Task 1

Pos Tagging

Nixon Andhika / 13517059
Ferdy Santoso / 13517116
Jan Meyer Saragih / 13517131

Data Source

Corpus NLTK

Task Description

POS tagging merupakan task untuk memberi label tertentu untuk setiap token atau kata pada sebuah teks. Label mengindikasikan kategori grammar dari kata sehingga dapat menunjukkan hubungan antar kata pada sebuah kalimat

Flow Modul

1. Import Data
2. Pemisahan data menjadi word dan tag
3. Pembentukan internal vocabulary (dictionary)
4. Sequencing data teks dan data tag menjadi integer
5. Splitting data menjadi data latih dan data tes
6. Padding sequence menjadi ukuran sama
7. Melakukan one-hot encoding untuk data tag
8. Pembangunan dan pelatihan model
9. Penyimpanan model ke file
10. Load model
11. Melakukan POS tagging

Teknik Digunakan

1. Preprocessing: sequencing, sequence padding, one-hot encoding
2. Eksperimen: RNN, LSTM, Bidirectional LSTM
3. POS Tagging: Bidirectional LSTM

Imports

Imports

In [1]:
```
import numpy as np
import pickle
import os
from nltk.corpus import wordnet, brown, treebank, conll2000
from keras.models import Sequential, Model, load_model
from keras.layers import (
    InputLayer,
    LSTM,
    Embedding,
    TimeDistributed,
    Dense,
    Bidirectional,
    Activation,
    Dropout,
    SimpleRNN
)
from keras.preprocessing.text import Tokenizer
from keras.preprocessing.sequence import pad_sequences
from keras.optimizers import Adam
from keras.utils.np_utils import to_categorical
from keras import backend

from sklearn.model_selection import train_test_split
```

Constants

In [2]:
```
TEST_SIZE = 0.1
VAL_SIZE = 0.15
EPOCH_COUNT = 3
BATCH_SIZE = 128
```

Dataset

Dataset yang digunakan adalah dataset dari nltk corpus library. Di dataset, setiap kata telah dilabeli dengan POS Tag.

In [3]:
```
treebank_corpus = treebank.tagged_sents(tagset='universal')
brown_corpus = brown.tagged_sents(tagset='universal')
conll_corpus = conll2000.tagged_sents(tagset='universal')
tagged_sentences = treebank_corpus + brown_corpus + conll_corpus
```

Preprocessing

Separate Word and Tag

Dataset yang diimpor memiliki data berupa tuple (word, tag) sehingga perlu dipisah terlebih dahulu. Setiap sentence words (list of word) dimasukkan ke variabel X sedangkan setiap sentence

tags (list of tags) dimasukkan ke variabel Y. Selain itu, dibentuk list semua kata unik dari dataset yang disimpan dalam variabel words dan list semua tag unik yang disimpan dalam variabel tags.

In [4]:
```
X = []
Y = []

for sentence in tagged_sentences:
    words_temp = []
    tags_temp = []
    for pair in sentence:
        words_temp.append(pair[0])
        tags_temp.append(pair[1])
    X.append(words_temp)
    Y.append(tags_temp)

words = set([word.lower() for sentence in X for word in sentence])
tags = set([tag for sentence in Y for tag in sentence])
```

Tokenization

Dilakukan tokenisasi terhadap variabel X yang berisi sentence words dan variabel Y yang berisi sentence tags. Tokenisasi dilakukan dengan Tokenizer dari Keras. Dilakukan fit_on_texts untuk membentuk vocabulary index dari setiap kata.

In [5]:
```
# Tokenizing words
word_tokenizer = Tokenizer(lower=True, oov_token='<<OOV>>')
word_tokenizer.fit_on_texts(X)

# Tokenizing tags
tag_tokenizer = Tokenizer(lower=False)
tag_tokenizer.fit_on_texts(Y)
```

Text to Sequence

Hasil tokenisasi yang masih berupa kata kemudian diubah menjadi sekuens integer menggunakan texts_to_sequences. Hasil yang didapatkan adalah hasil perubahan setiap kata menjadi indeksnya pada kamus dari Tokenizer. Untuk Tokenizer yang digunakan ke tag, ditambahkan satu entry '<>' = 0 karena akan dilakukan padding dengan nilai 0.

