Bürünsel, Sözcüksel ve Biçimbilgisel Bilgiyi Kullanan Eş-Eğitim ile Türkçe Konuşma Dilinin Otomatik Cümle Bölütlemesi

Dr. Doğan Dalva

F.M.V. Işık Üniversitesi, Fen Bilimleri Enstitüsü. Tez Danışmanları: Doç. Dr. Ümit Güz, Doç. Dr. Hakan Gürkan.

Özet

Cümle bölütleme işlevi, standart Otomatik Konuşma Tanıma sistemlerinin çıkışından elde edilen işlenmemiş kelime dizisi biçimindeki veriyi cümlelere ayırarak zenginleştirmeyi amaçlayan bir işlemdir. Cümle bölütlemede kullanılan standart yöntemler, model eğitimi aşamasında oldukça fazla miktarda el ile etiketlenmiş veriye ihtiyaç duyar. Bu çalışmada çok bakışlı yarı öğreticili yöntemler geliştirerek, daha az el ile etiketlenmiş veri ile standart yöntemlere göre daha yüksek başarımın sağlanması hedeflenmektedir.

Giriş

Otomatik Konuşma Tanıma sistemlerinin çıkışında, algılanan kelimeler herhangi bir noktalama işareti olmadan ham bir kelime listesi halinde elde edilmektedir. Elde edilen veri, küçük mesajlar için kullanışlı olabilir ancak uzun konuşmalar göz önünde bulundurulduğunda elde edilen verinin anlaşılması ve işlenmesi oldukça zordur. Ek olarak, Doğal Dil İşleme sistemleri, cümle bölütlemesi yapılmış veri tabanına ihtiyaç duyar.

Otomatik cümle bölütlemenin amacı her bir kelimenin ardında cümle sınırının olup olmadığını hipotez etmektir. Bu bağlamda, otomatik cümle bölütleme problemi ikili sınıf sınıflandırma problemi olarak ele alınabilir. Sözcüksel (lexical) bilgi cümle sonlarına ilişkin önemli ipuçları vermektedir. Ancak, sözcüksel bilginin bürünsel (prosody) ve biçimbilgisel (morphology) bilgi ile beraber kullanılması otomatik cümle bölütlemenin daha etkili yapılmasını sağladığı gösterilmiştir.

Türkçe dili için yapılmış önceki çalışmalarda bürünsel özellikler açık kaynaklı programlar kullanılarak çıkarılmış ve duraksama süreleri, enerji ve formant frekanslarından oluşan farklı bürünsel özellik gruplarının cümle bölütlemedeki başarımları kıyaslanmıştır. İlave olarak Türkçe dili için biçimbilgisel ve sözcüksel özellikler de açık kaynaklı yazılımlar kullanılarak çıkarılmış, bürünsel bilgi ile cümle bölütlemedeki başarımı kıyaslanmıştır.

Yöntem

Yarı-öğreticili öğrenme yöntemlerinin çalışma prensibi eğitim kümesini oldukça az miktarda el ile etiketlenmiş veri kümesi (in-domain labeled data) ve etiketlenmemiş veri kümesine (out-of domain unlabeled data) ayırmak ve kademeli olarak etiketlenmemiş veri kümesindeki verilerin sınıflarını hipotez ederek el ile etiketlenmiş veri kümesine aktarmaktır. Kendi kendine eğitim (self-training) ve eş eğitim (co-training) yarı-öğreticili öğrenme teknikleridir.

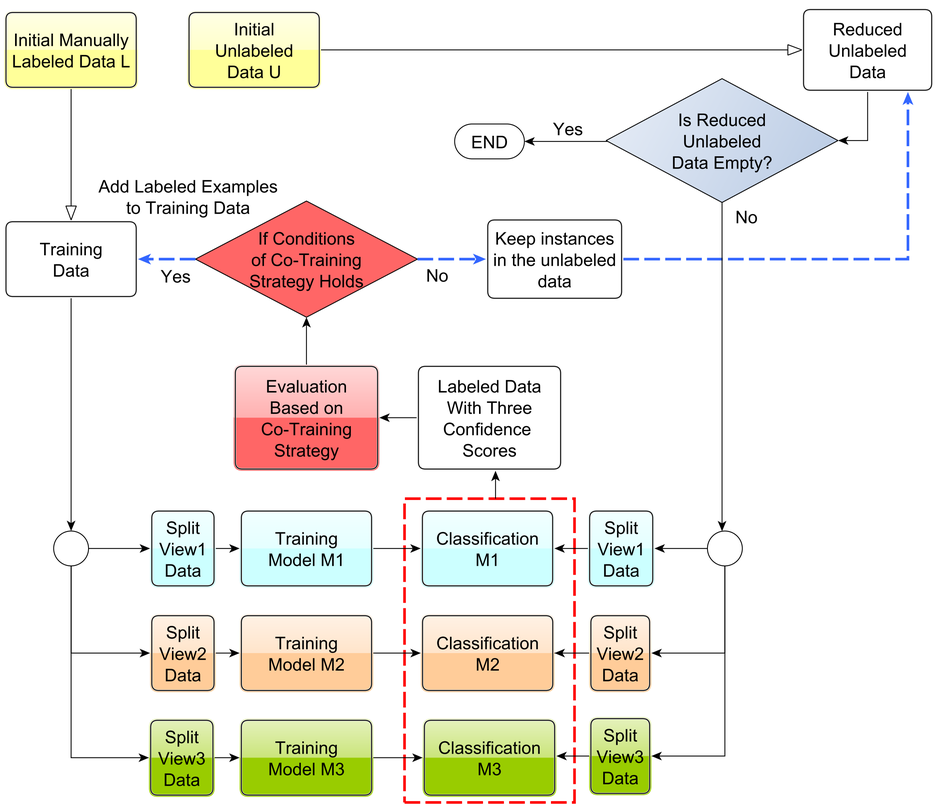
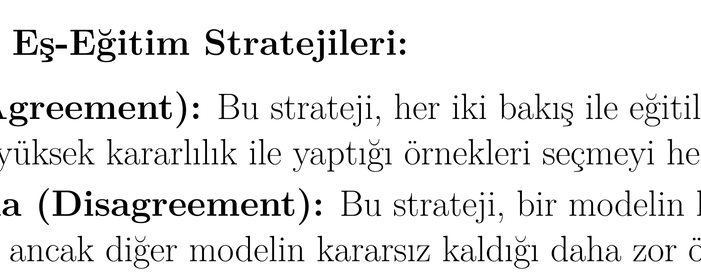


