WEEKLY REPORT

June 12, 2019

Mariama Drame
African Institute for Mathemathematical Science
African Master in Machine Intelligence

Contents

0.1 Highlights of the Week: Flair Experimentation

1. *Most significant or successful efforts of the week*
2. Dataset
3. Data loading and preprocessing
4. Flair: Introduction
5. Tried flair modules that are more appropriate for this task (Name Entity Recognition, Part of Speech tagging)
6. Discussion

0.1.1 Dataset

In this experimentation we use the 30M Factoid Question-Answer Corpus ([1] and [2]) which consists of 30M natural language questions in English and their corresponding facts in the knowledge base Freebase.

The dataset is formatted as a text file, where each line contains:

<subject> <relationship> <object> natural language question, where <subject>, <relationship> and <object> are the subject, relationship and object identifier in Freebase corresponding to the natural language question.

0.1.2 Data Loading and preprocessing

0.1.2.1 Data loading

To load the data we use Pandas. Pandas is a python open source library for data analysis. After loading the data, we just take the column containing the questions see that we don't use the answer in this work.

0.1.2.2 Data preprocessing

Most of the time text preprocessing require stop-words removal. But in our case a stop-word can be a keyword in a question (verb, pronoun,..). So we are not going to remove

them. Also we are not going to transform uppercase letter into lowercase because that can change the class of the word (e.g: from proper noun to just noun)

Instead We remove punctuation from the text and tokenize it. We use regular expressions to remove punctuation from the questions and Natural Language Toolkit to tokenize them

0.1.3 Flair: Introduction

Flair is a Natural Language Processing(NLP) library build on top of Pytorch and open sourced by Zalando Research. Flair comes with multiple NLP tasks

0.1.4 Flair: Name Entity Recognition (NER)

It can recognise whether a word represents a person, location, company or names in the text. The table below gives the NER extracted from the ten first questions in our dataset

	question	ner
0	what is the book e about ?	[]
1	in what release does the release track cardiac...	[]
2	what country is the debt from ?	[]
3	what songs have nobuo uematsu produced ?	[]
4	who produced eve-olution ?	[]
5	which artist recorded most of us are sad ?	[]
6	what movie is produced by warner bros. enterta...	[]
7	what is don graham known as ?	[]
8	what is an attraction near columbus ?	[]
9	what album was tibet released on ?	[]

By looking at the table above we see that NER is far from being sufficient to extract all the keywords in a given question. For example applying Flair NER on the question **Why did the U.S Invade Iraq?** we only get **U.S** and **Iraq** which are insufficient to understand the question. For a better understanding of this question the word **Invade** is as important as **U.S** and **Iraq**

Let's have a look at what we can extract from questions using Flair Part of Speech tagging

0.1.5 Flair: Part of Speech(PoS) Tagging

PoS is defined as the process of marking words in the text corresponding to their particular part of speech.

	question	pos
0	what is the book e about ?	what <WP> is <VBZ> the <DT> book <NN> e <NN> a...
1	in what release does the release track cardiac...	in <IN> what <WP> release <NN> does <VBZ> the ...
2	what country is the debt from ?	what <WDT> country <NN> is <VBZ> the <DT> debt...
3	what songs have nobuo uematsu produced ?	what <WDT> songs <NNS> have <VBP> nobuo <FW> u...
4	who produced eve-olution ?	who <WP> produced <VBD> eve-olution <NN> ? <.>
5	which artist recorded most of us are sad ?	which <WDT> artist <NN> recorded <VBD> most <J...
6	what movie is produced by warner bros. enterta...	what <WDT> movie <NN> is <VBZ> produced <VBN> ...
7	what is don graham known as ?	what <WP> is <VBZ> don <NN> graham <NN> known ...
8	what is an attraction near columbus ?	what <WP> is <VBZ> an <DT> attraction <NN> nea...
9	what album was tibet released on ?	what <WP> album <NN> was <VBD> tibet <NN> rele...