In [6]:
```
# Words sequencing
X_sequence = word_tokenizer.texts_to_sequences(X)

# Tags sequencing
Y_sequence = tag_tokenizer.texts_to_sequences(Y)

# Adding PAD tag to dictionary
tag_tokenizer.word_index['<<PAD>>'] = 0
```

Splitting Training Data and Test Data

Dilakukan split data menjadi training data dan testing data. Didefinisikan MAX_LENGTH untuk ukuran data yang akan dimasukkan ke network. Splitting dilakukan dengan train_test_split.

In [7]:
```
X_train_, X_test_, Y_train_, Y_test_ = train_test_split(X_sequence, Y_sequence, t

# Defining input layer size
MAX_LENGTH = len(max(X_train_, key=len))
```

Sequence Padding

Karena Keras membutuhkan ukuran yang didefinisikan lebih dulu, dilakukan padding hingga MAX_LENGTH untuk menyamakan ukuran setiap data.

In [8]:
```
X_train_ = pad_sequences(X_train_, maxlen=MAX_LENGTH, padding='pre')
X_test_ = pad_sequences(X_test_, maxlen=MAX_LENGTH, padding='pre')
Y_train_ = pad_sequences(Y_train_, maxlen=MAX_LENGTH, padding='pre')
Y_test_ = pad_sequences(Y_test_, maxlen=MAX_LENGTH, padding='pre')
```

One-Hot Encoding

One-Hot Encoding dilakukan untuk merepresentasikan index tag menjadi list of bit sehingga dapat lebih dimengerti oleh model machine learning. One-Hot Encoding dilakukan menggunakan to_categorical.

In [9]:
```
Y_train_ = to_categorical(Y_train_)
```

Defining Network Architecture

Arsitektur jaringan adalah sekuensial. Untuk eksperimen, dicoba model menggunakan RNN, LSTM, dan Bidirectional LSTM. Fungsi aktivasi yang digunakan adalah softmax dan digunakan layer Dropout untuk mengurangi overfitting.

RNN

In [10]:
```
rnn_model = Sequential()
rnn_model.add(InputLayer(input_shape=(MAX_LENGTH,)))
rnn_model.add(Embedding(len(word_tokenizer.word_index)+1, 128))
rnn_model.add(SimpleRNN(len(tag_tokenizer.word_index), return_sequences=True))
rnn_model.add(Dropout(0.1))
rnn_model.add(TimeDistributed(Dense(len(tag_tokenizer.word_index))))
rnn_model.add(Activation('softmax'))
rnn_model.summary()
```

```
Model: "sequential"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
embedding (Embedding)        (None, 180, 128)          7609600
_________________________________________________________________
simple_rnn (SimpleRNN)       (None, 180, 13)           1846
_________________________________________________________________
dropout (Dropout)            (None, 180, 13)           0
_________________________________________________________________
time_distributed (TimeDistri (None, 180, 13)           182
_________________________________________________________________
activation (Activation)      (None, 180, 13)           0
=================================================================
Total params: 7,611,628
Trainable params: 7,611,628
Non-trainable params: 0
_________________________________________________________________
```

In [11]:
```
rnn_model.compile(loss='categorical_crossentropy', optimizer=Adam(0.001), metrics
rnn_model.fit(X_train_, Y_train_, batch_size=BATCH_SIZE, epochs=EPOCH_COUNT, vali
```

```
Epoch 1/3
432/432 [==============================] - 119s 276ms/step - loss: 0.4589 - acc
uracy: 0.9329 - val_loss: 0.1240 - val_accuracy: 0.9782
Epoch 2/3
432/432 [==============================] - 118s 273ms/step - loss: 0.0857 - acc
uracy: 0.9858 - val_loss: 0.0529 - val_accuracy: 0.9910
Epoch 3/3
432/432 [==============================] - 122s 281ms/step - loss: 0.0441 - acc
uracy: 0.9928 - val_loss: 0.0339 - val_accuracy: 0.9932
```

