.

4.1 HyperLex Algoritması

HyperLex algoritması (Veronis, 2004) derlemler ya da metinler içerisindeki sözcük kullanımlarını herhangi bir sözlüğe başvurmadan otomatik olarak çıkartan bir yöntemdir. Bu yönüyle tamamen derlem tabanlı bir yaklaşımdır. Algoritma ile sözcük kullanımları belirlenirken sözcüklerin birlikte gözlenme durumlarından faydalanılarak *Küçük Dünya* özelliği (Watts ve Strogatz, 1998) gösteren bir çizge yapısı oluşturulmaktadır. Yöntemin daha önce geliştirilmiş olan ve sözlük kullanımının olmadığı vektör tabanlı benzer yöntemlerden temeldeki farkı, çok seyrek gözlenen kullanımların çizge oluşturulurken elenmesidir. Bu eleme işlemi,

ilerleyen bölümlerde tanımlanmakta olan merkez düğümlerin ve yüksek yoğunluklu bileşenlerin kullanımı ile sağlanmaktadır. Algoritma ilk olarak Web bilgiye erişim sistemleri üzerinde ve belirsizlik derecesi yüksek sözcükler için sınanmıştır.

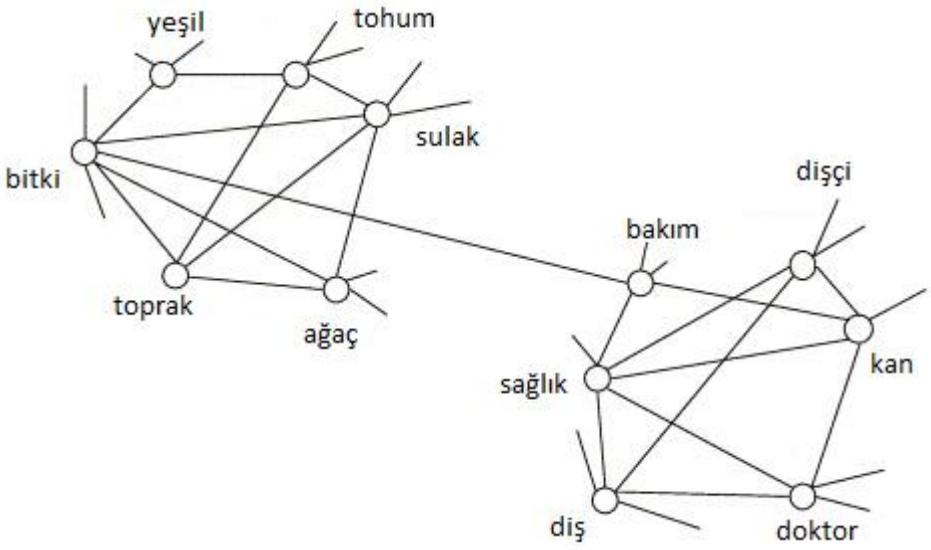
Bilgiye erişim sistemi kapsamında yapılan çalışma için belirsizlik derecesi oldukça yüksek olarak değerlendirilen 10 sözcük belirlenmiştir (Veronis, 2004). Her sözcük için Web sayfalarından derlenen alt derlemler oluşturulmuştur. Sözcüklerin öncelikle tekil ve daha sonra çoğul halleri sorgu olarak kullanılmıştır. Elde edilen sonuçlardan iki defa gözlenenler ve "Sayfa Bulunamadı" hatası gibi sözcüğü içermeyen sonuçlar elenmiştir.

Seçilmiş olan 10 adet hedef sözcüğü içeren paragraflar elde edilerek *Cordial Analyser* (http://www.synapse-fr.com) ile etiketlenmiştir. Aynı zamanda bazı son işleme programlarından da faydalanılmıştır. Çalışmada sadece isim ve sıfatlar değerlendirmeye alınmıştır. Buna neden olarak İngilizce *start* (*başlamak*) , *can* (*-ebilmek, -abilmek*) gibi eylemlerin çok genel kullanımlara sahip olmaları ve performansı oldukça olumsuz yönde etkilemeleri gösterilmiştir. Ancak bu durumun geçici bir yaklaşım olduğu, gelecekte yapılan çalışmalarla üstesinden gelineceği kaydedilmiştir. Paragraflar sonraki aşamada filtrelenerek, belirteç ve edatlar ayrılmış, aynı zamanda işlev sözcükler ve Web'in yapısıyla ilgili olan (menü, home, link, http vd.) bazı sözcükler çıkarılmıştır. Tüm derlemdeki toplam örnek sayısı 10'dan az olan sözcükler elenmiştir. Filtreleme işleminden sonra kalan sözcük sayısının 4'ten az olduğu bağlamlar ihmal edilmiştir.

Birliktelik matrisi filtrelenmiş olan alt derlemlerden oluşturulmuştur. Aynı paragrafta geçen sözcük çiftlerinin birlikte gözlendiği kabul edilmiştir. Değerlendirmeye alınacak sözcük çiftleri için birlikte gözlenme sayısının alt değeri 5 olarak seçilmiştir. Oluşturulan derlemler üzerinde yapılan çalışmalar sonucunda algoritma ile ilgili olarak döndürülmesi beklenen sonuçlardan çok küçük bir bölümün atlandığı görülmüştür. Tutturma ve bulma değerlerinin (%73 sınır başarım değeri için sırasıyla %97 ve %82 tutturma ve bulma değerleri) oldukça yüksek başarımda elde edildiği gözlenmiştir.

4.2 Sözcükler ve Küçük Dünya Modeli

İlerleyen bölümde açıklandığı gibi anlam belirsizliği taşıyan her sözcük için bir çizge oluşturulması mümkündür. Çizgedeki düğüm noktalarını anlam belirsizliğine sahip sözcükle birlikte gözlenen diğer sözcükler oluşturmaktadır. Sözcüklerin birlikte gözlendiği bağlam, hedef sözcüğü çevreleyen belirli bir pencere boyu, paragraf ya da cümle olabilmektedir. Birlikte gözlenen her A ve B gibi iki düğüm noktası ise bir kenar ile birbirine bağlamaktadır. Örneğin Şekil 4.1'de anlam belirsizliği taşıyan *kök* sözcüğü için oluşturulmuş çizge yapısında ağaç ve sulak sözcükleri aşağıda verildiği gibi benzer bağlamlarda geçtiği için birbirlerine bağlı olacaktır.



Şekil 4.1 : Kök sözcüğü için çizge örneği.

*Arazi **sulak** olduğu için **bitki kökleri** derine uzanmıyordu.*

Bu tür yapıların çizge teoreminde oldukça önemli bir araştırma konusu olan *Küçük Dünya* özelliği gösterdiği bilinmektedir (Watts ve Strogatz, 1998). Watts ve Strogatz (1998), küçük dünya yapısı tanımlarken iki ölçüt ortaya koymuştur; *Karakteristik Yol Uzunluğu* (L) ve *Kümeleme Katsayısı* (C). L, çizgede herhangi iki düğüm noktası arasındaki en kısa yol değerlerinin ortalamasıdır. Bir çizge yapısındaki *i* ve *j* gibi iki düğüm noktası arasındaki en kısa mesafenin d_{min} (i, j) ve N'nin toplam düğüm noktası olduğu kabulü ile denklem 4.1 ortaya konmaktadır.

$$L = \frac{1}{N}\sum_{i=1}^{N} d_{min}(\mathrm{i},\mathrm{j}) \qquad (4.1)$$

Her *i* düğüm noktası için, düğüm noktasının komşularının Γ(i) bağlantılarının E(Γ(i)) oranına eşit olan yerel C_i katsayısı tanımlanmaktadır. Örneğin *i* düğümünün 4 komşusu bulunuyorsa, en fazla bağlantı sayısı $\binom{|\Gamma(i)|}{2} = 6$ olacaktır. Eğer bu bağlantılardan 5 tanesi mevcutsa, Ci = 5 / 6 ≈ 0,83 olarak bulunacaktır. C katsayısının evrensel değeri ise yerel değerlerin ortalaması olarak elde edilmektedir.

$$C = \frac{1}{N}\sum_{i=1}^{N} \frac{|E(\Gamma(i))|}{\binom{|\Gamma(i)|}{2}} \tag{4.2}$$

Bu katsayı çizge yapısının tamamen bağlantısız veya tam bağlı olma durumuna göre 0 ve 1 değerleri arasında değişmektedir. Ortalama derecesi *k* olan (E çizgedeki kenar sayısı olmak üzere, düğüm noktası başına ortalama kenar sayısı E/N) rastgele bir çizge yapısını ele aldığımızda L_{rand} ve C_{rand} aşağıdaki gibi ifade edilmektedir.

$$L_{rand} \sim \log(N) / \log(k) \tag{4.3}$$

$$C_{rand} \sim 2\ k / N \tag{4.4}$$

Örneğin, 1000 düğüm noktasına sahip 10000 kenarlı bir rastgele çizgenin ortalama k değeri 10 olacaktır. Karakteristik yol uzunluğu L_{rand} log(1000)/log(10) = 3, kümeleme katsayısı ise C_{rand} 10/1000 = 0,01 olarak bulunacaktır. Watts ve Strogatz (1998), küçük dünya modelini aşağıdaki ilişkilerle karakterize etmiştir.

$$L \sim L_{rand} \tag{4.5}$$

$$C >> C_{rand} \tag{4.6}$$

Denklem 4.4 sabit bir *k* ortalama değeri için, düğüm noktalarının üstel olarak artabildiğini göstermektedir. Bununla birlikte yol uzunluğu düzlemsel olarak artış göstermektedir. Bu durum küçük dünya modelini ilk olarak ortaya koyan Milgram (1967)'ın tezini desteklemektedir. Bir sosyal ağ çizge yapısında yer alan herhangi bir birey evrende milyarlarca birey olması durumunda bile çizgedeki herhangi bir diğer bireyden en fazla 6 adım uzaklıkta olmaktadır. Küçük dünya modeli ve rastgele çizge arasındaki fark denklem 4.5 ile gösterilmektedir (Veronis, 2004).

HyperLex yönteminin farklı anlam kullanımlarını temsil eden bir çizge yapısı oluşturmadaki kullanımının SABG alanına uyarlanabileceği düşüncesi ile İngilizce için yapılan bir çalışmada, yöntemin diğer denetimsiz yöntemlere nazaran daha iyi

sonuçlar verdiği görülmüştür (Agirre ve diğ, 2006). Çalışmada, Senseval-2 İngilizce sözcüksel örnek verisi (S2LS) üzerinden HyperLex parametrelerini iyileştirmek üzere yarı denetimli bir çalışma yapılmıştır. Elde edilen anlamlar ile S2LS resmi eğitim verisinin anlamları arasında bir eşleşme sağlanmıştır. Sonraki aşamada Senseval-3 (S3LS) İngilizce sözcüksel örnekler verisi için en iyi parametre değerleri benzer yarı denetimli yöntem ile kullanılmıştır. Ek olarak, sözü edilen çalışmada PageRank çizge tabanlı algoritması (Brin ve Page, 1998) SABG alanına uyarlanarak merkez düğümlerin elde edilmesinde farklı yöntemlerin karşılaştırılması sağlanmıştır.

4.3 Çizge Tabanlı Yöntemin Geliştirme Aşamaları

Çalışmamızda geliştirilen çizge yönteminde HyperLex algoritması temel alınarak benzer yapıda bir algoritma geliştirilmiştir. Yöntem dahilinde, çizge yapısının oluşturulması, merkez düğüm seçimi, yüksek yoğunluklu bileşenlerin belirlenmesi, düğüm noktası ağırlıklandırmaları ve belirsizlik giderimi gibi alt bölümler yer almaktadır. Çizge yapısı, HSD kapsamındaki aday sözcükler için, örneklerde yer alan sözcük çiftleri ele alınarak oluşturulmuştur. Çalışmamızda sözcük anlamları olarak çizge yapısındaki bağlantı düzeyi yüksek merkez düğümlerin bulunması hedeflenmiştir. Diğer bir deyişle merkez düğümler, çizge yapısındaki kümelerde diğer düğüm noktalarına oranla daha fazla bağlantıya sahip aynı zamanda kümeyi daha iyi temsil etme yeteneğindeki düğüm noktaları olarak karşımıza çıkmaktadır. Merkez düğümler sözcüklerin temel kullanımlarını, dolayısıyla sözcük anlamlarını ortaya koymaktadır.

4.3.1 Birliktelik çizgesinin oluşturulması

Çizge yapısının oluşturulmasında derlem olarak, HSD kullanılmış, Türkçe isim grubu için çizge yapıları elde edilmiştir. Birliktelik çizgesinin oluşturulmasında belirsizliği ortadan kaldırılmak istenen her hedef sözcük için, HSD'deki hedef sözcüklere ilişkin örnekler kullanılmıştır. Örnek paragraflardaki işlev sözcükler çıkarılmıştır. Derlemde toplam sıklık değeri 5'in altında olan sözcükler çizge yapısı oluşturulurken değerlendirme dışı bırakılmıştır. Her hedef sözcük için yaklaşık olarak 100 örnek içeren ve her birinde anlamı belirginleştirilmek istenen sözcüğün geçtiği derlemimizdeki sözcük birliktelikleri saptanmıştır. Bu işlem için hedef

sözcük haricinde aynı paragrafta yer alan sözcük çiftleri birbirine bağlı ve aralarında bir kenar bulunan düğüm noktaları olarak sayılmıştır.

