A Sequence Modeling Approach for Structured Data Extraction from Unstructured Text

Jayati Deshmukh, Annervaz K M, Shubhashis Sengupta

Accenture Technology Labs, Bangalore, India
[jayati.deshmukh, annervaz.k.m, shubhashis.sengupta]
@accenture.com

High performance. Delivered.

Overview

Introduction

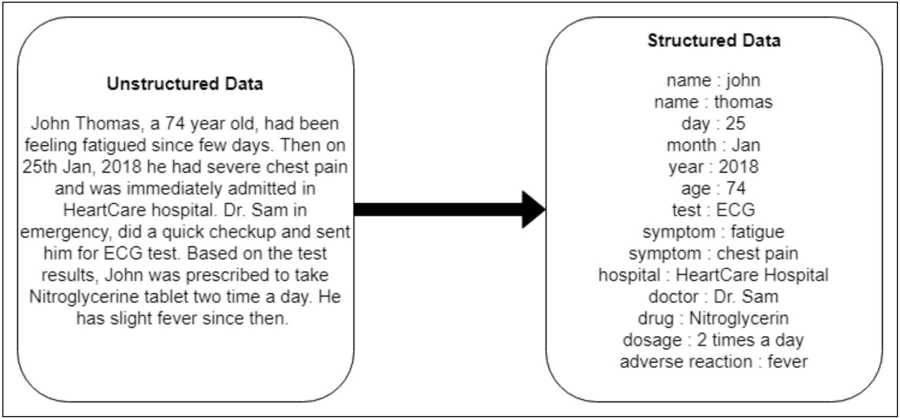
Motivation

- A lot of textual data is available in the form of documents which can be for a variety of purposes like documentation, reports and surveys, logs etc.
- Raw data is mostly useful only after extracting key information in a structured form.
- Structured data is concise, easy to store, search and retrieve for machine as well as human consumption.
- We look at the structured data extraction problem using two techniques: Seq2Seq models and sequence tagging models.

Applications

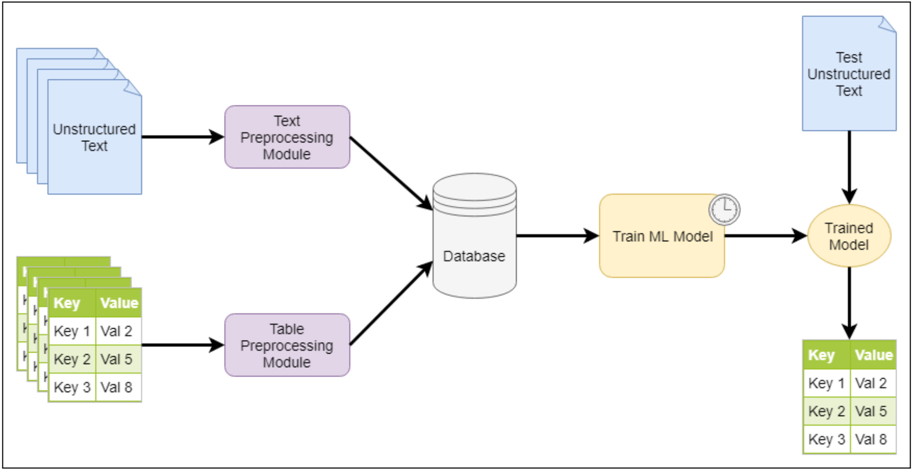
1. *Pharmacovigilance*[11] - adverse effects of prescribed drugs are reported by patients or medical practitioners in simple day to day language. This information is used to detect signals of adverse effects of drugs. Thus data has to be transformed into a structured format which is analyzed statistically for signals of adverse effects.
2. *Lease Abstraction* - largely manual inspection and validation of large commercial lease documents made for real estate deals is done by offshore experts and relevant information from the documents is extracted into a structured form. This structured information is further used for aggregate analytics and decision making by large real estate firms[1].

