Türkçe için Karşılaştırmalı bir Kelime Anlamı Belirginleştirme Uygulaması

Mehmet Ali Aksoy Tüysüz[1], Erdal Güvenoğlu[2]

[1] Maltepe Üniversitesi, Yazılım Mühendisliği Bölümü, İstanbul

[2] Maltepe Üniversitesi, Bilgisayar Mühendisliği Bölümü, İstanbul

aksoytuysuz@maltepe.edu.tr, erdalguvenoglu@maltepe.edu.tr

Özet: Kelime anlamı belirginleştirme (KAB), bir kelimenin bulunduğu bağlamda hangi anlamı ile kullanıldığının otomatik olarak belirlenebilmesidir. Makine çevirisi, bilgi çekme, içerik ve tematik analiz, dilbilgisi analizi, bağlantılı-metin tarama, konuşma işleme, metin işleme gibi bir çok alanda kelime anlamı belirginleştirmeden faydalanılabilmektedir. Bu çalışmada, Türkçe için Semeval-2007 çalıştayının sözcüksel örnekler kısmına katılan çalışmanın verileri kullanılarak makine öğrenmesi teknikleri uygulanmış ve elde edilen sonuçlar çalıştaydaki sonuçlarla karşılaştırılmıştır.

Anahtar Sözcükler: Kelime Anlamı Belirginleştirme, Sözcüksel Örnekler, Semeval, Makine Öğrenmesi

A Comparative Word Sense Disambiguation Application for Turkish

Abstract: Word Sense Disambiguation (WSD) can be defined as the process of automatically determining which sense or meaning of a word is used in a particular context. Machine translation, information retrieval, content and thematic analysis, grammatical analysis, speech processing, text processing/mining are amongst the fields that benefit from word sense disambiguation. In this study, machine learning techniques are applied to the Semeval-2007 workshop's Turkish Lexical Sample Task data and the obtained results are compared.

Keywords: Word Sense Disambiguation, Lexical Sample Task, Semeval, Machine Learning.

1. Giriş

Makine çevirisi alanındaki en eski ve zor problemlerden biri olan Kelime Anlamı Belirginleştirme (KAB) için bir çok yöntem önerilmiş ve uygulamalar yapılmıştır. Türkçe için yapılan KAB uygulamaları ve önerilen yöntem sayısı, varolan dilbilimsel kaynak azlığı gibi sebepler dolayısıyla (İngilizce düşünüldüğünde) göreli olarak azdır. Farklı diller için geliştirilen yöntemlerin değerlendirmesinin yapıldığı uluslararası bir çalıştay olan Senseval'in dördüncüsü Semeval adıyla anılmaktadır ve 2007 yılında yapılmıştır. İlk kez bu çalıştayda ve sözcüksel örnekler alanında Türkçe için bir çalışma yer almıştır[9]. Daha sonraki çalıştaylarda ise herhangi bir katılım olmamıştır.

Bu çalışmada, Semeval-2007 çalıştayına Türkçe için katılan çalışmanın verileri aynen kullanılarak makine öğrenmesi algoritmaları uygulanmış ve elde edilen sonuçlar, çalıştayda elde edilenler ile karşılaştırılmıştır.

Çalışmanın organizasyonu şu şekildedir: İlk olarak KAB ve Senseval/Semeval hakkında bilgi verilmiştir. Daha sonra gerçekleştirilen çalışma ayrıntıları açıklanmıştır. Son olarak da elde edilen sonuçlar karşılaştırmalı olarak değerlendirilmiş ve gelecekte yapılması düşünülen çalışmalar hakkında bilgi verilmiştir.

