New Domain, Major Effort?
How Much Data is Necessary to Adapt a Temporal Tagger to the Voice Assistant Domain

Touhidul Alam, Alessandra Zarcone & Sebastian Padó

Fraunhofer IIS
University of Stuttgart
Institute for Natural Language Processing
SPEAKER

Tagging temporal expressions (TEs)

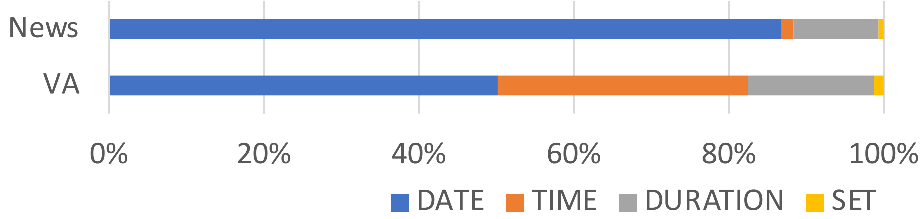
- identification and classification of TEs into types (**TE recognition**) and their conversion into a machine-readable value (**TE normalization**)
- typical tagset: TimeML/TIMEX3 (Pustejovsky et al. 2003)
- existing work considers news, social media, narrative or clinical domains, but TEs are crucial for the Voice Assistant (VA) domain

Book the room for tomorrow from 9 am, for 2 hours

TE	type	unit	normalized value
tomorrow	DATE	day	2021-06-20
9 am	TIME	hour	2021-06-20T9:00
2 hours	DURATION	hour	PT2H
empty tag	TIME	hour	2021-06-20T11:00

Scarcity of data for VA: Can we adopt a transfer learning approach?
How much data is necessary until performance flattens out?

News Datasets	VA Datasets
TE-3 (TBAQ+Silver, 800k tokens), TE-3 Simplified (290k), TE-3 Platinum (7k)	Snips (9,6k), PATE (5,6K)
• long, grammatical sentences • reference to past events • references between events	• short, concise, elliptical queries • reference to future events • fewer references between events



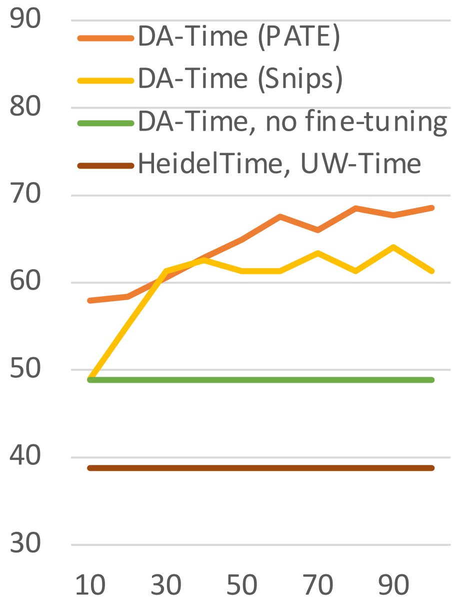
DA-Time - a hybrid temporal tagger for the VA domain

- **neural TE recognizer** (type + unit classification): DistilBERT embeddings + BiLSTM + CRF
- **rule-based TE normalizer:** based on recognizer output (type, unit) and dependency parses

Exp 1: in-domain (news: TE-3 Platinum)

- Extent comparable to other models
- Type and value worse
- DA-Time penalized (simplified training set)

Model	Extent	Type	Value
HeidelTime	90.7	83.3	78.1
UW-Time	91.4	85.4	82.4
DA-Time	90	81.1	71.3



Exp 2: out-of-domain (VA: PATE-test)

Transfer Learning (based on Felbo et al. 2017):
fine-tuning each layer sequentially (except embedding layer), freezing the other

- SOTA models perform worse out of domain (value F1 = 39)
- Even without fine-tuning, DA-Time (value F1 = 49) profits from
  - domain-specific normalizer
  - simplified news training set
- Looking at different amounts of fine-tuning data
  - improvements using Snips (after using 30% value F1 = 61)
  - best when fine-tuning on PATE-train (after using 10%, value F1 = 58, mostly for TIME)

Conclusions & Future Work

- Major improvements with only 10% of the VA data (in particular Value F1)
- Unit + type for efficient domain-specific normalization
- DA-Time as a baseline model for further neural-based research in the VA domain

TE recognizer available for academic use: https://github.com/audiolabs/DA-Time/