Example

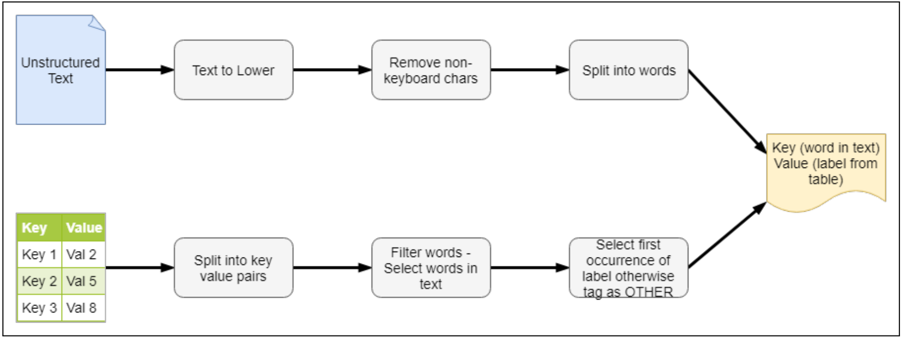

Unstructured Data
John Thomas, a 74 year old, had been feeling fatigued since few days. Then on 25th Jan, 2018 he had severe chest pain and was immediately admitted in HeartCare hospital. Dr. Sam in emergency, did a quick checkup and sent him for ECG test. Based on the test results, John was prescribed to take Nitroglycerine tablet two time a day. He has slight fever since then.
Structured Data
name : john
name : thomas
day : 25
month : Jan
year : 2018
age : 74
test : ECG
symptom : fatigue
symptom : chest pain
hospital : HeartCare Hospital
doctor : Dr. Sam
drug : Nitroglycerin
dosage : 2 times a day
adverse reaction : fever

System Diagram

accenture
High performance. Delivered.



Preprocessing and Input Generation

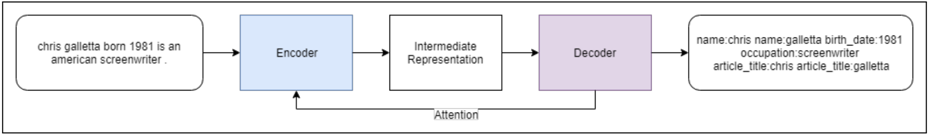


Novel Aspects

1. Use of seq2seq models for information extraction.
2. Improved variants of sequence tagging models with additional features like PoS and attention.
3. A multi-label sequence tagging model proposed.
4. Can be used for any domain where a parallel corpus of unstructured and structured data is available.
5. With the use of DL based seq2seq and sequence tagging models, this is a true machine learning based approach.

Models

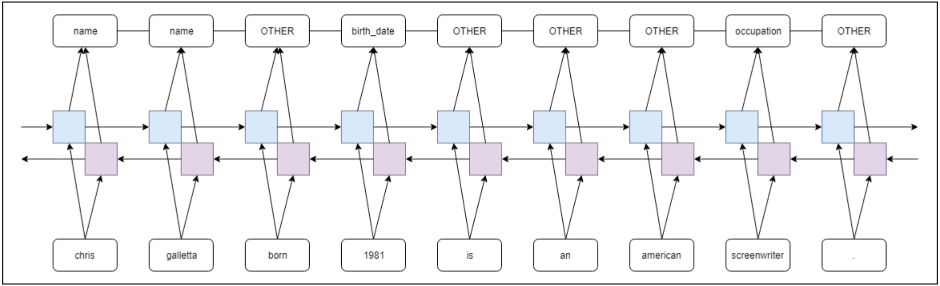
Seq2Seq Model Diagram



Seq2Seq Model

- Seq2seq models are end to end models which transform an input sequence into an output sequence.
- It consists of an encoder which takes the input and encodes it into an intermediate representation and a decoder which takes the intermediate representation as input and generates the output sequence one token at a time.
- Encoders and decoders structurally may be Recurrent Neural Networks like RNN, LSTM, GRU [3, 14]) or even Convolutional Neural Networks [7].
- seq2seq models were conceived for language translation task[3, 14]), where the input text is in one language like English and the output which is its translation, is in another language like French.

Sequence Tagging Model Diagram



Sequence Tagging Model

- Sequence tagging or labeling models tag all the tokens in the input sequence.
- It consists of recurrent neural network like RNN, LSTM, GRU and Convolutional Neural Network which reads input at token level and a conditional random field(CRF) [9] which takes as input the encoded features and generates corresponding tags for each token.
- Originally this model was tested on a variety of tasks like Part of Speech (PoS) tagging, chunking and Named Entity Recognition (NER) [8].