Out[11]:
```
<tensorflow.python.keras.callbacks.History at 0x24faabf6a30>
```

LSTM

In [12]:
```
lstm_model = Sequential()
lstm_model.add(InputLayer(input_shape=(MAX_LENGTH,)))
lstm_model.add(Embedding(len(word_tokenizer.word_index)+1, 128))
lstm_model.add(LSTM(256, return_sequences=True))
lstm_model.add(Dropout(0.1))
lstm_model.add(TimeDistributed(Dense(len(tag_tokenizer.word_index))))
lstm_model.add(Activation('softmax'))
lstm_model.summary()
```

```
Model: "sequential_1"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
embedding_1 (Embedding)      (None, 180, 128)          7609600
_________________________________________________________________
lstm (LSTM)                  (None, 180, 256)          394240
_________________________________________________________________
dropout_1 (Dropout)          (None, 180, 256)          0
_________________________________________________________________
time_distributed_1 (TimeDist (None, 180, 13)           3341
_________________________________________________________________
activation_1 (Activation)    (None, 180, 13)           0
=================================================================
Total params: 8,007,181
Trainable params: 8,007,181
Non-trainable params: 0
_________________________________________________________________
```

In [13]:
```
lstm_model.compile(loss='categorical_crossentropy', optimizer=Adam(0.001), metric
lstm_model.fit(X_train_, Y_train_, batch_size=BATCH_SIZE, epochs=EPOCH_COUNT, val
```

```
Epoch 1/3
432/432 [==============================] - 65s 149ms/step - loss: 0.1921 - accu
racy: 0.9512 - val_loss: 0.0280 - val_accuracy: 0.9917
Epoch 2/3
432/432 [==============================] - 64s 147ms/step - loss: 0.0196 - accu
racy: 0.9937 - val_loss: 0.0174 - val_accuracy: 0.9940
Epoch 3/3
432/432 [==============================] - 64s 147ms/step - loss: 0.0137 - accu
racy: 0.9951 - val_loss: 0.0158 - val_accuracy: 0.9944
```

Out[13]:
```
<tensorflow.python.keras.callbacks.History at 0x250c97e3c10>
```

Bidirectional LSTM

In [14]:
```
bi_lstm_model = Sequential()
bi_lstm_model.add(InputLayer(input_shape=(MAX_LENGTH,)))
bi_lstm_model.add(Embedding(len(word_tokenizer.word_index)+1, 128))
bi_lstm_model.add(Bidirectional(LSTM(256, return_sequences=True)))
bi_lstm_model.add(Dropout(0.1))
bi_lstm_model.add(TimeDistributed(Dense(len(tag_tokenizer.word_index))))
bi_lstm_model.add(Activation('softmax'))
bi_lstm_model.summary()
```

```
Model: "sequential_2"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
embedding_2 (Embedding)      (None, 180, 128)          7609600
_________________________________________________________________
bidirectional (Bidirectional (None, 180, 512)          788480
_________________________________________________________________
dropout_2 (Dropout)          (None, 180, 512)          0
_________________________________________________________________
time_distributed_2 (TimeDist (None, 180, 13)           6669
_________________________________________________________________
activation_2 (Activation)    (None, 180, 13)           0
=================================================================
Total params: 8,404,749
Trainable params: 8,404,749
Non-trainable params: 0
_________________________________________________________________
```

In [15]:
```
bi_lstm_model.compile(loss='categorical_crossentropy', optimizer=Adam(0.001), met
bi_lstm_model.fit(X_train_, Y_train_, batch_size=BATCH_SIZE, epochs=EPOCH_COUNT,
```

```
Epoch 1/3
432/432 [==============================] - 88s 204ms/step - loss: 0.1793 - accu
racy: 0.9480 - val_loss: 0.0231 - val_accuracy: 0.9933
Epoch 2/3
432/432 [==============================] - 87s 201ms/step - loss: 0.0154 - accu
racy: 0.9951 - val_loss: 0.0137 - val_accuracy: 0.9955
Epoch 3/3
432/432 [==============================] - 88s 203ms/step - loss: 0.0096 - accu
racy: 0.9968 - val_loss: 0.0123 - val_accuracy: 0.9959
```