2. Kelime Anlamı Belirginleştirme (KAB)

Makine çevirisi gibi bir takım doğal dil işleme (DDİ) uygulamaları doğrudan kullanılabilir iken doğrudan kullanılamayan ancak diğer DDİ uygulamaları tarafından performans iyileştirme vb. amaçlarla faydalanılan uygulamalar da mevcuttur. Kelime anlamı belirginleştirme (KAB) bunlardan biridir. KAB, kısaca birden fazla anlama sahip bir kelimenin, kullanıldığı bağlamdaki (context) anlamını bilgisayarlı/otomatik olarak belirleyebilmek olarak tanımlanabilir. Örneğin, Türkçe "yüz" kelimesinin sayı, insan organı (surat) ve yüzmek fiili olarak kullanıldığı farklı bağlamlar olabilir. Benzer şekilde İngilizce "light" kelimesinin sırasıyla isim, sıfat ve fiil olarak kullanımlarına göre anlamı "ışık", "hafif" ve "yakmak" şeklinde farklılıklar gösterecektir. Verilen örneklerden de anlaşılacağı üzere bir kelimenin hangi anlamıyla kullanıldığının doğru şekilde belirlenebilmesi, yüksek başarımlı makine çevirisi yapabilmek için gerekli bir işlemdir. Dolayısıyla, KAB için başlıca uygulama alanı makine çevirisidir. Ancak uygulama alanları bununla sınırlı değildir. DDİ'ye dayalı bir çok alanda performans iyileştirme vb. amacı ile bir alt modül olarak kullanılabilmektedir. Örneğin, internet üzerinde sorgulama ile veri çekme işlemi yapılacağı zaman kullanıcının kelimeyi hangi anlamı ile kullandığının belirlenebilmesi, getirilecek sonuçların konu ile ilgili olması açısından önemlidir. Bu amaçla da KAB işlemi uygulanabilir. En basit şekli ile birden fazla kelimeyi yanyana getirerek sözcük anlamını daha belirgin hale getirme işlemi kullanıcı tarafından KAB için yapılan bir yardım olarak düşünülebilir.

KAB zorluk seviyesi olarak "AI-complete" olarak değerlendirilmektedir. Yani yapay zekadaki tüm zor problemler çözüldükten sonra çözülebilecek bir problemdir [3]. Dolayısıyla, zor bir problemdir ve dilbilim alanında makine çevirisi ile birlikte uzun süredir uğraşılan eski problemlerden biridir.

KAB işlemi eski bir problem olduğu için bir çok farklı yöntem geliştirilmiştir. Geliştirilen/kullanılan yöntemler yapay zeka tabanlı yöntemler, bilgi tabanlı (knowledge-based) yöntemler ve derlem tabanlı (corpus-based) yöntemler ana başlıkları altında toplanmaktadır [1, 3]. Derlem tabanlı yöntemler de üzerinde anlam işaretlemesi yapılmış olup olmamasına göre kendi içinde sırasıyla denetimli (supervised) ve denetimsiz (unsupervised) olarak ikiye ayrılmaktadır.

KAB ile bir sözcüğün kullanıldığı bağlamda hangi anlamının aktive edildiğinin bulunması işlemi, anlamsal açıdan bir sınıflandırma işlemi olarak düşünülebilir. Bu sebeple makine öğrenmesi alanındaki bir çok algoritma bu alana uygulanmış ve bir kısmı son derece başarılı sonuçlar vermişlerdir. Denetimli derlem tabanlı yöntemler KAB alanında en başarılı olanlardır [1]. Üzerinde işaretleme yapılarak oluşturulmuş metin derlemeleri olan derlemlerden oluşturulan öğrenme kümesi (training set) üzerinden öğrenme işlemi yapan denetimli makine öğrenmesi teknikleri, önceden görmedikleri test verileri (test set) üzerinde başarılı anlam belirginleştime işlemleri yapabilmişlerdir. Sunulan çalışma için hazırlanan uygulama da denetimli derlem tabanlı ve makine öğrenmesi kullanan bir çalışmadır.

KAB için geliştirilen yöntemlerin değerlendirilmesi konusu da ayrı bir problem olmuştur. Araştırmacılar farklı kelimeler, anlam ayrımları, eğitim ve test kümeleri üzerinde çalışmış oldukları için farklı KAB yaklaşımlarının karşılaştırmalı olarak değerlendirmesi kolaylıkla yapılamamıştır. Örneğin ikili anlam ayrımları kullanan ve 12 kelime için işlem yapan bir algoritmanın bir sözlükten anlam ayrımları kullanan bir algoritma ile karşılaştırılması zordur [10]. Sistemlerin başarımlarının birbirleriyle karşılaştırılması konusundaki bu problem üç yılda bir düzenlenen Senseval çalıştayları ile çözülmeye çalışılmıştır. Çalıştayların ilk üç tanesi sırasıyla Senseval-1, Senseval-2, Senseval-3 adıyla anılmaktadır ve sadece KAB