Approach

Baseline - Seq2seq Models

accenture
High performance. Delivered.

- Input - Sentence
- Output - String which is a series of key-value pairs corresponding to the label-word pairs of the sentence
- Experiments have been performed with different combinations of RNN and CNN encoders and decoders.

Seq2seq Model Equations

$$z = enc(x)$$

$$h_t = dec(h_{t-1}, w_t)$$

$$s_t = g(h_t)$$

$$p_t = softmax(s_t)$$

$$i_t = argmax(p_t)$$

where, at $t = 1$

$$h_0 = z$$

$$w_0 = w_{sos}$$

Sequence Tagging Models

- Sequence tagging model reads the input word by word and simultaneously generates the corresponding label for the word.
- The sentence is split in words by spaces and then each word is tagged to a corresponding label. Only the first occurrence of label of a word is considered.
- If a word does not have any label then it is labeled as 'OTHER'.

Sequence Tagging Equations

$$h_f(t) = f(U_f x(t) + W_f h_f(t-1))$$

$$h_b(t) = f(U_b x(t) + W_b h_b(t-1))$$

$$h(t) = [h_f(t) : h_b(t)]$$

$$y(t) = g(Vh(t))$$

Modified Sequence Tagging Models

- Part of Speech (PoS) tags of words are highly correlated to the corresponding labels of each word. For example, names of persons or locations are nouns. PoS tag embeddings are randomly initialized. Then, word embeddings and PoS tag embeddings are concatenated and passed as input to the bi-LSTM.
- While generating label for the current word, not all the words of the input are equally important. Words nearby to the current word are more important as compared to words farther off from the current word. Thus, every word has different importance or weight while generating the label of current word. This word level weight on the input sentence is known as self-attention.

Multi-label Sequence Tagging model

At the output layer, instead of using softmax we use sigmoid which normalizes each of the label prediction scores between 0 and 1 independently. Hamming loss is defined as the fraction of wrong labels to total number of labels.

Let y_t be the vector of true labels and y_p be the vector of independent probabilities of predicted labels. Then hamming loss (HL) is computed as follows:

$$HL = y_t \ XOR \ y_p$$

$$HL_{diff} = average(y_t * (1 - y_p) + (1 - y_t) * y_p)$$

Let a word have true labels as $[1, 0, 0, 1]$ and the model predicts the labels $[0.9, 0.1, 0.2, 0.9]$, then hamming loss in this case is computed as $avg([1, 0, 0, 1] * [0.1, 0.9, 0.8, 0.1] + [0, 1, 1, 0] * [0.9, 0.1, 0.2, 0.9])$ or $avg(0.1 + 0.1 + 0.1 + 0.2)$ or 0.125. It is a loss value, so better models have lower hamming loss.

Related Work

Related Work

Traditional Methods using Parsing and Rules

- Relationship extraction from raw text using dependency parse tree based methods [4, 13].
- Rule based methods [6, 2]

Information Extraction using Deep Learning Techniques

- Joint entity and relation extraction model [12]
- Attention based encoder-decoder model [5]

Experiments & Results

Experiment Details

- Used Wikipedia Infobox dataset [10] [1].
- It consists of total $728,321$ biographies, each having the first Wikipedia paragraph and the corresponding infobox, both of which have been tokenized.
- Given a paragraph or unstructured data, we try to generate the corresponding infobox or structured data.
- The dataset is split into three parts in the ratio 8:1:1 for train, validation and test.

