

<b>1) Intent classification</b>	<b>1</b>
<b>2) Chatbot geliştirme</b>	<b>3</b>
<b>3) FAQ forwarding</b>	<b>3</b>
<b>4) Query analysis (telaurea)</b>	<b>4</b>
<b>5) Email için şikayet, teşekkür sınıflandırma (somemto)</b>	<b>4</b>
<b>Gelecek işler:</b>	<b>4</b>

## 1) Intent classification

- Datasetler genellikle diyalog içerikli
  - Diyalogda konuşma, konu takibi yapmak intent detection'ın en zor aşamalarından biri.
  - Bu nedenle domain (telekom, alışveriş vs), dataset tipi (email, chat), intent tipleri ve intent bulunduğu ne yapılacağına net karar vermek gerekir.
  - Hazır bir dataset bulunuyor, ingilizce emailer içeriyor [9]. Metinlerin dağılımı, etiketleme bozukluğu nedeniyle sınıflandırma sonuçları beklenenden kötü.
- Intent tipleri
  - Pek çok araştırmada intent tipleri en genel olarak şu şekilde: request, commit, question, propose, request.
    - Daha farklı kategoriler de var: [4, 5]
  - Bunları action gerektirenler ve gerektirmeyenler şeklinde ayırıp ön bir sınıflandırma yapabiliriz.
- Öznitelikler
  - Speech acts
    - Speech act, dilde performatif sözleri tanımlamak için öne sürülüp geliştirilmiş bir teoridir. Bu teoriye göre, bir söz veya cümlenin öncelikle somut ve sade bir anlamı vardır (locutionary act), daha sonra kast ettiği bir anlam vardır (illocutionary act) ve son olarak sözün doğurduğu bazı sonuçlar (performative act) vardır. [6]

- Dildeki sözcük tipleri (fiiller, isimler..) ve önermeler tanımdaki aşamalara göre sınıflandırılmış, cümle anlamları bu çerçevede incelenmiş, dillerin yapısı daha açıklayıcı yollarla incelenmeye çalışılmış. Örneğin Searle'ün illocutionary act sınıflandırmasına göre beş tane speech act vardır: 1) assertive 2) directive 3) commissive 4) expressive 5) declaration [7]
- Speech act teorisinin yardımıyla intent bulma pek çok çalışmada denenmiş ve en çok da Searle'ün yaklaşımı izlenmiştir [4, 5]. Genel yaklaşım, cümledeki fiilin niteliklerini ayırtmak. Bunun için İngilizce fiiller üzerinde yapılmış ve intent recognition'da da denenmiş çalışmalar var. [2]'de yazarlar fiillerin sınıflandırılmış halde sunan birer lexicon'dan yararlanmışlar ve iyi sonuçlar almışlar. Bu çalışmada email'e yaklaşım, emailin cevap bekleme niteliği gibi konular incelenmiş ve iyi bir şema hazırlanmış, uygulamak iyi olur.
- Soruları ve speech act içeren yapıları tanımak için iyi bir POS tagger ve chunk parser gerekiyor.
- Model  
İki seçeneğimiz bulunuyor.
  - [8]'de tanıtilen neural network temelli sistemi uygulamak ve geliştirmek
    - İlk sonuçlar
    - İyi bir dataset hazırlamak gerek
  - [2] ve diğer klasik algoritmalar ve speech act yaklaşımını izleyen çalışmaların biri veya hibridini uygulamak, değiştirmek, geliştirmek
    - İyi bir dataset gerekir
    - Fiilleri türkçeye çevirmek gerekir
- Zaman planı
  - Önce modele karar verelim (0.5 gün)
    - Cümle veya doküman bazlı metin sınıflandırma
    - Emaili gönderen hangi tarafları inceleyeceğimizi belirleme
    - Email tarihçesi için bir hat belirleme, zincir emailere bakıp bakmayacağımıza karar vermeliyiz
  - Kategorileri net bir şekilde belirleyelim (request, info, commit gibi) (0.5 gün)
  - İyi bir dataset oluşturalım, gerekliyse etiketleyelim (5 gün)
  - Uygulama (5 gün)
  - Sonuçları inceleme (1 gün)

#### Kaynaklar

[1] Cohen, William W., Vitor R. Carvalho, and Tom M. Mitchell. "Learning to Classify Email into" Speech Acts". *EMNLP*. Vol. 4. 2004.