işlemine odaklanmıştır. Ancak 2007 yılında düzenlenen dördüncüsü ile birlikte anlambilimsel analiz (semantic analysis) de eklenerek kapsamı genişletilmiş ve Semeval adıyla anılmaya başlanmıştır. Ayrıca konferans isimlerinin sonuna sıra numarası değil yıl eklenmiştir. Dolayısıyla 2007 yılında düzenlenen dördüncü çalıştayın ismi Semeval-2007 olmuştur. Semeval-2010'da üç yıllık sürenin uzun olduğu gündeme gelmiş ve bir sonraki çalıştay 2012 yılında gerçekleştirilmiştir. Sürenin kısalmasının yanında, çalıştayın yapıldığı her yıl her işin/değerlendirmenin yapılmamasına da karar verildiği için (iki yıl sonra) 2012 yılında gerçekleştirilen Semeval'de hiç KAB sistemine yer verilmemiştir. Son olarak 2013'te Semeval gerçekleştirilmiştir ve 2014 hazırlıkları da bu çalışma hazırlandığı esnada devam etmektedir.

Senseval/Semeval çalıştaylarında KAB sistemleri için metinlerde geçen hemen hemen tüm kelimelerin anlamının belirginleştirilmeye çalışıldığı bütün sözcükler (all-words) ve seçilen bir grup sözcüğün anlamının belirginleştirmeye çalışıldığı sözcüksel örnekler (lexical sample) işleri bulunmaktadır. Ayrıca değerlendirme işleminde de sistemler kaba ayrımlı (coarse grained) ve ince ayrımlı (fine grained) olarak sınıflandırılmıştır. Kaba ayrımlı değerlendirmede, bir kelimenin alt anlamları da ana anlamı ile aynı olarak kabul edilmiş ve sistemler buna göre değerlendirilmiştir. İnce ayrımlı değerlendirmede ise alt-anlam, ana anlam gibi ayrımlar göz önüne alınmayıp sistemlerin tam olarak doğru anlamı bulmalarına puan verilmiştir. Aksi durumlara puan verilmemiştir. Semeval-2007'ye Türkçe için katılan sistem[9], sözcüksel örnekler kısmına katılmış ve hem kaba hem de ince anlam belirginleştirme sonuçları vermiştir.

Senseval/Semeval çalıştaylarında KAB işleri için üç ana safha bulunmaktadır. Bunlar sırasıyla eğitim, test ve değerlendirme safhalarıdır. Eğitim safhasında katılımcılar kendileri için hazırlanan eğitim verileri üzerinden çok anlamlı kelimeler için anlam çıkarımında bulunmaktadırlar. Test safhasında, katılımcıların kendilerine verilen test verileri için KAB işlemi yapmaları istenmektedir. Değerlendirme safhasında da katılan sistemlerin elde ettikleri sonuçlar değerlendirilmektedir. Dolayısıyla, Senseval/Semeval çalıştaylarında eğitim ve test verileri bulunmakta, sistemlerin performansı test verileri için elde ettikleri değerlere göre yapılmaktadır. Semeval-2007'de Türkçe için kullanılmak üzere hazırlanmış olan verilere internet üzerinden ulaşılabilmektedir.