4.3.2 Ağırlıklandırma

Önceki bölümde sözü edilen düğüm noktalarının birlikte gözlenme sıklıklarının hesaplanmasında formül 4.7'den faydalanılmıştır. $freq_{ij}$ metinde w_i ve w_j sözcüklerinin birlikte gözlenme sıklığını, $freq_j$ ise w_j'nin gözlenme sıklığını göstermektedir. P_{ij} ve P_{ji} olasılıklarından büyük olan değer formül 4.8'de kullanılarak ağırlık değerleri elde edilmiştir.

$$P(i|j) = \frac{freq_{ij}}{freq_j} \text{ ve } P(j|i) = \frac{freq_{ij}}{freq_i} \tag{4.7}$$

$$w_{ij} = 1 - max\{P(w_i|w_j), P(w_j|w_i)\}, \tag{4.8}$$

Ağırlık değeri iki sözcüğün birbiri ile ne derece sıkı bir ilişki içinde olduğunu göstermektedir. Bu değerin 1'den çıkarılması ile birlikte gözlenme olasılığı yüksek sözcükler için 0'a yakın değerler elde edilmekte, birlikte görülme olasılığı düşük olan sözcükler için ise bulunan w_{ij} değeri 1'e yakın olmaktadır. Diğer deyişle sözcükler birbiri ile sıkı bir ilişkiye sahipse bu değer sıfıra, aksi durumda ise bire yakın olmaktadır. İki sözcük her zaman birlikte gözleniyorsa ağırlık 0, hiç bir zaman birlikte gözlenmiyorsa 1 değerini almaktadır. Çalışmamızda bu durum göz önünde bulundurularak belirli bir ağırlık değerinin üzerindeki kenar bileşenleri değerlendirme dışı bırakılmış ve çizgeye dahil edilmemiştir. Sözcük çiftlerinin (baş-ağrı ve baş-inek) anlamsal uzaklığının ve birlikte gözlenme oranının hesaplanması ile ilgili bir örnek Çizelge 4.1'de gösterilmektedir. Çizelge 4.1'te verilen *inek* sözcüğünün tüm örnekleri *baş* sözcüğü ile birlikte gözlenirken, *ağrı* sözcüğünün belirli sayıda örneğinin *baş* sözcüğü ile birlikte gözlendiği görülmekte ve çizgedeki bağlantılara atanacak değerler örnekte gösterildiği gibi elde edilmektedir.

Çizelge 4.1 : Örnek sözcük çiftlerinin birlikte gözlenme sıklıkları.

	BAŞ	**~BAŞ**	**Toplam**		**BAŞ**	**~BAŞ**	**Toplam**
AĞRI	357	462	819	**İNEK**	63	0	63
~AĞRI	621	1470	2091	**~İNEK**	915	1932	2847
Toplam	978	1932	2910	**Toplam**	978	1932	2910

p(baş|ağrı) = 357 / 819 = 0.44 p(ağrı|baş) = 357 / 978 = 0.37 w = 1 – 0.44 = 0.56

p(baş|inek) = 63 / 63 = 1 p(inek|baş) = 63 / 978 = 0.06 w = 1 – 1 = 0

Yüksek *w* değerlerine sahip bağlantıların çıkarılması anlamsal ilişkileri güçlü bağlantıların çizgeye dahil edilmesi açısından önemlidir. Çalışmamızda *w* değerinin 0,8 ~ 0,9 değerlerinden büyük olduğu durumlarda ilgili bağlantılar göz ardı edilmiştir. Zayıf kenarların çıkarılmaması durumunda, özellikle derlem boyutu büyüdükçe tesadüf eseri bir arada gözlenen sözcük çiftlerinin de etkisiyle çizgenin her zaman tam bağlı bir yapıda olması kaçınılmaz olmaktadır.

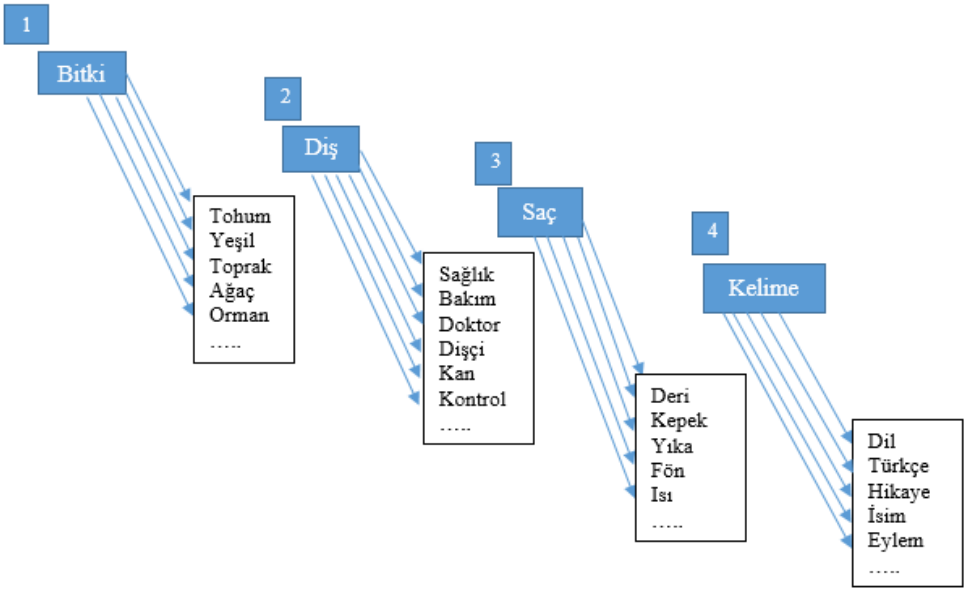
4.3.3 Yüksek yoğunluklu bileşenlerin bulunması

Yüksek yoğunluklu bileşenlerin bulunması iki adımdan oluşmaktadır. İlk adım farklı anlamsal özellikler için merkez düğüm işlevi gören bileşenlerin tespit edilmesidir. İlerleyen adımda ise bulunan merkez düğümlere bağlı olan bileşenler listelenmektedir. Çalışmamızda merkez düğümlerin belirlenmesi amacıyla, yinelemeli bir algoritma uygulanmıştır; her adımda en büyük görece dereceye sahip düğüm noktası merkez düğüm olarak seçilmiştir. Bu seçimde düğüm noktasının derecesi ve sıklığının orijinal metinde yüksek derecede ilişkili olması durumu belirleyici olmuştur. Seçilen merkez düğümün birinci dereceden komşuları sonraki adımda merkez düğüm seçilme şansını kaybetmektedir. Merkez düğümlerin anlatılan sıra ile belirlenmesine ilişkin bir gösterim Şekil 4.2'de verilmektedir. Algoritma bir sonraki merkez düğüm için gözlenen sıklık değeri belirli bir eşiğin altında kaldığında durmaktadır. Bu noktada seçili olan tüm merkez düğümlerin hedef sözcüğün olası anlamlarını temsil ettiği kabul edilmiştir. *Bitki* anlamı *kök* sözcüğünün ilk merkez düğümü olarak seçildikten sonra birinci derece komşuları olan *tohum*, *yeşil*, *toprak*, *ağaç* ve *orman* gibi sözcükler adaylık listesinden silinmiştir. Ardından *diş* sözcüğü ikinci anlam olarak belirlenmiş ve bu düğümün birinci derece komşuları olan *sağlık*, *bakım*, *doktor*, *dişçi* ve diğer sözcükler listeden çıkarılmıştır. Benzer işlemler tüm düğüm noktaları tamamlanana dek tekrarlanmıştır.

4.3.3.1 Merkez düğümlerin belirlenmesi

Birliktelik çizgesi oluşturulduktan sonra, bu çizgeye ilişkin merkez düğümleri bulmak üzere yinelemeli basit bir algoritma kullanılmıştır. Algoritma her adımda

çizge içerisindeki en yüksek göreli sıklığa sahip düğüm noktasını bulmaktadır. Eğer bu düğüm noktası belirli koşulları sağlıyorsa merkez düğüm olarak seçilmektedir. Bu koşullar sezgisel bir takım ölçütler ile belirlenmektedir. Bir düğüm noktası merkez düğüm olarak seçildikten sonra, komşuları bir sonraki merkez düğüm adayı olma şansını kaybetmektedir. Herhangi bir anda, sıradaki düğüm noktasına ilişkin sıklık değeri belirli bir eşik noktasının altında ise ya da hedeflenen merkez düğüm sayısına ulaşılmadıysa algoritma sonlanmaktadır.



Şekil 4.2 : Kök sözcüğü için komşulukların adım adım silinmesi.

Bu çalışma kapsamında merkez düğüm seçimi için HyperLex algoritmasında izlenen yaklaşımdan farklı olarak sözcük sıklıklarına ilişkin eşik değeri yerine önceden belirlenen bir merkez düğüm sayısı kullanılmıştır. HSD'deki örneklerimizin sayısı ve sözcük anlamlarının dağılımları bu seçimde etkili olmuştur. Merkez düğüm seçiminde sözcüklerin sıklıkları ve birbirleri ile ilişkileri göz önünde bulundurulmuş, bu seçimin iyileştirilmesinde etkili olduğu düşünülen ölçütler incelenmiştir. Bu aşamanın sonunda her bileşen grubu için bir merkez düğüm ve düğüme özgü ve en sık gözlenen komşular elde edilmiştir.

Şekil 4.3'de merkez düğüm seçiminde kullanılan algoritma adımları görülmektedir. Algoritma için çizge (Ç), sözcük sıklıklarını tutan bir dizi (Freq) ve merkez düğüm sayısı (n) değerleri kullanılmaktadır. Sözcükler azalan sıklık değerlerine göre V dizisine atanmaktadır. Her adımda en yüksek sıklığa sahip sözcük istenen koşulları sağlaması durumunda merkez düğüm olarak belirlenmekte ve H merkez düğüm

vektörüne atanmaktadır. Atama gerçekleştirildikten sonra V vektöründen hem sözcük hem komşuları çıkarılmaktadır. Bu yaklaşımla sonraki merkez düğüm seçiminde seçilen sözcük ve birinci dereceden komşularının elenmesi sağlanmış olmaktadır. Algoritma, bulunan merkez düğüm sayısı m, ulaşılmak istenen düğüm sayısı n'e eşit olduğunda sonlanmakta ve H merkez düğüm vektörü elde edilmiş olmaktadır. Sonraki adımda tüm bileşenlerin ilgili merkez düğüme dahil edilmesi sağlanmıştır.

```
MerkezDugumler(Ç, n, Freq)  {
    Ç: Birliktelik çizgesi
    n: Merkez düğüm sayısı
    Freq: Düğüm noktası sıklıklarını tutan dizi

     V ← Ç Çizgede azalan sıklık sırasındaki düğüm dizisi
     H ← 0 Merkez düğüm kümesi
     m ← 0 Merkez düğüm sayacı
     while V ≠ 0 and n > m {
           v ← V[0]
           if iyiAday(v)
           then  {
             H ← H U v
             V ← V – (v U Γ(v))
         m ← m + 1
     }
  }
return H
}
```

Şekil 4.3 : Merkez düğümlerin bulunmasında kullanılan algoritma.

4.3.3.2 Bileşenlerin ayrılması

Çizge yapısındaki yüksek yoğunluklu bileşenler ayrılarak kendilerine en yakın merkez düğüme bağlanmışlardır. Merkez düğümler belirlendikten sonra düğümlere özgü bileşenler ve diğer bileşenler merkez düğüme kademeli olarak bağlanmıştır. Bu işlem sırasında düğümler arası uzaklıklar göz önünde bulundurulmuş ve en yakın merkez düğüme bağlantı yapılması sağlanmıştır. Tüm düğüm noktaları ilgili merkez düğüme bağlandıktan sonra, merkez düğümler sıfır uzaklık ile hedef sözcüğe bağlanarak tüm çizge yapısına *En Küçük Kapsayan Ağaç* (EKKA) algoritması uygulanmıştır. EKKA algoritması ile birbirinden ayrıştırılmış olan bileşenlerin iç yapısındaki (hiyerarşik olarak merkez düğümlerin altında yer alan sözcükler) en kısa yolların bulunması sağlanmıştır.

Algoritmaya ilişkin adımlar Şekil 4.4'te verilmektedir. Merkez düğümler hedef sözcüğe 0 uzaklıkla bağlanmakta ve tüm çizge yapısına EKKA algoritması uygulanmaktadır. EKKA ağacının bulunmasında *Kruskal* algoritması (Kruskal, 1956) kullanılmıştır. Algoritmanın karmaşıklığı en kötü durumda EKKA ağacı karmaşıklığına eşittir. En kötü durum karmaşıklığı E kenar sayısı olmak üzere, O(E lg E) olarak verilmektedir.

```
Bilesenler(C, M, h) {
     C: çizge
     M: merkez düğümler kümesi
     h: hedef sözcük

    C′  ← C U h
      for each h in M {
               C′ ye 0 ağırlıkla kenar ekle <h, m>
}
   T ← MST(C′, h)
    return T
}
```

Şekil 4.4 : Bileşenlerin bulunması.

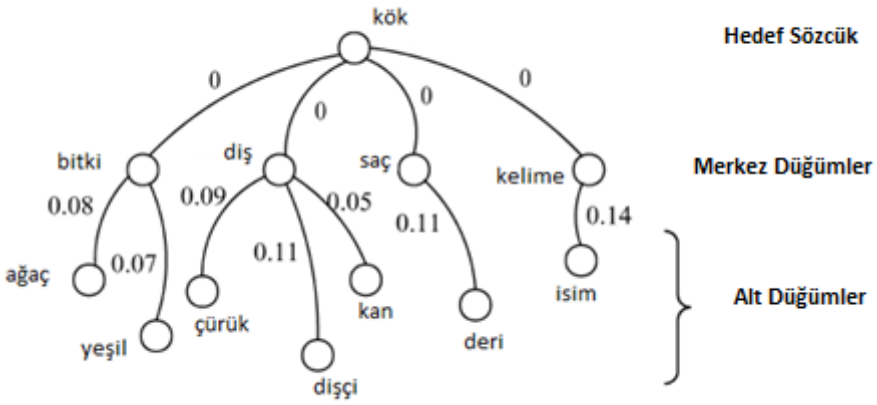
4.3.4 Belirsizlik giderme

EKKA yapısı hedef sözcüğe ilişkin örneklerin anlam belirsizliğinin giderilmesinde kullanılmaktadır. $W = (w_1, w_2, \dots, w_i, \dots, w_n)$ 'nin w_i'nin içinde hedef sözcüğün bir örneği olduğu bağlam olduğunu kabul edelim. Bir *s* derece vektörü (score vector) her $w_j \in W$ $(j \neq i)$ için hesaplanmaktadır. Aşağıda gösterilmekte olan hesaplamada, k numaralı bileşen s_k, ilgili merkez düğüm için elde edilen derece değerini vermektedir.

$$S_k = \begin{cases} \frac{1}{1+d(h_k, w_j)} & \text{eğer } h_k \ w_j' \text{nin MST}'\text{deki ebeveyni ise} \\ 0 & \text{ilk koşul geçerli değil ise,} \end{cases} \qquad (4.9)$$

Denklem 4.9'da, $d(h_k, w_j)$ kök düğüm h_k ile w_j düğümü arasındaki mesafeyi göstermektedir. İlerleyen adımda w_j ile eşleştirilen tüm $w_j \in W$ $(j \neq i)$ derece vektörleri toplanarak, en yüksek değeri alan merkez düğüm w_j en uygun anlam olarak belirlenmektedir.