[1] https://github.com/DavidGrangier/wikipedia-biography-dataset

Results

Table: Baseline Results - Seq2Seq Model

Model	Accuracy %	Perplexity
CNN Encoder Decoder	63.34	5.78
LSTM Encoder Decoder	68.42	3.95
LSTM Encoder Decoder with PoS	69.60	3.45

Table: Sequence Tagging Results

Model	Accuracy %	F1 Score %
biLSTM-CRF	79.34	65.00
biLSTM-CRF with PoS & Attention	82.82	62.32

Single-label Results

philip	mond	is	an	award-winning	dutch	film	director	and	cinematographer	.				
name	name	OTHER	OTHER	OTHER	OTHER	occupation	occupation	OTHER	occupation	OTHER				
w.	lamont	was	a	scottish	footballer	who	played	as	a	right	winger	.		
name	name	OTHER	OTHER	OTHER	OTHER	OTHER	OTHER	OTHER	OTHER	position	position	OTHER		
renan	luce	born	5	march	1980	,	paris	is	a	french	singer	and	songwriter	.
name	name	OTHER	birth_date	birth_date	birth_date	birth_place	birth_palce	OTHER	OTHER	OTHER	occupation	OTHER	occupation	OTHER

Multi-label Results

accenture
High performance. Delivered.

Table: Multi-Label Results

Word	Labels
begziin	article_title name
yavuukhulan	article_title image name
,	OTHER
1929-1982	OTHER
was	OTHER
a	OTHER
mongolian	nationality language
poet	occupation
of	OTHER
the	OTHER
communist	OTHER
era	OTHER
that	OTHER
wrote	OTHER
in	caption
mongolian	nationality language
and	OTHER
russian	language
.	OTHER

Conclusions & Future Work

Conclusions & Future Work

- Used multiple variants of sequence tagging models to extract structured data from unstructured data.
- Large publicly available dataset of Wikipedia Biographies has been used to convert the information available in paragraphs into structured format of infoboxes. However our models are generic and not dependent on the Wikipedia Infobox dataset. It should give similar results for any other similar dataset.
- Sequence tagging models further improved with additional features like PoS tags and attention.
- Multi-label sequence tagging model gave more complete results by giving multiple labels of words.

- In future we plan to experiment with other variations of the models and also try data of different domain.

References I

[1] K. M. Annervaz, Jovin George, and Shubhashis Sengupta. A generic platform to automate legal knowledge work process using machine learning. In *14th IEEE International Conference on Machine Learning and Applications, ICMLA 2015, Miami, FL, USA, December 9-11, 2015*, pages 396–401, 2015.

[2] Martin Atzmueller, Peter Kluegl, and Frank Puppe. Rule-based information extraction for structured data acquisition using textmarker. In *LWA*, pages 1–7, 2008.

[3] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014.

References II

[4] Aron Culotta and Jeffrey Sorensen. Dependency tree kernels for relation extraction. In *Proceedings of the 42nd annual meeting on association for computational linguistics*, page 423. Association for Computational Linguistics, 2004.

[5] Li Dong and Mirella Lapata. Language to logical form with neural attention. *arXiv preprint arXiv:1601.01280*, 2016.

[6] Katrin Fundel, Robert Küffner, and Ralf Zimmer. Relexrelation extraction using dependency parse trees. *Bioinformatics*, 23(3):365–371, 2006.

[7] Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N Dauphin. Convolutional sequence to sequence learning. *arXiv preprint arXiv:1705.03122*, 2017.

[8] Zhiheng Huang, Wei Xu, and Kai Yu. Bidirectional lstm-crf models for sequence tagging. *arXiv preprint arXiv:1508.01991*, 2015.

References III

[9] John Lafferty, Andrew McCallum, and Fernando CN Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. 2001.

[10] Rémi Lebret, David Grangier, and Michael Auli. Neural text generation from structured data with application to the biography domain. *arXiv preprint arXiv:1603.07771*, 2016.

[11] Anutosh Maitra, K. M. Annervaz, Tom Geo Jain, Madhura Shivaram, and Shubhashis Sengupta. A novel text analysis platform for pharmacovigilance of clinical drugs. In *Proceedings of the Complex Adaptive Systems 2014 Conference - Conquering Complexity: Challenges and Opportunities, Philadelphia, PA, USA, November 3-5, 2014*, pages 322–327, 2014.

[12] Makoto Miwa and Mohit Bansal. End-to-end relation extraction using lstms on sequences and tree structures. *arXiv preprint arXiv:1601.00770*, 2016.

References IV

[13] Frank Reichartz, Hannes Korte, and Gerhard Paass. Dependency tree kernels for relation extraction from natural language text. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 270–285. Springer, 2009.

[14] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112, 2014.

Questions?