2.1 Semeval-2007 Türkçe Sözcüksel Örnekler Çalışması

Senseval/Semeval çalıştaylarının yapısına uygun olarak Türkçe sözcüksel örnekler çalışması için de öğrenme ve test/değerlendirme verileri hazırlanmıştır. Hazırlanan verilerin türlere göre dağılımı şu şekildedir: İsim türünde on, fiil türünde on ve (çalışmada) diğer diye sınıflandırılan (sıfat, zarf gibi) türleri içine alan altı kelime mevcuttur. Bu verilerin hazırlanması sırasında sözlük olarak, Türk Dil Kurumu (TDK) tarafından oluşturulan ve internet üzerinden erişilebilen güncel Türkçe sözlük kullanılmıştır. Örnek cümlelerin elde edilmesi için ağırlıklı olarak ODTÜ derlemi ve Sabancı Ağaç Yapılı Derlemi kullanılmıştır. Ancak bazı kelimeler için yeterli miktarda örnekleme elde edilememesi durumunda dışarıdan da örnek cümlelerin eklenmesi sağlanmıştır. Örneklemeler için kullanılacak özellikler Türkçe KAB için iyi sonuçlar verdiği bilinenlerden seçilmiştir [7,8]. Bunlar üç kelime çeşidi için ayrı ayrı kodlanmıştır: Hedef kelime, hedef kelimenin öncesindeki ilgili kelimeler ve hedef kelime sonrasındaki ilgili kelime. Belirtilen kelimeler için kullanılan özellikler kelime kökü, kelime türü, hal bilgisi, hedef kelime ile ilişkisinin ne olduğu, sahiplik bilgisi, ontolojik olarak her seviyede aldığı değerler ana başlıkları altında toplanabilir. Semeval-2007 çalışması için çalışmayı hazırlayan grup tarafından amaca yönelik olarak üç seviyeli küçük bir ontoloji oluşturulmuştur. Önceki ve sonraki kelimelerin özelliklerinin sayısı tamamen aynı iken hedef kelimenin kodlanan özelliklerinin sayısı

daha azdır. Hedef kelime için kök haldeki tür bilgisinin yanında yapım eki/çekim eki almış halinin tür bilgisi de girilmiştir. Önceki ve sonraki kelimelerde ise bu iki özelliğin yanına düzeltilmiş bir tür bilgisi daha eklenmiştir. Ayrıca önceki ve sonraki kelimeler için ontolojik olarak üç seviyede aldıkları değerler de özellik olarak kodlanmış iken hedef kelimede bu bilgiye yer verilmemiştir. Her kelime için bir dosya hazırlanmış ve belirtilen alanlar TAB karakteri ile birbirinden ayrılmış olarak kodlanmıştır. Ayrıca her örnek için dosya adı olarak (Sabancı Ağaç Yapılı derlemindeki dosya adı gibi) nereden elde edildiği, belirtilen dosyada kaçıncı cümle olduğu ve kelime örnek cümlede birden fazla defa geçiyorsa hangisinin hedef kelime olarak alındığı bilgisi örneklemelerin başına eklenmiştir. Örnek cümleden önce de ince ve kaba anlam olarak belirlenen anlamlardan hangilerinin bu örnekle kodlandığı bilgisi bulunmaktadır. Özelliklerin değerleri İngilizce olarak kodlanmıştır. Örneğin tür bilgisi olarak isim için noun, fiil için verb karşılıkları kullanılmıştır. Örnek bir satır aşağıdaki gibidir.

00053223769.xml 5 1 ara noun abstraction quantity time noun noun nom fl collocatıon ara noun noun nom fl modıfier tak verb abstraction cognition process verb adj ? Fl modıfier 2 2 #ara ara aklıma takılan soruları da anlatır gibi ; ona mı ; kendime mi ; bilmeden sordum .#

KAB işlemi gerçekleştirmek için Naive Bayes yaklaşımına benzer bir istatistiki yöntem uygulanarak öğrenme verileri üzerinden anlamlar için olasılıklar belirlenmiştir. Test aşamasında da verilen özelliklere göre anlamların olasılıkları hesaplanmış ve en yüksek skoru alan üç anlam cevap olarak seçilmiştir. Elde edilen sonuçlar Tablo 1'de görülmektedir. Sonuçlar, her kelime türü için kaba ayrımlı ve ince ayrımlı KAB işlemleri sonucu elde edilen duyarlık (precision) ve geri çağırım/anımsama (recall) değerleri olarak verilmiştir.