Şekil 4.5'te "kök" hedef sözcüğü için oluşturulmuş bir EKKA yapısı görülmektedir. Ağaç yapısının birinci düzeyinde sözcüğün anlamlarını temsil eden merkez düğümler yer almaktadır.



Şekil 4.5 : Kök sözcüğüne ilişkin örnek EKKA yapısı.

Elde edilen ağaç yapısı anlam belirsizliği gidermede kullanılmaktadır. Hedef sözcüğün belirli bir metinde gözlenen her örneği için bu sözcüğü çevreleyen komşu sözcükler değerlendirmeye alınmakta ve ağaç içinde aranmaktadır. EKKA yapısı oluşturulurken bir sözcük sadece bir merkez düğüme bağlanmaktadır. Bağlamdaki her sözcük bir *s* derece dizisi değeri almaktadır. Bu değerler bir derece vektörü içinde tutulduğunda sadece altında bulunulan merkez düğüme ilişkin *i* numaralı göz değer almaktadır. Sözcüklere tek bir merkez düğümden ulaşılabildiği için, bu dizide sözcüğün altında yer aldığı merkez düğümü hariç diğer değerler 0 değerini almaktadır. Bağlamdaki tüm sözcükler için ortak bir derece vektörü alınarak, sözcüklerin aldığı değerleri vektördeki ilgili merkez düğüm gözüne yazılmaktadır. $d(h_i, v)$ uzaklığı ise merkez düğümler için 1 olarak alınırken, bu mesafe aranan sözcük ağaç yapısında merkez düğümden uzaklaştıkça düşmektedir. Hedef sözcüğün geçtiği bir örnek için, bağlamdaki sözcüklere ilişkin tüm derece vektörleri toplandığında en yüksek değeri alan merkez düğüm doğru anlam olarak seçilmektedir.

4.4 Denetimsiz SABG Yaklaşımlarında Değerlendirme

SABG kapsamında kullanılan tüm denetimsiz yöntemlerde değerlendirme için bazı ek bilgilere gereksinim duyulmaktadır. Seçeneklerden bir tanesi sözcüklerin atandığı

merkez düğümlerin elle kontrol edilmesidir. Bu yöntemin iki olumsuz yanı vardır; birincisi sözcüğe ilişkin her örneğin elle kontrolü külfetlidir. İkincisi, merkez düğümler ile sözcük anlamlarının eşleştirilmesi özellikle bu eşleşmeyi yapacak kişiler ve kendilerine sağlanan kısıtlı anlam listesi düşünüldüğünde zor bir işlemdir.

İkinci seçenek sistemin örneğin bilgiye erişim sistemi gibi bir uygulama içerisindeki başarımını değerlendirmektir. Bu oldukça ilgi çekici bir fikir olmakla birlikte, sistemin geliştirilmesi ve bazı durumlarda iyi ve kötü başarımın değerlendirilmesinde zorluklar ortaya çıkabilmektedir.

Üçüncü seçenek; elde edilen merkez düğümlerin (kümelerin) sözlük anlamları ile eşleşmelerini sağlayacak bir yöntem geliştirmektir. Etiketli derlemlerin bu eşleşme için kullanılması, sonuçların denetimli SABG ve diğer sistemlere ilişkin sonuçlarla karşılaştırılmasını olanaklı hale getirmektedir. Bununla birlikte sistem yarı denetimli olarak değerlendirilen bir şekle gelmektedir. Bu yaklaşımla eşleşme sağlanırken gürültü ve bilgi kayıpları oluşma olasılığı göz önünde bulundurulmalıdır.

Bir diğer seçenek ise elde edilen anlamların sözlük anlamlarına göre değerlendirilmesi işlemini kümeleme kullanarak gerçekleştirmektir. Burada bulunan anlamlar kümeler, sözlük anlamları ise sınıflardır. Bu doğrultuda düzensizlik ve saflık gibi ölçütler kullanılabilmektedir.

4.4.1 Merkez düğümlerin kümelenmesi ile değerlendirme

Bu değerlendirme seçeneğinde merkez düğümler örnek kümeleri, sözlük anlamları ise anlam sınıfları olarak karşımıza çıkmaktadır. Kümeleri anlam sınıfları ile karşılaştırmak için elle işaretlenmiş derlemler gerekmektedir. Sınama kümesi öncelikle bulunan anlamlar ile işaretlenmektedir. Kusursuz bir kümeleme ise aynı kümenin sadece bir anlam sınıfına ilişkin örnekleri bir araya getirdiği durumdur. Buradaki değerlendirme tamamen denetimsizdir.

Yapılan çalışmalarda bu değerlendirme şekli tercih edildiğinde düzensizlik, saflık ve F-ölçütü olmak üzere üç farklı ölçüt kullanılmaktadır (Agirre ve diğ, 2006; Zhao ve diğ, 2005). Düzensizlik çeşitli nesne sınıflarının kümelere göre dağılımını yorumlamaktadır. Genellikle, düzensizlik değerinin küçük olması kümeleme algoritması başarımının daha iyi olduğunu göstermektedir. Saflık ölçütü, genel

olarak bir sınıfa ilişkin örnekleri toplayan kümelerin derecesini ölçmektedir. Saflık değerinin büyük olması, kümeleme algoritması başarımının daha iyi olduğunu göstermektedir. F-ölçütü ise tutturma ve bulma değerlerinden elde edilmektedir. Tutturma değeri, bir küme için doğru olarak getirilmiş örnekleri, bulma ise bir küme için doğru olarak döndürülmüş örnek yüzdesini vermektedir.

4.4.2 Merkez düğümler ve sözlük anlamlarının eşleştirilmesi ile değerlendirme

Eşleştirme ile değerlendirme çizge yapısından bulunmuş sözcük anlamları ile sözlük anlamlarının eşleştirilmesi için ortaya koyulan bir çözümdür (Agirre ve diğ, 2006). SABG sistemi önce eğitim verisini yeni elde edilen anlam kümeleri (merkez düğüm etiketleri) ile etiketlemektedir. Sözlük anlam etiketleri ise anlam kümeleri ve sınıfları arası eşleştirme için, basitçe s_j anlamının h_i merkez düğümü ile kaç defa eşleştiği bilgisini tutan bir matris oluşturulmak üzere kullanılmaktadır. Sınama aşamasında SABG algoritması sınama kümesine uygulanmakta, merkez düğüm-sözlük anlamı matrisi ise en yüksek ağırlığa sahip anlamın seçilmesinde kullanılmaktadır.

4.5 Çizge Tabanlı Algoritmanın Gerçekleştirilmesi

Çizge tabanlı algoritmamızın gerçekleştirilmesinde ayrıntıları daha önce verilen ara aşamalardan oluşan genel algoritmaya ilişkin adımlar Şekil 4.6'da gösterilmektedir. Çalışmamızda öncelikle her hedef sözcük için yüz paragraftan oluşan örnekler kullanılarak çizge yapısı kurulmuştur.

Adım1: Birliktelik çizgesi G'nin oluşturulması.

Adım2: Merkez düğümlerin belirlenmesi

Adım3: Her düğüm noktasının kendine yakın merkez düğüme eklendikten sonra hedef sözcüğün çizgeye eklenerek EKKA algoritması uygulanması

Adım4: **Sınama:** Verilen bir bağlamdaki tüm sözcüklerin derece vektörüne eklenerek, en yüksek dereceyi alan bileşenin anlam olarak seçilmesi.

Şekil 4.6 : Algoritma genel adımları.

Öncelikle tüm veride toplam sıklığı belirli bir eşik değerinin üzerinde olan sözcükler değerlendirmeye alınmış, daha düşük değerlere sahip sözcükler göz ardı edilmiş ve çizge dışında bırakılmıştır. Bu sıklık değeri için çalışmamızda 5 – 10 aralığı kullanılarak deneyler yapılmış ve etkin değer aralığına ulaşılmıştır. Çalışmamızda sözcük ya da düğüm noktası sıklık değerleri 6-8 aralığında eşiğin daha düşük veya daha yüksek olduğu değerlere göre etkin bulunmuştur. Sonraki adımda göreli olasılık değerleri birlikte gözlenen sözcük çiftlerinden her biri için hesaplanmış ve iki değerin de belirlenen eşik değerinden büyük olduğu kenarlar elenmiştir. Bu adımda sözü edilen değerin büyüklüğü sözcük çiftleri arasındaki bağın zayıflığını göstermektedir. Zayıf bağlara sahip düğüm noktalarının elenmesi merkez düğümlerinin doğru seçilmesi açısından önemlidir. Örneğin merkez düğüm olmaya aday iki düğüm noktası arasındaki zayıf bağ sözü edilen eşik değerinin doğru seçilmesiyle elendiğinde ikinci düğüm merkez düğüm olma şansını kaybetmeyecektir. Aksi durumda sözcükler ilk seçilen merkez düğüme bağlı olacağından ikinci düğüm noktası gerekli kriterleri sağlaması durumunda bile sonraki merkez düğüm olarak bulunamayacaktır.

Merkez düğümler belirlenirken düğümler arasından en yüksek sıklığa sahip, derecesi ya da komşuluk düzeyi belirli sayıda ve ortalama ağırlık olarak belirlenen eşik değerinin üzerinde olan adaylar seçilmiştir. Merkez düğümler veri kümesindeki belli başlı anlamları temsil etmektedir. Örneğin *açık* hedef sözcüğü için elde edilen merkez düğümlerden bazıları; *öğretmen, renk, milyar* ve *baş* gibi sözcüklerden oluşmaktadır. Merkez düğümler hedef sözcüğün veri kümesindeki örneklerde bulunan *'öğretmen açığı', 'açık renk', 'cari açık'* ve *'başı açık'* gibi kullanımlarına karşılık gelmektedir. Bir diğer örnek olan *kök* hedef sözcüğü için *'bitki', 'diş', 'saç', 'kelime'* vb. sözcükler merkez düğümler olarak bulunmaktadır.

İlerleyen aşamada diğer düğümler kendine en yakın merkez düğüme bağlanmıştır. Bu aşamada anlamları temsil eden kümeler birbirinden ayrı ancak kümelerin içindeki çizge bağlantıları aktiftir. Çizge yapısını üzerinde arama yapabileceğimiz bir ağaç yapısına dönüştürmek için her merkez düğüm 0 uzaklıkla hedef sözcüğe bağlanarak EKKA algoritması uygulanmıştır. Bu şekilde elde edilen yapı, her sözcüğün belirli merkez düğüm altındaki dallardan birinde yer aldığı bir ağaç yapısına dönüşmüştür..

Sınama aşamasında ise cümle ya da paragraf gibi bir sınama örneğindeki her sözcük merkez düğümler altında aranmakta, bulunduğu yere olan uzaklık ilgili anlam için

ayrılan vektör gözüne eklenmektedir. Tüm paragraf için en yüksek değeri alan merkez düğüm, o bağlamdaki doğru anlamı vermektedir.

Çizge algoritmamızın değerlendirme aşaması için farklı yaklaşımlar denenmiştir. İlk yapılan çalışmanın değerlendirme aşamasında TDK sözlüğündeki anlam tanımları ile merkez düğümler ve bu düğümlere bağlı üyeler arasındaki örtüşmeler dikkate alınmıştır. Sözlük kapsamında sadece sözcük anlam tanımlamaları değil, örnekler de kullanılmıştır. Sözlük anlamı ve merkez düğüm düğüm örtüşmelerinin bulunması için aşağıda verilen yaklaşımlar sınanmıştır:

- Sınama verisindeki sözcük en küçük kapsayan ağaç üzerinde aranırken gezdiği düğümlerin sözlük anlamları ile eşleşmesine bakılması.
- Bir merkez düğüm altındaki tüm sözcüklerin sözlük anlamları ile örtüşmesine bakılması.
- Merkez düğümün kendisinin sözlük anlamları arasında aranması.

Yapılan çalışmalarda merkez düğümün kendisi ya da merkez düğümü temsil eden az sayıdaki sözcüğün doğru anlamla örtüşmede daha iyi sonuç verdiği gözlenmiştir. Bununla birlikte sözlükte yer alan anlam tanımlamalarının ve örneklerin her sözcük için yeterli düzeyde olmadığı görülmüştür. Örneğin bir sözcüğe ilişkin anlamların bir bölümü için sözlükte örnek mevcutken, bir çok anlam için örnek cümle sağlanmamıştır. Sözlük kullanımı ile anlam eşleştirmesinde doğru anlamın yakalanmasında karşılaşılan temel zorluk sözü edilen anlam örneklerinin azlığı olmuştur.

4.5.1 Parametrelerin ayarlanması

HyperLex algoritma davranışının sezgisel bir parametre kümesinden etkilendiği düşünülmektedir. Çalışmamızda bu parametre kümesine ilişkin en iyi değerlerin bulunmasına çalışılmıştır. Çizge tabanlı algoritmanın parametreleri iki gruba ayrılmaktadır: ilk grup birliktelik çizgesinin oluşumunu etkileyenler (p1-p4 arası), ikinci grup ise bu çizgeden merkez düğümlerin elde edilmesini etkileyen parametrelerdir (p5-p8 arası). Bu parametrelerden bir bölümü örneğin, p1 ve p2 çizge oluşumundan önce dikkate alınan parametrelerdir. Sözcüklere ve sözcük birliktelik sıklıklarına ilişkin değeri vermektedir. P3 parametresi, çizge yapısına ilişkin ağırlık değerleri hesaplandıktan sonra ihmal edilecek olan kenarların, diğer deyişle aralarındaki mesafenin uzak ve ağırlığın büyük olduğu sözcükleri birleştiren bağların

alt sınırını vermektedir. P4 parametresi, örneklerin uzunluğu ile ilgili olan parametredir. Çalışmamızda kullanılan örnekler belirli bir standartta ve yakın uzunlukta olduklarından bu parametre ihmal edilmiştir.