- [2] Goldstein, Jade, and Roberta Evans Sabin. "Using speech acts to categorize email and identify email genres." *System Sciences, 2006. HICSS'06. Proceedings of the 39th Annual Hawaii International Conference on*. Vol. 3. IEEE, 2006.
- [3] Ulrich, Jan, et al. "Regression-Based Summarization of Email Conversations." *Icwsn*. 2009.
- [4] Jeong, Minwoo, Chin-Yew Lin, and Gary Geunbae Lee. "Semi-supervised speech act recognition in emails and forums." *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 3-Volume 3*. Association for Computational Linguistics, 2009.
- [5] Dabbish, Laura A., et al. "Understanding email use: predicting action on a message." *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, 2005.
- [6] [https://en.wikipedia.org/wiki/Speech\\_act](https://en.wikipedia.org/wiki/Speech_act)
- [7] Searle, John R. "A taxonomy of illocutionary acts." (1975).
- [8] <https://chatbotsmagazine.com/contextual-chat-bots-with-tensorflow-4391749d0077>
- [9] Ulrich, Jan, Gabriel Murray, and Giuseppe Carenini. "A publicly available annotated corpus for supervised email summarization." *Proc. of AAIL Email-2008 Workshop, Chicago, USA*. 2008.

## 2) Chatbot geliştirme

- Intent classification modeli belirleyici olacak
- İki yaklaşım denemeliyiz:
  - Basit bir neural model
  - Basit bir kural set oluşturup klasik, kural tabanlı veya hibrid bir sistem oluşturmak
- İlk dataseti netleştirmek zorundayız
- Kapsam belirlenmeli
- Araştırma, basit sistemle ilk denemeler ve dataset oluşturma, sistemi kurma, birer hafta olacaktır; en az üç haftaya ihtiyacımız olacak.

## 3) FAQ forwarding

- Intent classification'ın belirleyici olabileceği bir iş olacak
- Daha önce geliştirdiğimiz sorguda anahtar kelimeyi bulan sistemi kullanabiliriz
- Keyword extraction ve text similarity uygulamak ilk aşamada işe yarar
- İki hafta çalışmak gerekir.

#### 4) Query analysis (telaura)

- İlk çalışmada, query'leri bir chunk parser kullanarak phrase'lere ayrıştırdık, isim yapılarını ürün, sıfat yapılarını ürünün özelliği kabul ettik. Yaklaşık 650 query içeren bir dataset verilmişti; bu dataset üzerinde sistemin precision ölçümüyle performansı ~%50, çok iyi değil. Sonuçları incelediğimizde beklenen yakalamaları yapamadığımızı gördük.
- Duckling'le aldığımız sonuçları inceleyelim.
- İki sistemi birleştirip birleştiremeyeceğimize karar vermeliyiz
- Ürün isimleri, özelliklerini ayrıntılı içeren daha iyi bir dataset ve phrase'lere skor atayabilecek bir sistemle daha iyi sonuçlar alabiliriz.
- Bir haftalık bir çalışma gerekiyor.

#### 5) Email için şikayet, teşekkür sınıflandırma (somento)

- 1000 metinlik bir veriset etiketlenmiş.
  - Beş kategori: şikayet, teşekkür, öneri, spam, diğer
- Beş kategorilik sette ilk sınıflandırma sonuçları, accuracy: 0.82 ve f-score: 0.48
  - Öneri olarak etiketlenmiş sadece 17 metin var, sistem öneri bulamıyor, bu da performansı azaltıyor.
- İki kategoriye düşürünce (şikayet / diğerleri) performansın 0.85'e çıktığını gördük.
- 37K'lık bir email veri seti var. 1000-2000 etiketlenmiş verisetiyle eğitilmiş modelin bu büyük seti etiketlemesini sağlayacağız, böylece train setimizi artırmış olacağız.
- Bir gün sürebilir.

#### Gelecek işler:

- Sentiment analysis
  - Neural model deneyelim
  - Aspect based sentiment analysis (cümlede neye karşı ne tür sentiment var) için bir yapı oluşturalım. Keyword extraction, intent classification işlerinden faydalanacağız.
  - Metin yapısına, phrase, cümle, doküman, karar verelim. Bu yapılarla verisetleri bulalım, oluşturalım.
- POS tagging ve chunk parsing
  - Her çalışmada gereken ve iyi çalışması sistemi etkileyen araçlar
  - Türkçede bu araçlarda eksikiz ve tamamlamamız gerek.
    - Syntaxnet kurulumu

- Nltk'deki chunk parser modelini türkçe treebank ile eğiterek türkçe chunker elde edebiliriz
  - Treebank'te lisans kısıtı var
- İngilizce için iyi çalışan araçlar var
- Fransızca için araştırmalıyız
- Spacy kullanımı ve uyarlama
  - İngilizce ve fransızca araçları dikkatle deneyelim, ölçelim
  - Türkçe için uyarlama düşünelim veya kendi sistemimizi bu yapıyla uyumlulaştırabiliriz.