Out[15]:
```
<tensorflow.python.keras.callbacks.History at 0x250cafb7340>
```

Analisis

Berdasarkan hasil eksperimen, terlihat bahwa arsitektur Bidirectional LSTM menghasilkan nilai akurasi yang paling tinggi. Nilai akurasi masing-masing arsitektur adalah 99.28% untuk RNN, 99.51% untuk LSTM, 99.68% untuk Bidirectional LSTM. Hal ini mungkin terjadi karena kompleksitas arsitektur yang berbeda. Arsitektur RNN memiliki jaringan yang paling sederhana dibandingan LSTM dan Bidirectional LSTM. LSTM dan Bidirectional LSTM menggunakan multiple

network layer sehingga lebih dapat memahami hubungan antar data. Bidirectional LSTM yang merupakan extension dari LSTM memiliki akurasi yang lebih baik dari LSTM karena lebih cocok untuk sequence classification problem seperti pada pelatihan model POS tagging.

Pada RNN, juga terdapat vanishing gradient problem yang dapat menyebabkan jaringan berhenti training karena nilai gradient yang sangat kecil. LSTM dan Bidirectional LSTM menggunakan identity function untuk mengatasi masalah tersebut sehingga model terus dilatih yang menyebabkan nilai akurasi mungkin meningkat.

Saving model and Tokenizer

Model yang dihasilkan kemudian disimpan ke sebuah file h5 dan Tokenizer serta MAX_LENGTH disimpan ke file pickle.

In [16]:
```
bi_lstm_model.save("model/bi_lstm_model.h5")

pickle_files = [word_tokenizer, tag_tokenizer, MAX_LENGTH]

if not os.path.exists('PickledData/'):
    os.makedirs('PickledData/')

with open('PickledData/data.pkl', 'wb') as f:
    pickle.dump(pickle_files, f)
```

Doing Pos Tagging

Test Data

In [17]:
```
test_samples = [
    ['skyrim', 'nt', 'good', 'game', 'without', 'mods', 'fact', 'might', 'pay',
    ['addictive', 'game', 'ever', 'made'],
    ['counter', 'strike', 'even', 'fight', 'highly', 'trained', 'american', 'anti
]
```

Load Model and Tokenizer

In [18]:
```
def load(path):
    with open(path, 'rb') as f:
        word2int, tag2int, MAX_LENGTH = pickle.load(f)
        return word2int, tag2int, MAX_LENGTH
```

Tagging Method

Untuk melakukan tagging, digunakan method pos_tag(). Algoritma diawali dengan me-load model

dari file h5 dan me-load variabel Tokenizer serta MAX_LENGTH dari file pickle. Input yang berupa list of list of token kemudian diubah menjadi sekuens integer menggunakan Tokenizer yang di-load. Sekuens yang didapat di-padding hingga sebesar MAX_LENGTH. Model kemudian melakukan prediksi dengan menggunakan predict().

Kamus tag yang ada pada tag_tokenizer (Tokenizer POS TAG) di-reverse sehingga key menjadi index dan value menjadi kata POS TAG. Hasil tag didapatkan dengan memanggil sequences_to_tags() dengan parameter hasil prediksi dan kamus yang telah di-reverse. sequence_to_tags() akan mengembalikan value dari key dengan key berupa indeks prediksi dengan probabilitas terbesar.

Setelah hasil tag didapatkan, dilakukan pengecekan panjang sentence asli dengan panjang sentence tag. Jika sentence tag yang didapatkan kurang panjang, dilakukan penambahan 'NOUN' di depan karena berdasarkan percobaan terdapat beberapa kasus 'NOUN' di awal sentence hilang. Setelah panjang keduanya sama, dibentuk tuple (word, tag) yang disimpan pada list result. Hasil list result kemudian dikembalikan sebagai hasil sentence yang telah di-tag.