Çizelge 4.2 : Çizgeye ilişkin parametreler.

Pr	**Açıklama**
p1	En küçük kenar sıklığı (sözcük birliktelikleri)
p2	En küçük düğüm noktası sıklığı (sözcükler)
p3	Ağırlığı eşik değerinin üzerinde olan kenarların elenmesi
p4	Belli değerden daha az sözcük içeren bağlamın işlem görmemesi
p5	Merkez düğüm için en az komşu düğüm sayısı
p6	Merkez düğüm komşuları için en büyük ortalama ağırlık
p7	Merkez düğümlere ilişkin en küçük gözlenme sıklığı
p8	Seçilen merkez düğüm sayısı.

Çalışmamızda derlemdeki her düğüm noktasına karşılık gelen sözcüğün en küçük sıklık değeri için 5 – 10 arasındaki değerler sınanmış ve 6 – 8 değer aralığının daha iyi sonuçlar verdiği görülmüştür. Aynı zamanda iki düğüm noktası arasındaki ağırlığın 0,60 – 0,70 değerleri arasında olduğunda daha iyi sonuç verdiği gözlenmiştir. Kenarlara ilişkin ağırlık değeri büyük tutulduğunda zayıf bağlara sahip bilgi değeri olmayan sözcüklerde çizge yapısına katılabilmekte ve çizgedeki düğümler arası komşuluğu olumsuz yönde etkilemektedir. Aynı zamanda düğümler arasındaki zayıf bağlar merkez düğüm seçimini de etkileyebilmektedir. Çalışmamızda merkez düğüm sayısı beş anlamı içine alacak şekilde belirlenmiş, derlemin boyutu da göz önünde bulundurularak bu değerin belirlenmesinde bir sonlandırma koşulu kullanılmamıştır.

4.5.2 HyperLex algoritması ile elde edilen sonuçlar ve değerlendirme

Çalışmamızda Türkçe HSD isim kümesi kullanılarak çizge yapısı oluşturulmuştur. Anlamların işaretli olduğu derlemde 10 katlı çapraz doğrulama kullanılmış, örnek kümesinin sınama için ayrılan bölümü çizge oluşturmada kullanılmamıştır. Sözcük birliktelikleri saptanarak ağırlıklar belirlenmiş, merkez düğümler ortaya çıkarılmış ve hedef sözcük çizgeye eklenerek EKKA algoritması uygulanmıştır.

Elde edilen merkez düğümlerin sözlük anlamı karşılığını bulmak için merkez düğüm/sözlük anlamı matrisinden yararlanılarak anlamlar arası dönüşüm

sağlanmıştır. Önceki bölümde sözü edilen merkez düğüm sözlük anlamı eşleşmeleri bulmada tercih edilebilecek farklı seçenekler denenmiş, matris kullanımı dışındaki yöntemlerin kullanışlı olmadığı görülmüştür. Örneğin, sözlük anlamları ve ağaç yapısında merkez düğümlerin altında bulunan düğüm noktalarının kesişimlerine bakılmış ancak örtüşmelerin yetersiz olduğu gözlenmiştir. Bu durum, sözlükteki anlam tanımlarının çoğu durumda kısa ve yetersiz olmasından, tüm tanımlar için sözlükte örnek bulunmamasından kaynaklanmaktadır. Şekil 4.7'de TDK sözlüğünde "kök" sözcüğü için tanımlı anlamlar ve karşılık gelen örnekler görülmektedir. Sözlükte kök sözcüğüne ilişkin dört anlam için örnek verilmiştir. Anlam belirginleştirilmesi adayı diğer sözcükler için de benzer durum mevcut olup, sözlük anlamı – merkez düğüm eşleşmesine bakıldığında karşılaşılan temel zorluk olduğu görülmüştür.

	Sözcük_ID:31717 (Anlamlar)
1	Bitkileri toprağa bağlayan ve onların, topraktaki besi maddelerini emmesine yarayan klorofilsiz bölüm
2	Süsende olduğu gibi yer üstüne sap çıkaran çok yıllık yer altı gövdesi
3	Bazı şeylerde dip bölüm
4	Sapıyla çıkarılan bitkilerde tane
5	Dip, temel, esas
6	Kaynak, köken
7	Bir kimseyi bir yere bağlayan manevi temel güçlerin bütünü
8	Sözcüğün her türlü ek çıkarıldıktan sonra kalan anlamlı bölümü: Yaptırmak kelimesinde kök, yap- bölümüdür
9	Olağan şartlarda çevresinden yalıtılamayan ancak birçok tepkimede nitelik değiştirmeden geçebilen atom kümesi
10	Denklemde bilinmeyenin yerine konulduğunda uygun düşen gerçek veya birleşik değer

	Sözcük_ID:31717 (Örnekler)
3	Diş kökü.
4	Üç kök maydanoz.
5	Ta gölden başlayan tipi ve fırtına Şebnem'in sıcak evini kökünden sarsıyordu.
6	Ölenle, son zamanları gevşeyen, azalan fakat kökleri mazinin sağlamlığı içinde kalan eski bir aşinalığım vardı.

Şekil 4.7 : TDK sözlüğü anlam – örnek eşleşmeleri.

Çalışma kapsamında önceki bölümde verilen farklı parametre değerleri sınanmıştır. Bu parametre değerlerinde öncelikli olarak merkez düğümlerin doğru oluşmasını etkileyen parametreler üzerinde durulmuştur. Örneğin iki düğüm noktasını bağlayan bir kenar değeri 1'e yakınsa bu iki sözcüğün ilişkisinin zayıf olduğunu göstermektedir. Elenecek kenar ağırlıkları için büyük değerler seçmek bağlantısı zayıf iki düğüm noktasının bağlı kalmasını sağlayacak ve önceki bölümde değinilen nedenden ötürü merkez düğümlerin doğru olarak elde edilmesini engelleyecektir. Bu doğrultuda merkez düğümler etrafındaki düğüm noktalarının iyi bir şekilde ayrıştırılması önemlidir. Yapılan deneyler doğrultusunda seçilen parametre değerleri Çizelge 4.3'te verilmektedir.

Çizelge 4.3 : Çizge tabanlı yöntem parametre değerleri.

	Sınanan Aralık	**Tercih Edilen Değer**
P1	-	-
P2	5 - 10	6 - 7
P3	0,6 – 0,9	0,7
P4	-	-
P5	5 - 10	5
P6	0,7	0,7
P7	-	-
P8	5	5

Çizge tabanlı benzer yöntemlerle yapılan çalışmalarda isim kümeleri değerlendirmeye alınmıştır. Eylemler üzerinde elde edilen sonuçlara ilişkin başarım düzeyinin yeterli olmadığı kaydedilmiştir. Çalışmamızda ise eylem grubu anlam sayılarının fazlalığı ve bir anlama karşılık gelen örnek sayısının isim kümesine göre azlığı nedeniyle kapsam dışında bırakılarak algoritmamız Türkçe isim grubu üzerinde sınanmıştır. Bununla birlikte derlemimizde isim grubu içinde örnek sayılarının sınırlı olduğu anlamlar elenmiştir. Aynı zamanda Türkçe isim grubu üzerinde yapılan deneylerde derlemimizdeki aday sözcüklere ilişkin anlam dağılımlarının değişken olması nedeniyle, dağılımın dengeli olduğu sözcükler değerlendirmeye alınmıştır. Çalışma kapsamında ele alınan sözcükler için bulunan merkez düğüm anlamları Çizelge 4.4'de görülmektedir. İkinci sütunda görülen anlamların her biri sözcüğün bir kullanımına karşılık gelmektedir. Merkez düğüm

seçimi için kullandığımız parametrelerin sözcük anlamlarını iyi temsil eden düğüm noktalarını ortaya koyduğu gözlenmiştir.

Çizelge 4.4 : Çizge tabanlı yöntem sözcük anlamları.

Sözcük	**Merkez Düğümler:**
Açık	Renk, öğretmen, bütçe, baş, deniz
Baskı	Kitap, insan, oyun, gazete, gün
Baş	Ağrı, devlet, yıl, açık, sığır
Derece	Okul, sıcak, açı, başarı, gör
Dünya	İnsan, büyük, çocuk, gör, ülke
Göz	Gör, geç, dolap, delik, şans
Kök	Aile, diş, bitki, ek, saç
Kör	Göz, ışık, siyaset, bıçak, başbakan
Ocak	Yemek, yıl, maden, iç, ülkü
Yaş	Doğum, ıslak, zor, ağla, zaman

Deneylerde 10 katlı çapraz doğrulama kullanılarak veri kümesinin farklı bölümleri üzerinden çizge oluşturulmuş ve sınama kümesi üzerinden elde edilen doğruluk değerlerinin ortalaması elde edilmiştir. Çizelge 4.5 sözcükler için elde edilen doğruluk değerlerini göstermektedir.

Çizelge 4.5 : Çizge tabanlı yöntem başarım oranları – I (%).

Sözcük	**Baskın Anlam**	**Doğruluk**
Açık	27	69
Baskı	30	62
Baş	30	66
Derece	38	64
Dünya	18	60
Göz	51	79
Kök	28	71
Kör	53	68
Ocak	26	65
Yaş	52	73

Çizelge 4.6'da ise sözcüklerin ortalama doğruluk değeri, saflık ve düzensizlik ölçütleri ile birlikte verilmektedir. Ek B'de detayları verilmekte olan saflık ve düzensizlik ölçütleri de çalışmamızda elde edilen kümelerin kalitesinin ölçülmesinde faydalanılan ölçütler olmuştur. Bir kümeye ilişkin saflık derecesi verilen kümede elde edilen sınıfların tek sınıfa ait olma derecesini ölçmektedir. Merkez düğüm etiketi olarak elde ettiğimiz kümelerin saflığını bulmada kullanılmaktadır. Düzensizlik ölçütü ise merkez düğüm etiketi ile etiketlediğimiz örneklerin dağılımına göre hesaplanmaktadır. Bir kümedeki farklı sınıfa ilişkin örnekler eşit dağılımda bulunduğunda düzensizlik bir olarak bulunmakta, saflık derecesinin yüksek olduğu ve tüm örneklerin aynı sınıftan olduğu durumda düzensizlik sıfır olmaktadır.

Çizelge 4.6 : Çizge tabanlı yöntem başarım oranları - II.

	HSD İsim Kümesi
Düzensizlik	0,37
Saflık	0,71
Doğruluk	0,68

4.6 Bölüm Sonucu

Bu çalışma kapsamında Türkçenin yapısı ve mevcut kaynakları göz önünde bulundurularak etkin bir denetimsiz yöntem geliştirilmesi hedeflenmiştir. Çizelge 2.1'de özetlenen genel yöntem sınıflarına ilişkin üstünlük ve eksik yönler bize günümüze dek yapılan çalışmalara ilişkin bulguları yöntem sınıfları temelinde yansıtmaktadır. Bununla birlikte yöntem temelinde değerlendirme yapıldığında denetimsiz yöntem sınıfına dahil olan çizge tabanlı yöntemlerin diğer denetimsiz yöntemlere kıyasla oldukça tatmin edici ve denetimli yöntem sonuçlarıyla karşılaştırılabilir sonuçlar ürettiği gözlenmektedir.

Çalışmamızın önceki bölümlerinde de kullandığımız HSD'nin denetimsiz çizge tabanlı bir algoritma ile kullanıma elverişli olması, belirsizlik taşıyan sözcüklere ilişkin anlamların dinamik olarak derlemden çıkarılabilir olması, dolayısıyla farklı alanlara uyarlanabilmesi yöntemi tercih edilir yapan etmenler olmuştur. Yöntemin temelinde aynı bağlamda gözlenen sözcük bilgilerine dayanarak bir ağ yapısı oluşturulması ve sözcüklerin ne derece sıkı ilişki içinde bulunduğu ile ilgili ölçütlerin çizge yapısının daha iyi temsilinde kullanılması yatmaktadır. Diğer yöntemlerde

kullanılan düzlemsel sözcük vektör temsillerine göre hiyerarşik bir temsil oluşturan çizge yapısının sözcükler arasındaki ilişkiyi daha iyi gösterdiği düşünülmektedir.

Temel adımlarında HyperLex yönteminin esas alındığı yöntemimizde öncelikle komşuluklar belirlenerek ağırlık hesaplamaları yapılmış ve en iyi temsil gücüne sahip merkez düğümlerin elde edilmesinde parametreler üzerinde deneyler yapılmıştır. Çalışmamız ve deneyler sonucunda elde ettiğimiz bilgiler, çizge tabanlı algoritmamızda tanımlanmış olan sekiz parametrenin ilk dördünün çizge oluşumu ve performans açısından oldukça etkili olduğunu göstermiştir. Merkez düğümlerin doğru elde edilmesi başarımda özellikle etkili olduğu için p5-p8 arası parametre değerlerinin de etkin aralığı sınanmıştır. Parametreler derlemimizin özellikleri ve boyutu göz önünde bulundurularak değerlendirilmiştir.

Yapılan deneylerde derlem boyutu göz önünde bulundurularak sabit merkez düğüm sayısı kullanılmıştır. Ortalama doğruluk ve saflık derecesinin sırasıyla %68 ve %71 düzeyinde, ortalama düzensizliğin ise %37 olduğu görülmüştür. Değerlendirme aşamasında sözlük anlam eşleştirme matrisinden faydalanılması yöntemi yarı denetimli olarak değerlendirilebilecek bir denetim seviyesine getirmekte ancak anlamları temsil eden merkez düğümler tamamen denetimsiz olarak elde edilmektedir.