Tür	**İnce ayrım**		**Kaba ayrım**	
	Duyarlık	Geri çağırım	Duyarlık	Geri çağırım
İsim	0.15	0.50	0.65	0.43
Fiil	0.10	0.38	0.56	0.50
Diğer	0.13	0.50	0.57	0.44
Ortalama	0.13	0.46	0.59	0.46

Tablo 1. Semeval-2007 Türkçe Sözcüksel Örnekler Çalışmasının Sonuçları

Tablodan da görüldüğü gibi elde edilen sonuçlar çok yüksek bir başarımı göstermemektedir. Durum, ilgili çalışmanın yazarları tarafından öğrenme verisinin küçüklüğü sebebi ile beklenen bir sonuç olarak değerlendirilmektedir. Ayrıca, Türkçe için veri bulmakta yaşanan sorunlar göz önüne alındığında mevcut veri büyüklüğünün imkanlar dahilinde elde edilebileceğin en iyisi olduğu belirtilmektedir [9].

3. Uygulama

Geliştirilen uygulama basitçe şu adımlardan oluşmaktadır.

- Semeval-2007 sözcüksel örnekler Türkçe kategorisi için hazırlanan verilerin elde edilmesi.
- Elde edilen verilerin makine öğrenmesi algoritmaları ile değerlendirilebilmesi için WEKA yazılımının istediği biçime dönüştürülmesi.
- WEKA yazılımı aracılığı ile Semeval-2007 eğitim verilerinin kullanılarak algoritmaların eğitilmesi
- Bir önceki adımda gerçekleştirilen öğrenme işlemi ile Semeval-2007 test verileri üzerinde KAB işleminin gerçekleştirilmesi
- Elde edilen sonuçların Semval-2007'de elde edilen sonuçlarla karşılaştırılması ve değerlendirilmesi.

Uygulamada, Semeval-2007 çalıştayında kullanılan veriler üzerinde sadece ince ayrım yapılmak üzere makine öğrenmesi teknikleri uygulanmıştır. KAB ile ince ayrımlı anlamları belirlemek daha zor bir işlemdir. Durum Tablo 1'de verilen sonuçlarda da görülmektedir.

İlerleyen kısımlarda gerçekleştirilen uygulamanın adımları konusunda daha ayrıntılı bilgiler verilmektedir.

3.1 WEKA Yazılımı ve Makine Öğrenmesi Algoritmaları

Makine öğrenmesi algoritmalarından faydalanmak amacıyla Yeni Zelanda'daki Waikato Üniversitesi'nin Java programlama dilini kullanarak hazırladığı ve GNU lisansı ile özgür yazılım olarak kullanıma sunduğu WEKA yazılımından faydalanılmıştır [4].

İlk iş olarak, Semeval-2007 verilerinin dosya formatı WEKA yazılımı tarafından kullanılan ARFF formatında olmadığı için verilerin istenen biçime dönüşümü gerekmiştir. Bu amaçla Python programlama dili kullanılarak eldeki verileri ARFF formatına dönüştüren bir çevirici yazılım geliştirilmiştir. Verilerin çevrimi esnasında formatlar arası uyumsuzluklar tek tek bulunarak giderilmiştir. Örneğin, ARFF dosya formatında her bir özelliğin alabileceği tüm değerlerin en başta bir liste şeklinde yazılması gerekmektedir. Bu sebeple önce Semevel-2007 verisi içinde kullanılan tüm özellikler ve bu özellikler için kullanılan tüm değerlerin bulunması gerekmiştir. Benzer şekilde ARFF formatında birden fazla kelimeden oluşan değerlerin tek tırnak içinde bildirilmesi gerektiği için bu şekilde verilmiş olan değerler de bulunarak ARFF dosyasına aktarılırken gerekli düzeltme işlemi yapılmıştır. Ayrıca, veri dosyalarının barındırdığı fazladan/istenmeyen bir takım karakterler ayrıştırma işlemleri konusunda zaman alıcı problemler olmuştur.

İkinci iş olarak, kullanılacak makine öğrenmesi algoritmalarına karar verilmiştir. Semeval-2007'ye katılan çalışmada kullanılan istatistiki tekniğin Naive Bayes yaklaşımına benzemesi sebebi ile ilk olarak Naive Bayes seçilmiştir. Daha sonra çok karmaşık olmayan karar ağacı algoritması seçilmiş ve öğrenme verileri üzerinde öğrenme işlemi gerçekleştirilerek modeller kaydedilmiştir.