Figure: Önerilen Yarı Öğreticili Öğrenme Yöntemi

İki bakışlı Eş-Eğitim Stratejileri:

- **Uzlaşma (Agreement):** Bu strateji, her iki bakış ile eğitilen modelin de aynı hipotezi yüksek kararlılık ile yaptığı örnekleri seçmeyi hedefler.
- **Uzlaşamama (Disagreement):** Bu strateji, bir modelin kararlı bir şekilde hipotez ettiği, ancak diğer modelin kararsız kaldığı daha zor örnekleri seçmeyi hedefler.
- **Self-Combined:** Bu strateji birinci aşamada her iki modelin ayrı ayrı en kararlı hipotez ettiği örnekleri seçmesine izin verir. Birinci aşamada seçilen örnekler arasından her iki modelin de aynı hipotezde bulunduğu örnekler ikinci aşamada seçilen örnekleri oluşturur. [1]

Önerilen Üç-bakışlı Eş-Eğitim ve Kurul Tabanlı Stratejiler: Uzlaşma, Uzlaşamama ve Self-Combined stratejilerinin aşağıdaki tablo ve figürde gösterildiği biçimde geliştirilmesi veya birleştirilmesi ile üç bakışlı eş eğitim ve kurul tabanlı stratejiler geliştirilmiştir. [2]



(a) Yapı 1 (b) Yapı 2

Figure: Önerilen Üç-Bakışlı Eş-Eğitim ve Kurul Tabanlı Stratejilerin Yapıları

Table: Önerilen Üç-Bakışlı Eş-Eğitim Stratejileri (Strateji 1 - 7), Kurul Tabanlı Stratejiler (Strateji 8 ve 9) ve Yapıları

Strateji	Yapı	Level 1	Level 2
Strateji 1	Yapı 2	Agreement	-
Strateji 2	Yapı 1	Agreement	Disagreement
Strateji 3	Yapı 2	Self-Combined	-
Strateji 4	Yapı 1	Agreement	Self-Combined
Strateji 5	Yapı 1	Self-Combined	Disagreement
Strateji 6	Yapı 1	Disagreement	Self-Combined
Strateji 7	Yapı 1	Self-Combined	Agreement
Strateji 8	Yapı 2	Kurul Tabanlı Öğrenme	
Strateji 9	Stratejiler arası Kurul Tabanlı Öğrenme (Strateji 2,34,5,6,8)		

Sonuç

Bu çalışmada Amerika'nın Sesi Türkçe Haber Yayınına ilişkin veri setleri kullanıldı. Kullanılan veri tabanında toplam 104458 kelime ve 6881 cümle sınırı bulunmaktadır. Toplam veri setinin %60ı eğitim kümesini, %20si geliştirim kümesini ve %20si test kümesini oluşturmaktadır.

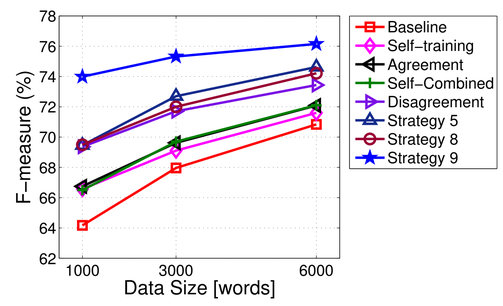


Figure: El ile Etiketlenmiş 1000, 3000 ve 6000 Kelimelik Veri Bulunurken Elde Edilen Ortalama F-measure Başarımları.

Bu çalışmada cümle bölütleme için yeni çok bakışlı yarı öğreticili stratejiler önerildi. Bu stratejiler ile az sayıda el ile etiketlenmiş veri varken eski çalışmalarda önerilen iki bakışlı yöntemlere göre daha yüksek başarımların elde edildiği gösterildi.

Referanslar

[1] U. Guz, S. Cuendet, D. Hakkani-Tür, and G. Tur.
Multi-view semi-supervised learning for dialog-act segmentation of speech.
IEEE Trans. Audio, Speech, Lang. Process., 18:320–329, 2010.

[2] D. Dalva, U. Guz, and H. Gurkan.
Effective semi-supervised learning strategies for automatic sentence segmentation.
Pattern Recognition Letters, 2017.

Teşekkür

Bu çalışma TÜBİTAK ARDEB (3501 Kariyer Geliştirme Programı, Proje No: 107E182 ve 1001 Bilimsel ve Teknolojik Araştırma Projeleri Programı, Proje No: 111E228) ve Işık Üniversitesi Bilimsel Araştırma Projeleri Fonu (Proje No: 09A301 ve 14A201) tarafından desteklenmiştir.

İletişim

- Dogan Dalva: dogan.dalva@isikun.edu.tr
- Ümit Güz: umit.guz@isikun.edu.tr
- Hakan Gürkan: hakan.gurkan@btu.edu.tr