5. DEĞERLENDİRMELER VE SONUÇ

Bu tez çalışmasında DDİ alanının en önemli alt dallarından biri olan SABG konusu ele alınmıştır. Sözcük anlam belirsizliğinin giderilmesi konusunda günümüze dek pek çok çalışma yapılmıştır. Ancak Türkçe için yapılmış olanların sayısı azdır. Çalışmaların hali hazırda devam etmesi SABG alanında kullanılacak uygun kaynakların seçilmesi, anlamların çıkarılması veya işaretlenmesi ve değerlendirme yöntemleri gibi pek çok noktada yapılan incelemelerin devam etmesinin bir sonucudur. Değerlendirme aşamasında yapılan çalışmalar Senseval ve benzer çalıştaylar kapsamında sınanmaktadır. Bununla birlikte SABG, esasen bir ara adım olduğu için gereksinim duyulduğu uygulamalarda da ne derece etkin olduğunun belirlenmesi oldukça önemli bir noktadır.

SABG alanında İngilizce ve sınırlı sayıdaki diğer diller için geliştirilen algoritma ve izlenen yaklaşımlar değerlendirildiğinde önemli ilerleme kaydedilmiş olduğu görülmektedir. Konu kapsamında geliştirilen yöntemler tüm diller için yol gösterici olsa da diller kendi aralarında önemli farklılıklar göstermektedir. Türkçede sözcüklerin aldıkları ekler göz önünde bulundurulduğunda oldukça zengin bir dil olduğu gözlenmektedir. Diller arasındaki yapısal farklılıklar benzer özellikteki dillere uygun yöntemler geliştirilmesini ve yöntemler kapsamında dile özgü özelliklerin kullanılmasını zorunlu kılmaktadır. Bunun yanında diller arası çeviri ya da özet çıkarma gibi kendi aralarında çok farklı özellik gösteren uygulamalarda SABG'ye gereksinim duyulması konuyu zorlaştıran unsurlardan bir tanesidir. Bu durum farklı uygulamaların farklı duyarlılık ve kapsamda anlam belirsizliği gidermeye gereksinim duymasından ve tüm uygulamalar için kullanılabilir ortak bir yöntem geliştirmedeki zorluktan kaynaklanmaktadır.

Bu çalışma kapsamında öncelikle dildeki kaynak kısıtları göz önünde bulundurularak Türkçe için anlam işaretli bir derlem hazırlanmıştır. Derlemimizin kullanıma açılarak Türkçe için yapılacak ileriki çalışmalarda faydalı olacağı düşünülmektedir. Bu özel derlem anlam belirsizliği yüksek 15 Türkçe isim ve eylemden oluşmaktadır. Derlem hazırlama aşamasının ardından ilk olarak denetimli yöntemler kapsamındaki

algoritmalar derlemimiz üzerinde sınanmıştır. HSD derlemi aynı zamanda denetimsiz yöntemlerle yapılan çalışmalarda çizge yapısının oluşturulmasında da kullanılmıştır.

5.1 Yöntemlerin Karşılaştırılması

Denetimli yöntemlerle yapılan çalışmalarda tüm özellik grupları için aynı algoritma kümesi sınanmıştır. Bu algoritmalar NB, IBk, FT, J48 ve DVM yöntemlerinden oluşmaktadır. Tablo 5.1 algoritmalara ilişkin sonuçların farklı özellik ve deneyler için ortalama doğruluk değerlerini en baskın anlamın yüzdesi ile birlikte göstermektedir.

Çizelge 5.1 : Algoritmalara ilişkin ortalama başarım (%).

	EBA	**NB**	**Ibk**	**J48**	**FT**	**DVM**
İsim	33,47	65,61	58,18	63,51	71,49	68,70
Eylem	23,60	52,11	48,41	60,58	63,06	58,86

Çizelge 5.1'deki sonuçlardan ağaç yapılı algoritmaların FT başta olmak üzere isim ve eylem grupları için daha iyi sonuçlar verdiği gözlenmiştir. Yöntemler ayrı ayrı değerlendirildiğinde de FT, J48 ve DVM sonuçlarına ilişkin başarımın NB ve IBk yöntemine oranla daha yüksek olduğu gözlenmiştir.

Değerlendirme ölçütlerinde hata matrisleri ile doğru sınıflandırılan örnek sayısı yanında tutturma ve bulma değerleri de kullanılmıştır. Deneylerde 10 katlı çapraz doğrulama kullanılmıştır. Deneyler kapsamında sınanan aday sözcüklerin anlam dağılımları değişebilmektedir. Örnekler içerisinde dağılımı az olan anlamların tutturma ve bulma değerleri daha düşük olmaktadır. Aynı zamanda anlam dağılımları dengesiz olduğunda sınır başarım değeri yüksek ancak başarımdaki artış daha düşük olmaktadır. Dağılımların dengeli olmaması ve anlamlardan birinin baskın olmasının diğer anlamların örnek sayısını azaltması açısından da olumsuz etkisi olduğu düşünülmektedir.

5.2 Özelliklerin Karşılaştırılması

Denetimli yöntemlerle yapılan çalışmalarda farklı özellik grupları incelenmiştir. Çalışma sonuçlarından konumsal özelliklerin sözcük kesesi özelliklerine kıyasla

sözcük anlamlarını daha iyi ayrıştırdığı görülmüştür. Yapılan dört farklı deney sınıfı kapsamında konumsal özellikler, sözcük kesesi özellikleri, konumsal ve sözcük kesesi özellikleri derlemimiz üzerinde sınanmıştır. Son grupta ise iki özelliğin birlikte kullanılmasının yanında CFS ile özellik azaltımına gidilmiştir. Elde ettiğimiz sonuçlar, konumsal özelliklerin sözcük kesesi özelliklerine göre daha başarılı sonuçlar ürettiğini ortaya koymuştur. İki özellik birlikte kullanıldığında ise başarım artmıştır. Çizelge 5.2'de denetimli yöntemler ile edilen en yüksek doğruluktaki sonuçlar (SK + KÖ özellikleri ve özellik azaltımı) en baskın anlam alt sınırı (EBA) gözetilerek hesaplanan performans artışı ile birlikte verilmektedir.

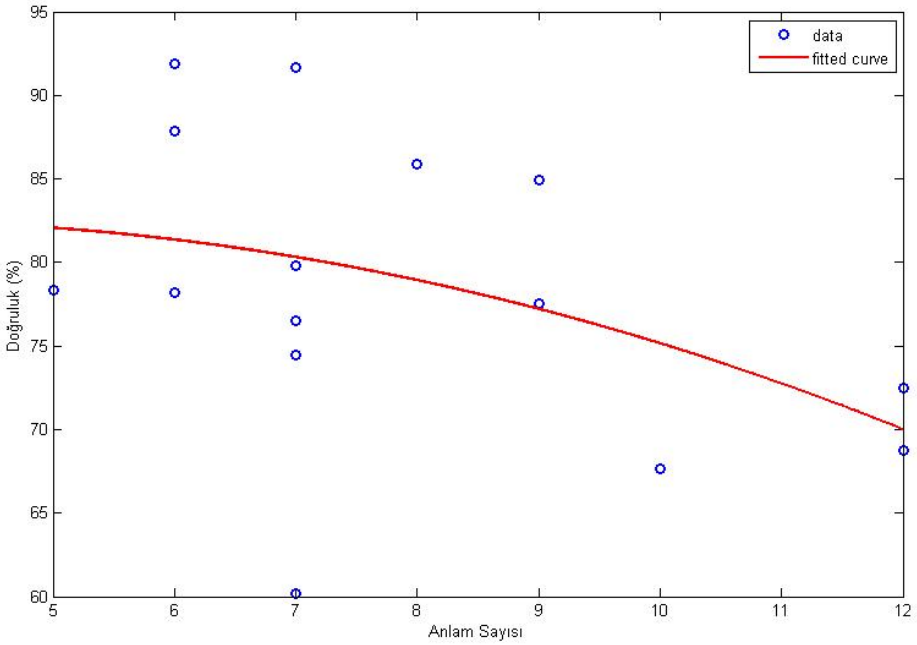
Çizelge 5.2 : Denetimli yöntemlerde en yüksek başarım (%).

	EBA	**NB**	**Ibk**	**J48**	**FT**	**DVM**
İsim	33,47	76,71 **(↑43,24)**	77,89 **(↑44,42)**	69,22 **(↑35,75)**	78,42 **(↑44,95)**	78,91 **(↑45,44)**
Eylem	23,60	64,46 **(↑40,86)**	69,71 **(↑46,11)**	70,29 **(↑46,69)**	72,70 **(↑49,10)**	74,03 **(↑50,43)**

Çizelge 5.2'deki deney sonuçları farklı yöntemlere ilişkin artış değerlerinin bir birine yakın ve kabul edilebilir düzeyde olduğunu göstermektedir. Bununla birlikte, Türkçe isim ve eylem gruplarında en fazla artışın DVM yönteminde olduğu gözlenmektedir.

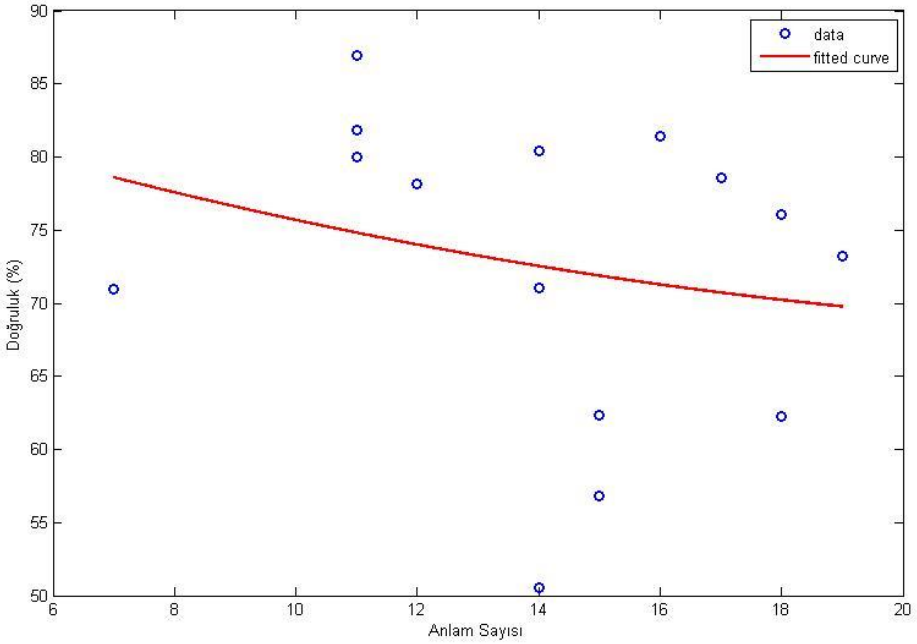
5.3 Çok Anlamlılık

Türkçe anlamsal belirsizlik derecesi yüksek bir dildir. Çalışmamızın başında belirttiğimiz gibi, Türkçe isim ve eylemler için ortalama çok anlamlılık derecesi sırasıyla 10,67 ve 26,53 olarak hesaplanmıştır. Anlam sayıları da sonuçlar değerlendirilirken göz önünde bulundurulması gereken önemli parametrelerdendir. Şekil 5.1 ve Şekil 5.2'de FT algoritması için HSD anlam sayıları ve anlam sayılarına karşılık elde edilen doğruluk değerine göre çizilen grafikler, verinin uydurulduğu ikinci dereceden denkleme ilişkin katsayılar ile birlikte görülmektedir.



Şekil 5.1 : İsim grubu için doğruluk – anlam sayısı ilişkisi.

f(x) = p1*x^2 + p2*x + p3
Katsayılar (%95 güven aralığı):
p1 = -0.1691 (-1.354, 1.016)
p2 = 1.153 (-19.63, 21.94)
p3 = 80.53 (-5.762, 166.8)



Şekil 5.2 : Eylem grubu için doğruluk – anlam sayısı ilişkisi.

f(x) = p1*x^2 + p2*x + p3
Katsayılar (%95 güven aralığı):
p1 = 0.02629 (-0.4842, 0.5368)
p2 = -1.419 (-15.3, 12.47)
p3 = 87.24 (-4.334, 178.8)

5.4 Diğer Çalışmalar ve Karşılaştırmalar

Denetimli yöntemler kapsamında yapılan çalışmalar sırasında ayrıca hedef sözcüğü çevreleyen etkin pencere boyu SK özelliklerinde isim ve eylem grupları için araştırılmıştır. İsim ve eylem grupları için uygun pencere boyu ±5 olarak belirlenmiştir. Aynı zamanda SK özellikleri elde edilirken daha etkin olan özellik sayısı saptanmıştır. Bu değer isim grubunda en fazla bilgi taşıyan ilk 100 sözcük olarak bulunurken, eylem grubunda etkin SK özelliği sayısı 75 olarak bulunmuştur. Denetimli yöntemler kapsamında biçimbilimsel özellik gruplarının anlam belirsizliği giderme üzerindeki etkisi incelenmiştir. Bu gruplar kişi ekleri, sahiplik ve durum eklerini içine almaktadır.

Elde ettiğimiz sonuçlar Türkçe için yapılan önceki çalışmalarla karşılaştırılmıştır. Yapılan karşılaştırmanın güvenilirliği için sözü edilen çalışmayla aynı derlem kullanılmış ve özelliklerimiz ODTÜ-Sabancı ağaç yapılı derleme uyarlanarak sınanmıştır. Çalışmamızdaki tutturma değeri önceki çalışmaya göre isim ve eylem grupları için sırasıyla %0,46 ve %0,54 oranında, bulma değerleri ise %0,07 ve %0,28 oranında daha başarılı bulunmuştur. Bu artışta kullanmış olduğumuz özelliklerin daha etkin ve pencere boyunun daha uygun olmasının etkili olduğu düşünülmektedir.

