End-to-end information extraction without token-level supervision

Most state-of-the-art information extraction approaches rely on token-level labels to find the areas of interest in text. Unfortunately, these labels are time-consuming and costly to create, and consequently, not available for many real-life IE tasks. To make matters worse, token-level labels are usually not the desired output, but just an intermediary step. End-to-end (E2E) models, which take raw text as input and produce the desired output directly, need not depend on token-level labels. We propose an E2E model based on pointer networks, which can be trained directly on pairs of raw input and output text. We evaluate our model on the ATIS data set, MIT restaurant corpus and the MIT movie corpus and compare to neural baselines that do use token-level labels. We achieve competitive results, within a few percentage points of the baselines, showing the feasibility of E2E information extraction without the need for token-level labels. This opens up new possibilities, as for many tasks currently addressed by human extractors, raw input and output data are available, but not token-level labels.

General information
Publication status: Published
Organisations: Department of Applied Mathematics and Computer Science, Cognitive Systems, Copenhagen Center for Health Technology, Tradeshift Network Ltd. Copenhagen, University of Copenhagen
Contributors: Palm, R. B., Hovy, D., Laws, F., Winther, O.
Number of pages: 5
Publication date: 2017
Peer-reviewed: Yes

Publication information
Journal: ArXiv
Volume: 2017
Article number: 1707.04913
Original language: English
Electronic versions:
1707.04913.pdf
Source: FindIt
Source ID: 2390594786
Research output: Contribution to journal › Journal article – Annual report year: 2017 › Research › peer-review