Flair define the PoS of every single word in the text based on its definition and its context. So to extract meaningful words from questions we need to choose the classes to which the keywords of a question can belong

Define those classes is not obvious because a class can contains some words that are import for the understanding of a question and other that are not at all. To understand the different classes, we use [3] and select the most relevant classes for this work. Below, the list of classes we chose:

- NN, NNS, NNP, NNPS : for nouns (Proper noun, plural noun,..)
- PRP$: for possessive pronoun
- JJ, JJS, JJR: for adjectives
- RB, RBR, RBS: for adverbs
- VB, VBD, VBG, VBN, VBZ : for verbs
- IN: for conjunction-subordinating (before, after, between, ..)
- CD: cardinal number
- FW: for foreign words
- MD: for modals

We combine result from the PoS with the NER result we got in the previous part. The table below gives part of the result obtained

	question	keywords
0	which country is sonia lafuente from ?	[country, is, sonia, lafuente, from]
1	what country is lazare meerson from ?	[country, is, lazare, meerson, from]
2	which country is j. neil schulman from ?	[country, is, neil, schulman, from]
3	which country is sixmile , alabama located in ?	[country, is, sixmile, alabama, in]
4	which country is colton jobke from ?	[country, is, colton, jobke, from]
5	where was the film alone in the dark filmed in ?	[wa, film, alone, in, dark, in]
6	what kind of drug is treatment set ts331975 7....	[kind, of, drug, is, treatment, set, ts331975,...
7	whats the name of a track that was done in the...	[whats, name, of, track, wa, in, artist, béla,...
8	what is the genre of the album beyond the blue ?	[is, genre, of, album, beyond, blue]
9	name a singer ?	[name, singer]

To illustrate how difficult is to choose the different classes to keep for our study, we noticed for example at index 2 the word **connect** is not in the selected keywords and yet it is the most important in this question. To solve this problem, we added the class **VB** (verb, base form) to the list of classes selected.

	question	keywords
0	how do achaeologist find remains?	[do, achaeologist, find, remains]
1	how to set up a wireless connection?	[set, wireless, connection]
2	How I connect to la guardia community college?	[connect, to, la, guardia, community]
3	how do I make yahoo my default email for docum...	[do, make, yahoo, my, default, email, for, doc...
4	what is sufism?	[is, sufism]
5	What's your fav Italian place in Palo Alto are...	[Whats, your, fav, Italian, place, in, Palo, A...
6	why do humans have pupils?	[do, humans, have, pupils]
7	ABC newsman Had Body Armor, why don't our troo...	[ABC, newsman, Had, Body, Armor, dont, our, tr...
8	How do I tell if my mail and messages have bee...	[do, tell, if, my, mail, messages, have]
9	how do you move songs from limewire onto an ip...	[do, move, songs, from, limewire, onto, ipod, ...

The previous problem is solved but we have now some words that are not necessarily important for the understanding of the question. For example we notice that the word **do** is selected anytime it appear in a question (e.g. indexes 3, 6, 8, 9)

0.1.6 Discussion

Flair is a very powerful library to extract Name Entity Recognition and Part of Speech tagging given that it takes the context of the word into account. But not all important

words in a text can be extracted using NER. Knowing that PoS also only tags each word of the text to a given class, we need to choose which classes to care about.

This choice is not obvious since words in the same class may be important for one question and not for another. So not taking a class into account can lead to the absence of keywords in a question but adding a class can also add a lot of unwanted words.

0.2 Things I Need Help with and my Plan to Resolve

Given that the result we got from Flair is pretty good, I think it will be a good idea to use it and train a NMT in order to generalize it to a QA domain:

1. Collect a huge volume of questions dataset
2. use Flair on a part of them to extract keywords
3. Now we get our data (questions) and target (keywords), we train a Neural machine translation on this dataset
4. we evaluate the model on the questions without target
5. train our clustering model to identify the type of question

0.2.1 Next step

start working on the clustering from **Sunday, 9^{th} June 2019 Possible type of questions:**

- what
- who
- how
- which
- how much, how many
- where
- why

- when
- can, could, would
- is, was, were
- define, list

Bibliography

[1] academictorrents.com/details/973fb709bdb9db6066213bbc5529482a190098ce/techhit=1&filelist=1

[2] http://arxiv.org/pdf/1603.06807.pdf

[3] https://www.clips.uantwerpen.be/pages/mbsp-tags

[4] https://github.com/zalandoresearch/flair.

[5] Introduction to Flair for NLP: https://www.analyticsvidhya.com/blog/2019/02/flair-nlp-library-python/

[6] Contextual String Embeddings for Sequence Labeling, Zalando Research