Denetimli yöntemler ve farklı özellik grupları üzerinde yapılan çalışmalardan sonraki aşamada ise çizge tabanlı ve denetimsiz bir yöntem geliştirilmiştir. Denetimsiz yöntem kapsamında geliştirilen algoritmada çizge yapısının oluşturulmasında hazırlamış olduğumuz Türkçe derlem kullanılmıştır. Çizge yapısının ortaya çıkarılmasında kullandığımız HSD etiketli bir derlem olmakla birlikte yöntem çizge yapısını tamamen denetimsiz bir şekilde ortaya çıkarmaktadır. Çizge yapısı her belirsiz sözcük için oluşturulmaktadır. Belirsiz sözcüğe ilişkin paragraflardan düğüm noktaları ve kenarlara ilişkin ağırlık değerleri elde edilmiş, parametre kümesi üzerinde yaptığımız deneyler sonucu merkez düğümlerin çıkarılmasını da etkileyen bu parametre listesinin etkin değer aralıkları araştırılmıştır. Sonuçların değerlendirilmesi için farklı yaklaşımlar üzerinde çalışılmıştır. Kullandığımız derlem

anlam işaretli bir derlem olduğu için etiketleme ve sonuçların değerlendirilmesinde merkez düğüm ve TDK sözlüğü anlam eşleşme matrisinden faydalanılmıştır. Bu yönüyle yaklaşım yarı denetimli bir özellik kazanmıştır. Bununla birlikte çalışmamızdan elde ettiğimiz değerler, denetimli yöntem sonuçları ile karşılaştırılabilir düzeyde bulunmuş ve başarılı sonuçlar ürettiğini göstermiştir.

KAYNAKLAR

Adalı, E. (2012). Doğal Dil İşleme (Natural Language Processing). *Türkiye Bilişim Vakfı Bilgisayar Bilimleri ve Mühendisliği Dergisi*, 2012. 6(6).

Agirre, E. (1999). *Formalization of concept-relatedness using ontologies: Conceptual Density*. (Ph.D. thesis). University of the Basque Country.

Agirre, E., de Lacalle, O. L., ve Soroa, A. (2014). Random walks for knowledge-based word sense disambiguation. *Computational Linguistics*, 40(1), 57-84.

Agirre, E. ve Martinez, D. (2001). Knowledge sources for word sense disambiguation. *Text, Speech and Dialogue*. Springer Berlin Heidelberg.

Agirre, E., Martinez, D., Lacalle, O.L. ve Soroa, A. (2006). Two graph-based algorithms for state-of-the-art WSD. *in Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Agirre, E. ve Rigau, G. (1996). Word sense disambiguation using conceptual density. *In Proceedings of the 16th conference on Computational linguistics-Volume 1* (pp. 16-22). Association for Computational Linguistics.

Agirre, E. ve Soroa, A. (2009). Personalizing pagerank for word sense disambiguation. *In Proc. of EACL*, pages 33–41.

Atsushi, F., Kentaro, I., Takenobu, T. ve Hozumi, T. (1996). To what extent does case contribute to verb sense disambiguation? *in Proceedings of the 16th conference on Computational linguistics-Volume 1*. Association for Computational Linguistics.

Banerjee, S. ve Pedersen, T. (2002). An adapted Lesk algorithm for word sense disambiguation using WordNet, *In Proceedings of the Third International Conference on Intelligent Text Processing and Computational Linguistics*, Mexico City, February.

Baskaya, O., Sert, E., Cirik, V. ve Yüret, D. (2013). Ai-ku: Using substitute vectors and co-occurrence modeling for word sense induction and disambiguation. *Proceedings of SemEval (2013)*: 300-306.

Bataa, B. ve Altangerel, K. (2012). Word sense disambiguation in Mongolian language. *In Strategic Technology (IFOST), 7th International Forum on. 2012*. IEEE.

Bordag, S. (2006). Word sense induction: Triplet-based clustering and automatic evaluation. in *Proceedings of the 11th EACL*. 2006.

Boser, B.E., Guyon, I.M. ve Vapnik, V.N. (1992). A training algorithm for optimal margin classifiers. *In Proceedings of the 5th Annual Workshop on Computational Learning Theory (Pittsburgh, PA)*. 144–152.

Brin, S. ve Page, L. (1998). The anatomy of a large-scale hypertextual Web search engine. *Computer networks and ISDN systems*, 1998. **30**(1): p. 107-117.

Brody, S., Navigli, R. ve Lapata, M. (2006). Ensemble methods for unsupervised WSD. *In Proceedings of the 44th Annual Meeting of the Association for Computational Linguistics joint with the 21st International Conference on Computational Linguistics* (COLING-ACL, Sydney, Australia). 97–104.

Bruce, R. ve Wiebe, J. (1994). Word-sense disambiguation using decomposable models. *In Proceedings of the 32rd Annual Meeting of the Associtaion for Computational Linguistics, 139-146.*

Bruce, R., Wilks, Y., Guthrie, L., Slator, B. ve Dunning, T. (1992). NounSense - A Disambiguated Noun Taxonomy with a Sense of Humour. *Research Report MCCS-92-246*. Computing Research Laboratory, New Mexico State University.

Budanitsky, A. ve Hirst, G. (2001). Semantic distance in WordNet: An experimental, application-oriented evaluation of five measures. *In Workshop on WordNet and Other Lexical Resources* (Vol. 2, pp. 2-2).

Chen, X., Liu, Z. ve Sun, M. (2014). A unified model for word sense representation and disambiguation. *In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 1025-1035).

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement* Vol.20, No.1, pp.37-46

Collins, M. (2004). Parameter estimation for statistical parsing models: Theory and practice of distributionfree methods. *In New Developments in Parsing Technology*, H. Bunt, J. Carroll, and G. Satta, Eds. Kluwer, Dordrecht, The Netherlands, 19–55.

Cowie, J., Guthrie, J. ve Guthrie, L. (1992). Lexical disambiguation using simulated annealing. *In Proceedings of the 14th conference on Computational linguistics-Volume 1* (pp. 359-365). Association for Computational Linguistics.

Daelemans, W., Van Den Bosch, A. ve Zavrel, J. (1999). Forgetting exceptions is harmful in language learning. Mach. Learn. 34, 1, 11–41.

Decadt, B., Hoste, V., Daelemans, W. ve Van Den Bosch, A. (2004). GAMBL, genetic algorithm optimization of memory-based WSD. *In Proceedings of the 3rd International Workshop on the Evaluation of Systems for the Semantic Analysis of Tex*t (Senseval-3, Barcelona, Spain). 108–112.

Domingos, P. ve Pazzani, M. (1997). On the optimality of the simple Bayesian classifier under zero-one loss. *Machine learning* 29.2-3 (1997): 103-130.

Dongen, S. M. van. (2000). *Graph clustering by flow simulation.* (Doktora Tezi), http://dspace.library.uu.nl/handle/1874/848.

Erkan, G. ve Radev, D.R. (2004). LexRank: Graph-based lexical centrality as salience in text summarization. *J. Artif. Intell. Res. (JAIR)*, 2004. 22: p. 457-479.

Escudero, G., Marquez , L. ve Rigau, G. (2000a). Naive Bayes and exemplar-based approaches to word sense disambiguation revisited. *In Proceedings of the 14th European Conference on Artificial Intelligence* (ECAI, Berlin, Germany). 421–425.

Escudero, G., Marquez , L. ve Rigau, G. (2000b). On the portability and tuning of supervised word sense disambiguation. I*n Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora* (EMNLP/VLC, Hong Kong, China). 172–180.

Fleiss, J.L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, Vol. 76, No. 5 pp. 378--382

Florian, R., Cucerzan, S., Schafer, C. ve Yarowsky, D. (2002). Combining classifiers for word sense disambiguation. *J. Nat. Lang. Eng*. 8, 4, 1–14.

Fujii, A., Inuı, K., Tokunaga, T. ve Tanaka, H. (1998). Selective sampling for example-based word sense disambiguation. *Computat. Ling*. 24, 4, 573–598.

Gale, W.A., Church, K.W. ve Yarowsky, D. (1992a). A method for disambiguating word senses in a large corpus. *Computers and the Humanities, 1992*. **26**(5-6): p. 415-439.

Gale, W.A., Church, K.W. ve Yarowsky, D. (1992b). One sense per discourse. In *Proceedings of the workshop on Speech and Natural Language*. 1992. Association for Computational Linguistics.

Gale, W.A., Church, K.W. ve Yarowsky, D. (1992c). Work on statistical methods for word sense disambiguation. *In Proceedings AAAI Fall Symposium on Probabilistic Approaches to Natural language, Cambridge, MA, 54-60.*

Golub, G.H. ve van Loan C.F. (1989). *Matrix computations.* The John Hopkins University Press, Baltimore, MD.

Göz, İ. (2003). *Yazılı türkçenin kelime sıklığı sözlüğü*. Vol. 823. 2003: Türk Dil Kurumu.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P. ve Witten IH. (2009). The WEKA data mining software: an update. *ACM SIGKDD explorations newsletter 2009*; 11: 10-18.

Hinton, G.E. ve Salakhutdinov, R. R. (2006). Reducing the dimensionality of data with neural networks. *Science*, 313(5786), 504-507.

Hirst, G. (1987). Semantic Interpretation and the Resolution of Ambiguity. Cambridge University Press. Cambridge, England.

Hirst, G. ve St-Onge, D. (1998). Lexical chains as representations of context for the detection and correction of malapropisms. *WordNet: An electronic lexical database*, 305, 305-332.

Hoste, V., Hendrickx, I., Daelemans, W. ve Van Den Bosch, A. (2002). Parameter optimization for machine learning of word sense disambiguation. *J. Nat. Lang. Eng.* 8, 4, 311–325.

Ide, N.M. ve Veronis, J. (1990). Very large neural networks for word sense disambiguation. *In Proceedings of the 9th European Conference on Artificial Intelligence*, ECAI90, pp. 366 - 368.

Ilgen, B., Adali, E. ve Tantug, A. (2012). The impact of collocational features in Turkish Word Sense Disambiguation. In *Intelligent Engineering Systems (INES), 2012 IEEE 16th International Conference on*. 2012. IEEE.

Ilgen, B., Adali, E. ve Tantug, A. (2013). A Comparative Study to Determine the Effective Window Size of Turkish Word Sense Disambiguation Systems, *in Information Sciences and Systems 2013*. Springer. p. 169-176.

Joachims, T. (1998). Text categorization with support vector machines: Learning with many relevant features. *In Proceedings of the 10th European Conference on Machine Learning* (ECML, Heidelberg, Germany). 137–142.

Kelly, E.F. ve Stone, P.J. (1975). Computer recognition of English word senses. Vol. 13. 1975: North-Holland.

Keok, L.Y. ve NG, H.T. (2002). An empirical evaluation of knowledge sources and learning algorithms for word sense disambiguation. *In Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing* (EMNLP, Philadelphia, PA). 41–48.

Klein, D., Toutanova, K., Ilhan, T. H., Kamvar, S. D. ve Manning, C. D. (2002). Combining heterogeneous classifiers for word-sense disambiguation. *In Proceedings of the ACL workshop on Word Sense Disambiguation: Recent Successes and Future Directions* (Philadelphia, PA). 74–80.

Kruskal, J.B. (1956). On the shortest spanning subtree of a graph and the traveling salesman problem. *In: Proceedings of the American Mathematical Society*, volume 7, pp. 48-50.

Leacock, C., Miller, G.A. ve Chodorow, M. (1998). Using corpus statistics and WordNet relations for sense identification. *Computational Linguistics*, 24(1), 147-165.

Leacock, C., Towell, G. ve Voorhees, E. (1993). Corpus-based statistical sense resolution. *In proceedings of the ARPA Human Language Technology Workshop.*

Lee, H., Baek, D.H. ve Rim H.C. (1997). Word sense disambiguation based on the information theory. *In Proceedings of Research on Computational Linguistics Conference*.

Lesk, M. (1986). Automatic Sense Disambiguation: How to tell a Pine Cone from an Ice cream cone. *In Proceedings of the 1986 SIGDOC Conference*.

Li, H. ve Takeuchi, J. (1997). Using evidence that is both strong and reliable in Japanese homograph disambiguation. *SIG-NL, Information Processing Society of Japan*, 1997: p. 53-59.

Lin, D. (1997). Using syntactic dependency as local context to resolve word sense ambiguity. *In Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics (pp. 64-71). Association for Computational Linguistics.*

Lin, D. (1998). Automatic retrieval and clustering of similar words. *In Proceedings of the 17th International Conference on Computational linguistics (COLING, Montreal, P.Q., Canada). 768–774.*

Lin, D. ve Pantel, P. (2002). Discovering word senses from text. I*n Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (Edmonton, Alta., Canada). 613– 619.*

Luk, K.A. (1995). Statistical sense disambiguation with relatively small corpora using dictionary definitions. *In Proceedings of the 33rd Annual Meetings of the Association for Computational Linguistics*, pp. 181-188.

McRoy, S. (1992). Using Multiple Knowledge Sources for Word Sense Discrimination. *Computational Linguistics*, 18(1).

Mihalcea, R. (2004). Co-training and self-training for word sense disambiguation. *In Proceedings of the Conference on Computational Natural Language Learning (CoNLL-2004).*

Mihalcea, R. (2005). Unsupervised large-vocabulary word sense disambiguation with graph-based algorithms for sequence data labeling. *In Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing* (pp. 411-418). Association for Computational Linguistics.

Mihalcea, R. (2006). Knowledge-based methods for WSD. Word Sense Disambiguation: Algorithms and Applications, 107-131.

Mihalcea, R. ve Faruque, E. (2004). Senselearner: Minimally supervised word sense disambiguation for all words in open text. *In Proceedings of ACL/SIGLEX Senseval* (Vol. 3, pp. 155-158).

Mihalcea, R. ve Moldovan, D. I. (1999). A method for word sense disambiguation of unrestricted text. *In Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*(pp. 152-158). Association for Computational Linguistics.

Mihalcea, R. ve Tarau, P. (2004). TextRank: Bringing order into texts. *In Proceedings of EMNLP*. 2004. Barcelona, Spain.