Son olarak, öğrenilen modeller ile test verileri üzerinde gerekli işlemler yapılmış ve Tablo 2 ve 3'te verilen değerler elde edilmiştir.

Tür	İnce ayrım	
	P	R
İsim	0.6152	0.5823
Fiil	0.4625	0.4379
Diğer	0.56	0.5843
Ortalama	0.54623	0.58484

Tablo 2. Naive Bayes Algortiması Kullanılarak Elde Edilen Sonuçlar

Tür	İnce ayrım	
	P	R
İsim	0.5498	0.5898
Fiil	0.418	0.5348
Diğer	0.5143	0.608
Ortalama	0.494	0.57753

Tablo 3. Karar Ağacı Algoritması Kullanılarak Elde Edilen Sonuçlar

4. Değerlendirme

Gerçekleştirilen uygulama aracılığıyla (Tablo 2 ve 3'ten de) görüldüğü üzere, makine öğrenmesi teknikleri ile Semeval-2007'de elde edilenlerden daha başarılı sonuçlar elde edilmiştir. Durum, Senseval/Semeval çalıştaylarında ortaya çıkan tabloya paraleldir. Yani makine öğrenmesi teknikleri KAB amacıyla uygulandığında başarılı sonuçlar üretmektedirler.

5. Gelecekte Yapılması Planlanan Çalışmalar

Bu bildiri ile sunulan çalışmanın bir dizi çalışmanın başlangıcı olması planlanmaktadır. Öncelikle, Semevel-2007 'de Türkçe için elde edilen sonuçların başarımının düşük olmasının bir sebebi olarak gösterilen örnekleme yetersizliğinin giderilmesi düşünülmektedir. Bu amaçla, o tarihlerde varolmayan Türkçe Ulusal Derlemi kullanılabileceği değerlendirilmektedir [2]. Benzer şekilde, kullanılan özelliklerin her ne kadar daha önceki çalışmalarla elde edilen sonuçlara göre seçildiği söylenmiş olsa da yeni örneklemelerle birlikte özellikler eklenmesi ve/

veya varolanların bazılarının çıkarılması ile elde edilen sonuçların değişip değişmediğinin görülmesi de düşünülmektedir. Çünkü daha önce yapılan KAB çalışmaları göstermiştir ki her özellik her durumda fayda sağlamadığı gibi bazı durumlarda negatif etki de yapabilmektedir [5].

6. Kaynaklar

[1] Agirre E., Edmonds P., Introduction, in Agirre and Edmonds (eds.) Word Sense Disambiguation: Algorithms and Applications, 1-28, Springer, (2006)

[2] Aksan, Y. et al., Construction of the Turkish National Corpus (TNC), In Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC 2012), İstanbul, Turkiye, (2012)

[3] Ide, N., Veronis, J., "Word Sense Disambiguation: The State of the Art", Computational Linguistics, (1998)

[4] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, Ian H. Witten, The WEKA Data Mining Software: An Update; SIGKDD Explorations, Volume 11, Issue 1, (2009)

[5] Mihalcea, R., Instance Based Learning with Automatic Feature Selection Applied to Word Sense Disambiguation, COLING '02 Proceedings of the 19th international conference on Computational linguistics - Volume 1, Pages 1-7, (2002)

[6] Oflazer, K., Say, B., Tur, D. Z. H. and Tur, G., Building A Turkish Treebank, Invited Chapter In Building And Exploiting Syntactically-Annotated Corpora, Anne Abeille Editor, Kluwer Academic Publishers, (2003)

[7] Orhan Z. and Altan Z.. Effective Features for Disambiguation of Turkish Verbs, IEC'05, Prague, Czech Republic: 182-186, (2005)

[8] Orhan, Z. and Altan, Z. Impact of Feature Selection for Corpus-Based WSD in Turkish, LNAI,Springer-Verlag, Vol. 4293: 868-878, (2006)

[9] Orhan, Z., Çelik, E., Demirgüç, N.,SemEval-2007 Task 12: Turkish Lexical Sample Task, Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007), Prague, Czech Republic, Association for Computational Linguistics, 59-63, (June, 2007,)

[10] Stevenson, M., Word Sense Disambiguation: The case for combinations of knowledge sources, CSLI Publications, (2003)