In [19]:
```
def pos_tag(token_list):
    bi_lstm_model = load_model("model/bi_lstm_model.h5")
    word_tokenizer, tag_tokenizer, MAX_LENGTH = load('PickledData/data.pkl')

    input_sequences = word_tokenizer.texts_to_sequences(token_list)
    input_sequences = pad_sequences(input_sequences, maxlen=MAX_LENGTH, padding='
    predictions = bi_lstm_model.predict(input_sequences)

    reverse_tag_map = dict(map(reversed, tag_tokenizer.word_index.items()))
    tag_result = sequences_to_tags(predictions, reverse_tag_map)

    result = []
    for i in range(len(token_list)):
        if (len(token_list[i]) != len(tag_result[i])):
            diff = len(token_list[i]) - len(tag_result[i])
            if (diff > 0):
                for j in range(diff):
                    tag_result[i].insert(0, 'NOUN')
        result.append(list(zip(token_list[i], tag_result[i])))

    return result

def sequences_to_tags(predictions, tag_map):
    tag_result = []
    for prediction in predictions:
        not_padding = False
        tag_list = []
        for index in prediction:
            tag = tag_map[np.argmax(index)]
            if (tag != "<<PAD>>"):
                not_padding = True
            if (not_padding):
                tag_list.append(tag)

        tag_result.append(tag_list)

    return tag_result
```

Doing Pos Tag

In [20]:
```
result = pos_tag(test_samples)
for res in result:
    print(res)
```

```
[('skyrim', 'NOUN'), ('nt', 'NOUN'), ('good', 'ADJ'), ('game', 'NOUN'), ('witho
ut', 'ADP'), ('mods', 'NOUN'), ('fact', 'NOUN'), ('might', 'VERB'), ('pay', 'VE
RB'), ('mods', 'NOUN'), ('make', 'VERB'), ('bugthesda', 'NOUN'), ('s', 'NOUN'),
('game', 'NOUN'), ('playable', 'ADJ'), ('rubbish', 'NOUN')]
[('addictive', 'NOUN'), ('game', 'NOUN'), ('ever', 'ADV'), ('made', 'VERB')]
[('counter', 'NOUN'), ('strike', 'NOUN'), ('even', 'NOUN'), ('fight', 'NOUN'),
('highly', 'NOUN'), ('trained', 'ADV'), ('american', 'VERB'), ('antiterrorist',
'ADV'), ('team', 'VERB'), ('using', 'ADJ'), ('latest', 'NOUN'), ('military', 'N
OUN'), ('technology', 'NOUN'), ('battle', 'ADJ'), ('group', 'NOUN'), ('really',
'NOUN'), ('madmen', 'NOUN'), ('possessing', 'NOUN'), ('crude', 'ADV'), ('bomb',
'ADJ'), ('surplus', 'NOUN'), ('ussr', 'ADJ'), ('s', 'NOUN'), ('army', 'NOUN'),
('supplies', 'NOUN'), ('despite', 'NOUN'), ('training', 'NOUN'), ('technology',
'NOUN'), ('terrorists', 'ADP'), ('still', 'NOUN'), ('good', 'NOUN'), ('chance',
'NOUN'), ('blowing', 'ADV'), ('market', 'ADJ'), ('therefore', 'NOUN'), ('much',
'NOUN'), ('like', 'NOUN'), ('real', 'ADV'), ('life', 'ADJ'), ('game', 'ADJ'),
('currently', 'ADJ'), ('full', 'NOUN'), ('hackers', 'NOUN'), ('fly', 'ADV'),
('top', 'ADJ'), ('map', 'NOUN'), ('unless', 'NOUN'), ('hack', 'ADJ'), ('like',
'NOUN'), ('getting', 'ADP'), ('aerial', 'NOUN'), ('teabag', 'ADJ'), ('please',
'NOUN'), ('play', 'ADJ'), ('better', 'NOUN'), ('counter', 'VERB'), ('strike',
'VERB'), ('sauce', 'ADJ'), ('counter', 'NOUN'), ('strike', 'NOUN'), ('go', 'VER
B'), ('game', 'NOUN'), ('game', 'NOUN'), ('day', 'VERB'), ('exists', 'NOUN'),
('historical', 'NOUN'), ('purposes', 'NOUN'), ('remember', 'VERB'), ('times',
'ADJ'), ('internet', 'NOUN'), ('cafe', 'VERB'), ('mosque', 'NOUN'), ('full', 'N
OUN'), ('game', 'NOUN')]
```