Mihalcea, R., Tarau, P. ve Figa, E. (2004). Pagerank on semantic networks with application to word sense disambiguation. *In Proc. of COLING.*

Mihăltz, M. (2005). Towards A Hybrid Approach to Word-Sense Disambiguation in Machine Translation. *In RANLP-2005 Workshop: Modern Approaches in Translation Technologies.*

Milgram, S. (1967). The small world problem. *Psychology today*, 2(1), 60-67.

Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D. ve Miller, K. (1990). Five Papers on WordNet. *Special Issue of International Journal of Lexicography*, 3(4).

Miller, G.A., Chodorow, M., Landes, S., Leacock, C. ve Robert G.T. (1994). Using a semantic concordance for sense identification. *In Proceedings of the ARPA Human Language Technology Workshop.*

Mooney, R.J. (1996). Comparative experiments on disambiguating word senses: an illustration of the role of bias in machine learning. *In Eric Brill, Kenneth Church, Editors, Proceedings of the Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Somerset, New Jersey, 82-91.*

Moro, A., Raganato, A. ve Navigli, R. (2014). Entity linking meets word sense disambiguation: a unified approach. *Transactions of the Association for Computational Linguistics*, 2, 231-244.

Murata, M., Utiyama, M., Uchimoto, K., Ma, Q. ve Isahara, H. (2001). Japanese word sense disambiguation using the simple Bayes and support vector machine methods. *In The Proceedings of the Second International Workshop on Evaluating Word Sense Disambiguation Systems* (pp. 135-138). Association for Computational Linguistics.

Navigli, R. (2006). Online word sense disambiguation with structural semantic interconnections. *In Proc. of EACL.*

Navigli, R. (2009). *Word sense disambiguation: A survey.* ACM Computing Surveys (CSUR), 2009. 41(2): p. 10.

Navigli, R. ve Ponzetto, S.P. (2010). BabelNet: Building a very large multilingual semantic network. *In Proceedings of the 48th annual meeting of the association for computational linguistics. 2010. Association for Computational Linguistics.*

Navigli, R. ve Ponzetto, S.P. (2012). BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 2012. 193: p. 217-250.

Navigli, R. ve Velardi, P. (2005). Structural semantic interconnections: a knowledge-based approach to word sense disambiguation. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 2005. 27(7): p. 1075-1086.

Ng, H.T. (1997). Getting serious about word sense disambiguation. *In Proceedings of the ACL SIGLEX Workshop on Tagging Text with Lexical Semantics: Why, What, and How?* (Washington D.C.). 1–7.

Ng, H.T. ve Lee, H.B. (1996). Integrating multipe knowledge sources to disambiguate word sense: an exemplar-based approach. *In Proceedings of the 34th Annual Meetings of the association for Computational Linguistics*, pp. 40-47.

Niu, C., Li, W., Srihari, R. ve Li, H. (2005). Word independent context pair classification model for word sense disambiguation. *In Proceedings of the 9th Conference on Computational Natural Language Learning (CoNLL, Ann Arbor, MI).*

Oflazer, K. (1994). Two-level description of Turkish morphology. *Literary and linguistic computing*, 1994. 9(2): p. 137-148.

Orhan, Z. (2006). *Türkçe Metinlerdeki Anlam Belirsizliği Olan Sözcüklerin Bilgisayar Algoritmaları İle Anlam Açıklaştırması*. Doktora Tezi. 2006.

Orhan, Z., Çelik, E. ve Demirgüç, N. (2007). SemEval-2007 task 12: Turkish lexical sample task. *In Proceedings of the 4th International Workshop on Semantic Evaluations*. Association for Computational Linguistics.

Patwardhan, S., Banerjee, S. ve Pedersen, T. (2003). Using measures of semantic relatedness for word sense disambiguation. *In Computational linguistics and intelligent text processing* (pp. 241-257). Springer Berlin Heidelberg.

Pedersen, T. ve Bruce, R. (1997a). A new supervised learning algorithm for word sense disambiguation. *In Proceedings of the 14th National Conference on Artificial Intelligence (AAAI-97), Providence, RI, 254-267.*

Pedersen, T. ve Bruce, R. (1997b). Distinguishing word senses in untagged text. *In Proceedings of the 1997 Conference on Empirical Methods in Natural Language Processing (EMNLP, Providence, RI). 197– 207.*

Quinlan, J.R. (1986). Induction of decision trees. *Mach. Learn*. 1, 1, 81–106.

Quinlan, J.R. (1993). Programs for Machine Learning. *Morgan Kaufmann*, San Francisco, CA.

Resnik, P.S. (1993). Selection and information: a class-based approach to lexical relationships. IRCS Technical Reports Series, 200.

Resnik, P. (1995). Using information content to evaluate semantic similarity in a taxonomy. arXiv preprint cmp-lg/9511007.

Rigau, G., Atserias, J. ve Agirre, E. (1997). Combining Unsupervised Lexical Knowledge Methods for Word Sense Disambiguation. *Proceedings of ACL-EACL*, Madrid, Spain.

Rivest, R.L. (1987). Learning decision lists. *Mach. Learn*. 2, 3, 229–246.

Say, B., Zeyrek, D., Oflazer, K. ve Özge, U. (2002). Development of a corpus and a treebank for present-day written Turkish. *In Proceedings of the eleventh international conference of Turkish linguistics*.

Schütze, P. (1992). Dimensions of meaning. *In Supercomputing' 92. Proceedings of the 1992 ACM/IEEE Conference on Supercomputing. IEEE Computer Society Press*, Los Alamitos, CA. 787–796.

Schütze, P. (1998). Automatic word sense discrimination. *Computat. Ling.* 24, 1, 97–124.

Shinnou, H. (2001). Learning of word sense disambiguation rules by Co-training, checking co-occurrence of features. 自然言語処理, 64(5), 145-5.

Shinnou, H. ve Sasaki, M. (2003). Unsupervised learning of word sense disambiguation rules by estimating an optimum iteration number in the EM algorithm. *In Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4*. 2003. Association for Computational Linguistics.

Sinha, R. ve Mihalcea, R. (2007). Unsupervised graph-based word sense disambiguation using measures of word semantic similarity. *In Proc. of ICSC.*

Tsatsaronis, G., Vazirgiannis, M. ve Androutsopoulos, I. (2007). Word sense disambiguation with spreading activation networks generated from thesauri. *In Proc. of IJCAI*, pages 1725–1730.

Url-1 <*http://www.senseval.org*>, erişim tarihi 15.08.2015.

Url-2 < *http://www.synapse-fr.com* >, erişim tarihi 15.08.2015.

Véronis, J. (2004). Hyperlex: lexical cartography for information retrieval. *Computer Speech & Language*, 2004. 18(3): p. 223-252.

Walker, D. (1987). Knowledge resource tools for accessing large text files. *In Machine Translation: Theoretical and Methodological Issues.*

Watts, D.J. ve Strogatz, S.H. (1998). Collective dynamics of 'small-world'networks. *nature*, 393(6684), 440-442.

Weiss, S.F. (1973). Learning to disambiguate. *Information Storage and Retrieval*, 1973. 9(1): p. 33-41.

Widdows, D. ve Dorow, B. (2002). A graph model for unsupervised lexical acquisition. *In Proceedings of the 19th international conference on Computational linguistics-Volume 1*. 2002. Association for Computational Linguistics.

Wiriyathammabhum, P., Kijsirikul, B., Takamura, H. ve Okumura, M. (2012). Applying Deep Belief Networks to Word Sense Disambiguation. arXiv preprint arXiv:1207.0396.

Yarowsky, D. (1992). Word sense disambiguation using statistical models of Roget's categories trained on large corpora. *In Proceedings of the Fifteenth International Conference on Computational Linguistics, pp. 189-196.*

Yarowsky, D. (1993). One sense per collocation. *In Proceedings of the workshop on Human Language Technology*. Association for Computational Linguistics.

Yarowsky, D. (1995). Unsupervised word sense disambiguation rivaling supervised methods. *In Proceedings of the 33rd annual meeting on Association for Computational Linguistics.* Association for Computational Linguistics.

Yoon, Y., Seon C.N., Lee S. ve Seo J. (2006). Unsupervised word sense disambiguation for Korean through the acyclic weighted digraph using corpus and dictionary. *Information processing & management*, 2006. 42(3): p. 710-722.

Yuret, D. ve Türe, F. (2006). Learning morphological disambiguation rules for Turkish. *In Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics.* Association for Computational Linguistics.

Zhao, Y., Karypis, G. ve Fayyad, U. (2005). Hierarchical clustering algorithms for document datasets. *Data Mining and Knowledge Discovery*, 2005. 10(2): p. 141-168.

Zipf, G. K. (1949). *Human Behavior and the Principle of Least Efiort*.Cambridge, MA: Addison—Welsey.

EKLER

EK A: Geliştirilen Uygulama
EK B: Değerlendirme Yöntemleri
EK C: Terimler Sözlüğü

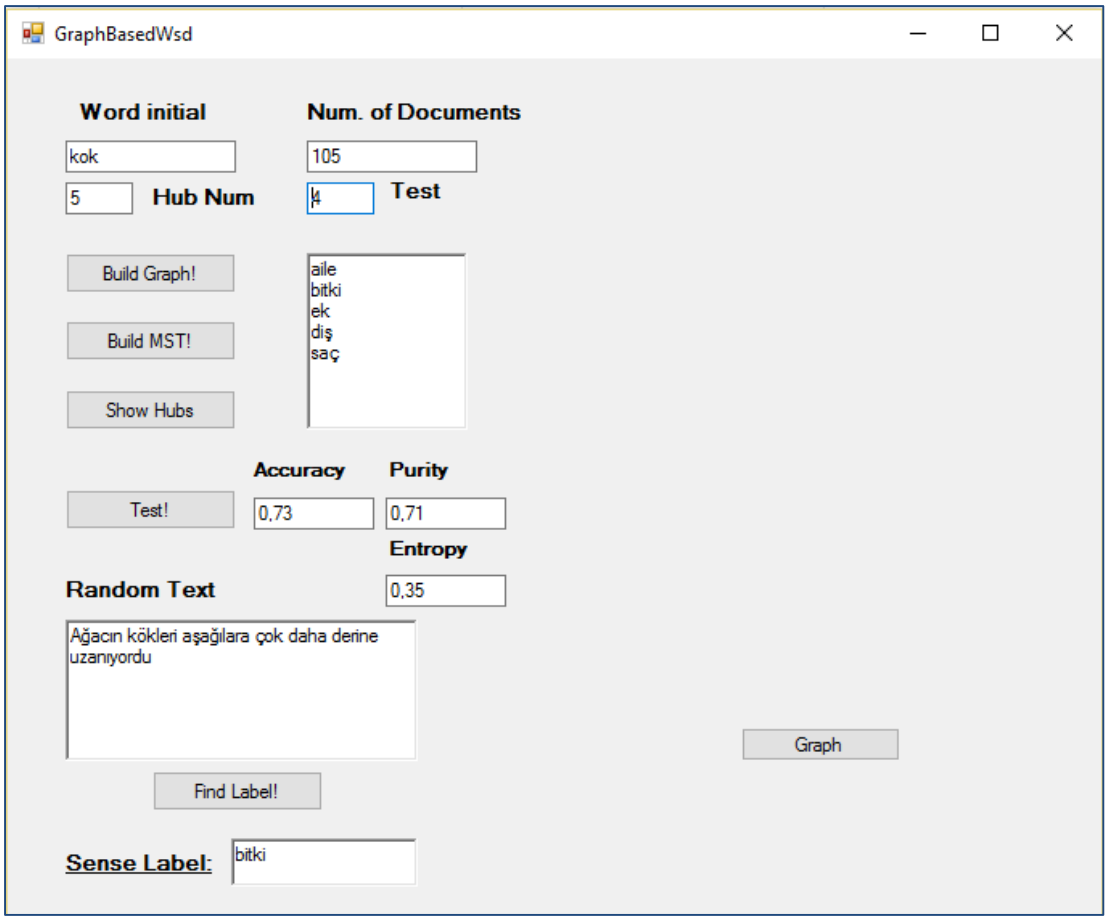
EK A

Geliştirilen Uygulama

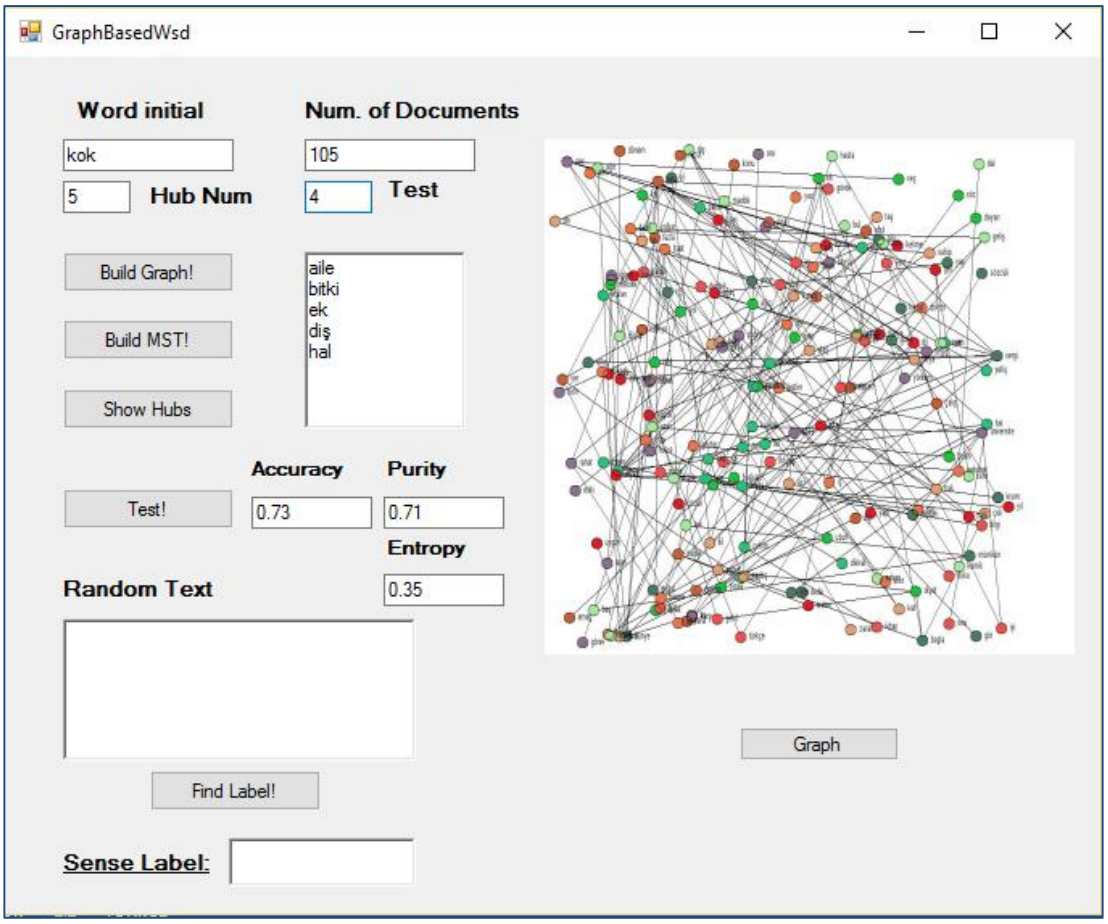
Bu bölümde çizge tabanlı algoritmaya ilişkin çıktı ve anlatımlar yer almaktadır. Şekil A.1'de kök sözcüğüne ilişkin yüz beş örnek ele alınmaktadır. Program merkez düğüm sayısını kullanıcının girebileceği şekilde tasarlanmıştır. Test bölümünde görülen alan, sınama için ayrılacak örnekleri göstermektedir. İlk örnekten son örneğe kadar tüm dokümanlar içinde ona bölümünden kalanı dört olan sıralamadaki örnekler (dört, on dört, yirmi dört, otuz dört…yüz dört numaralı örnekler) eğitim aşamasında atlanmakta ve sınama için kullanılmaktadır. Birden ona kadar farklı test değerleri için program çalıştırılarak on katlı çapraz doğrulama uygulanmaktadır.

Uygulamaya sonradan eklenen bir modül ise kullanıcı tarafından rastgele girilecek metinlerde kullanılan sözcük anlamını belirlenen merkez düğümlere atamaktadır. Örneğin, kullanıcı "Ağacın kökleri aşağılara çok daha derinlere uzanıyordu" cümlesini yazdığında, belirlenen merkez düğüm 2 numaralı anlam olan bitki olarak bulunmaktadır. Programın esas işleyişinde biçimbilimsel olarak analiz edilmiş ve belirsizlik giderimi yapılmış örnekler kullanılmaktadır. Bu nedenle metin kullanıcı tarafından girildiğinde öncelikle sözcük gövde biçimlerinin elde edilmesi ve sonrasında EKKA üzerinde aranabilmesi için Zemberek modülü kullanılmıştır.

Programda ayrıca oluşan ağaç yapısı görsel olarak ta oluşturulmaktadır. Görsel bölümle ilgili kodlama Matlab'da yapılmış, C# içerisinden çağrılarak kullanımı sağlanmıştır. Şekil A.2 kök sözcüğü için oluşturulan örnek ağaç yapısını göstermektedir.



Şekil A.1 : Kök sözcüğü örnek ekran görüntüsü – 1.



Şekil A.2 : Kök sözcüğü örnek ekran görüntüsü – 2.

EK B

B.1 Değerlendirme Yöntemleri

Çalışmamızda yapılan değerlendirmelerde 10 katlı çapraz doğrulama kullanılmıştır. ÇD bir modele ilişkin doğruluk değerinin sınanmasında kullanılan bir tekrar örnekleme tekniğidir. ÇD veriyi eşit büyüklükte olacak şekilde (F1, F2,….Fk) gibi k parçaya ayırır. Yapılan her bir deneyde Fi kümesi sınama kümesi olarak kullanılırken, geriye kalan (k-1) küme eğitim aşamasında kullanılmaktadır. Deneyler tamamlandıktan sonra sonuçların ortalaması ve standart sapması hesaplanır. Yapılan çalışmalarda k değerinin çok küçük ya da çok büyük olduğu durumlarda ortaya çıkabilecek olumsuzluklar olabileceği belirtilmiştir. Örneğin küçük k değerlerinin, doğruluk değerinin güvenilirliğini düşürmesi söz konusu olabileceği gibi, bu değer çok büyük seçildiğinde varyans ve hesaplama zamanının artmasına yol açabilmektedir. Deneysel çalışmalar k = 10 olarak seçildiğinde sözü edilen olumsuzluklara yok açmayan uygun sonuçlar elde edildiğini göstermiştir.

Elde edilen sonuçların analizinde kullanılan hata matrisi bir sınıflandırıcının ne derece iyi tahmin edebileceğini gösteren bir tablodur. Hata matrisi ile çıkan hata oranının basit şekilde sunulması yerine ayrıntılı bilgi vermesi olumlu bir yönüdür. Elde edilen sonuçlar Doğru Pozitif (DP), Doğru Negatif (DN), Yanlış Pozitif (YP) ve Yanlış Negatif (YN) seçenekleri ile ifade edilmektedir.

Çizelge B.1 : Hata matrisinde kullanılan terimler.

DP : A sınıfına ait olup A sınıfı olarak seçilen örnekler
DN : A sınıfına ait olmayan ve A sınıfı olarak seçilmeyen örnekler
YP : A sınıfına ait olmayan, ancak A sınıfı olarak seçilen örnekler
YN : A sınıfına ait olan, ancak A sınıfı olarak seçilmeyen örnekler

Çalışmamızda hata matrisinden elde edilen doğruluk, tutturma ve bulma gibi farklı değerlendirme ölçütleri kullanılmıştır:

Tutturma (T): A sınıfına ait olan örneklerin A sınıfı olarak seçilen tüm örneklere oranıdır.

$$T = (DP) / (DP + YP) \quad (B.1)$$

Bulma (B): A sınıfına ait olarak seçilen örneklerin A sınıfındaki tüm örneklere oranıdır.

$$B = (DP) / (DP + YN) \quad \text{(B.2)}$$

Tutturma ve bulma değerlerinin her ikisinin de yüksek olması hem doğruluk oranını hem de seçilen örneklerin seçilmesi gereken örneklerin ne kadarını kapsadığını göstermesi açısından önemlidir.

Doğruluk (D): A sınıfına ait olarak seçilen örneklerin A sınıfındaki tüm örneklere oranıdır.

$$D = (DP+DN)/(DP+YN+DN+YP) \quad \text{(B.3)}$$

B.2 Denetimsiz Değerlendirme Ölçütleri

Düzensizlik: Kümelerin bir sınıftan örnek içermesine ilişkin derece hesaplanır. i kümesine ilişkin düzensizlik aşağıdaki gibi hesaplanmaktadır.

$$\boldsymbol{E_i = -\sum_{j=1}^{l} \frac{n_{ij}}{n_i} log\left(\frac{n_{ij}}{n_i}\right)} \quad \text{(B.4)}$$

(B.4)'te n_{ij} j kümesindeki i sınıfına ait noktaların sayısıdır. n_i küme içindeki örnek sayısı l ise sınıf sayısıdır. k toplam küme sayısı ve n toplam örnek sayısı olmak üzere, toplam düzensizlik (B.5) ile verilmektedir.

$$E = \sum_{i=1}^{k} \frac{n_i}{n} E_i \quad \text{(B.5)}$$

Saflık: Oluşturulan bir kümenin tek sınıftan örnek içerme derecesini, diğer deyişle saflık derecesini verir. i kümesi için saflık değeri aşağıdaki gibi bulunmaktadır.

$$\boldsymbol{Pur_i = \max_j \frac{n_{ij}}{n_i}} \quad \text{(B.6)}$$

Toplam saflık değerini veren bağıntı:

$$\boldsymbol{Pur = \sum_{i=1}^{k} \frac{n_i}{n} Pur_i} \quad \text{(B.7)}$$

olarak bulunacaktır.

EK C

Bu bölümde tez kapsamında kullanılan Türkçe terimlerin literatürde yer alan İngilizce karşılıkları verilmektedir.

10 Katlı Çapraz Doğrulama	10 Fold Cross Validation
Ağırlıklı Oylama	Weighted Voting
Alt Sınıf Çıkarımı	Subcategorization Acquisition
Anlamsal İşaretleme	Semantic Annotation
Anlamsal Rollerin Tanımlanması	Identification of Semantic Roles
Ayrıştırma	Parsing
Bağlam Kümeleme	Context Clustering
Beklenti Maksimizasyonu	Expectation Maximization
Belirtici Sözcükler	Indicative Words
Bellek Tabanlı Öğrenme	Memory Based Learning
Benzetilmiş Tavlama	Simulated Annealing
Biçimbilim	Morphology
Bilgi Edinim Darboğazı	Knowledge Acquisition Bottleneck
Bilgi Tabanlı Yöntemler	Knowledge Based Methods
Bilgisayarla Okunabilir Sözlükler	Machine Readable Dictionaries
Bilgisayarlı Çeviri	Machine Translation
Bilgisayarlı Dilbilim	Computational Linguistics
Bilgiye Erişim	Information Retrieval
Birliktelik Yöntemleri	Ensemble Methods
Birlikte Eğitme	Co-training
Birliktelik Çizgesi	Cooccurance Graph
Bölgesel Özellikler	Local Features
Bulma	Recall
Çekirdek Küme	Seed Set
Çoğunluk Oylaması	Majority Voting
Çok Dilli İşaretleme	Multilingual Annotation
Denetimli Öğrenme	Supervised Learning
Denetimsiz Öğrenme	Unsupervised Learning
Destek Vektör Makineleri	Support Vector Machines
Doğal Dil İşleme	Natural Language Processing
Doğruluk	Accuracy
Düzensizlik	Entropy
En Baskın Anlam	Most Frequent Sense
En Fazla Olabilirlik	Maximum Likelihood
En Fazla Düzensizlik Birleşimi	Maximum Entropy Combination
En Küçük Kapsayan Ağaç	Minimum Spanning Tree
Etiketli	Labeled
Etiketsiz	Unlabeled
Evrensel Özellikler	Global Features
Gizli Anlamsal İndeksleme	Latent Semantic Analysis
Görünüm Bilgisi	Lexical
Gürültü	Noise

Bir Söz Öbeği İçin Bir Anlam	One sense per collocation
İçerik Sözcük	Content Word
İşlev Sözcük	Stop Word
İşlevsel Ağaç	Functional Tree
Karakteristik Yol Uzunluğu	Characteristic Path Length
Karar Ağacı	Decision Tree
Kaynak Kısıtlı Diller	Resource Poor Languages
Kendiliğinden Eğitme	Self-training
Küçük Dünya	Small World
Kümeleme Katsayısı	Clustering Coefficient
Makine Öğrenmesi	Machine Learning
Mantıksal Biçimler	Logic Forms
Olasılık Karışımı	Probability Mixture
Önyükleme	Bootstrapping
Örnek Tabanlı Öğrenme	Exemplar Based Learning
Saflık	Purity
Sınır Başarım	Baseline
Sıralama Tabanlı Birleştirme	Rank-based Combination
Söz Öbeği	Collocation
Sözcük Anlam Ayrıştırma	Word Sense Discrimination
Sözcük Anlam Belirsizliği Giderme	Word Sense Disambiguation
Sözcük Etiketleyici	POS tagger
Sözcük Kesesi	Bag of Words
Sözcük Kümeleme	Word Clustering
Sözcüksel Örnek	Lexical Sample
Tek Anlamlı Yakın Sözcükler	Monesemous Relatives
Tekil Değer Ayrışımı	Singular Value Decomposition
Tüm Sözcükler	All Words
Tutturma	Precision
Uçtan Uca Uygulama	End to End Application
Veri Seyrekliği	Data Sparseness
Yapısal Anlamsal Bağlantılar	Structural Semantic Interconnections
Yarı Denetimli Öğrenme	Semi Supervised Learning
Yüksek Yoğunluklu Bileşen	High Density Component

ÖZGEÇMİŞ

Ad-Soyad : Bahar İLGEN

Doğum Tarihi ve Yeri : İzmir, 28.06.1977

E-posta : b.ilgen@iku.edu.tr , baharilgen@gmail.com

ÖĞRENİM DURUMU:

- **Lisans** : 2000, Dokuz Eylül Üniversitesi, Mühendislik Fakültesi, Çevre Mühendisliği Bölümü
- **Yükseklisans** : 2006, Ege Üniversitesi, Uluslararası Bilgisayar Enstitüsü, Bilgisayar Bilimleri

DOKTORA TEZİNDEN TÜRETİLEN YAYINLAR, SUNUMLAR VE PATENTLER:

- **Ilgen, B.,** Adali, E., and Tantug, A. 2015. Exploring Feature Sets for Turkish Word Sense Disambiguation. *Turkish Journal of Electrical Engineering and Computer Sciences*, DOI: 10.3906/elk-1408-77
- **İlgen, B.,** & Adalı, E. 2014. Exploring the Effect of Bag-of-Words and Bag-of-Bigram Features on Turkish Word Sense Disambiguation. *Türkiye Bilişim Vakfı Bilgisayar Bilimleri ve Mühendisliği Dergisi*, 8(8).
- **Ilgen, B.,** Adali, E. and Tantug, A. 2013. A Comparative Study to Determine the Effective Window Size of Turkish Word Sense Disambiguation Systems, *In Information Sciences and Systems 2013*. 28-29 October 2013. Paris, France. Springer. p. 169-176.
- **Ilgen, B**., Adali, E. and Tantug, A. 2012. Building up Lexical Sample Dataset for Turkish Word Sense Disambiguation. Innovations in Intelligent Systems and Applications (INISTA) IEEE International Symposium, July 2-4, 2012. Trabzon, Turkey.

- **Ilgen, B**., Adali, E. and Tantug, A. 2012. The impact of collocational features in Turkish Word Sense Disambiguation. Intelligent Engineering Systems (INES), 2012 IEEE 16th International Conference, June 13-15, 2012, Lisbon, Portugal.