

MEF UNIVERSITY

**DEVELOPING A RELATION EXTRACTION MODEL
FROM ENGLISH WIKIPEDIA ARTICLES FOR
INFORMATION BOXES**

Capstone Project

Kaan Karabal

İSTANBUL, 2018

MEF UNIVERSITY

**DEVELOPING A RELATION EXTRACTION MODEL
FROM ENGLISH WIKIPEDIA ARTICLES FOR
INFORMATION BOXES**

Capstone Project

Kaan Karabal

Advisor: Asst. Prof. Şeniz Demir

İSTANBUL, 2018

MEF UNIVERSITY

Name of the project: Developing a Relation Extraction model from English Wikipedia articles for Information Boxes

Name/Last Name of the Student: Kaan Karabal

Date of Thesis Defense: 26/12/2018

I hereby state that the graduation project prepared by Your Name (Title Format) has been completed under my supervision. I accept this work as a "Graduation Project".

26/12/2018

Advisor's Name (Asst. Prof. Şeniz Demir)

Şeniz

I hereby state that I have examined this graduation project by Your Name (Title Format) which is accepted by his supervisor. This work is acceptable as a graduation project and the student is eligible to take the graduation project examination.

26/12/2018

Director
of
Big Data Analytics Program

We hereby state that we have held the graduation examination of _____ and agree that the student has satisfied all requirements.

THE EXAMINATION COMMITTEE

Committee Member

Signature

1. Your Advisor's Name

Asst. Prof. Dr. Şeniz DEMİR

Şeniz

2.

.....

Academic Honesty Pledge

I promise not to collaborate with anyone, not to seek or accept any outside help, and not to give any help to others.

I understand that all resources in print or on the web must be explicitly cited.

In keeping with MEF University's ideals, I pledge that this work is my own and that I have neither given nor received inappropriate assistance in preparing it.

Name	Date	Signature
Kaan Karabal	26.12.2018	

EXECUTIVE SUMMARY

Developing a Relation Extraction model from English Wikipedia articles for Information Boxes

Kaan Karabal

Advisor: Asst. Prof. Şeniz Demir

December, 2018, 30 pages

With the beginning of internet era in the mid '90s, the source of knowledge started to change rapidly from written sources such as the well-known Encyclopedia Britannica to online sources. One of the first online sources of the free encyclopedia was called Interpedia¹ initiated by Rick Gates on October 25, 1993. The idea behind Interpedia would let anyone contribute by writing articles and submitting to the database to be published which is very similar with "Wikipedia", today's most used online & free encyclopedia.

As the volume of online sources increased rapidly, different techniques are needed to provide accurate information extracted from these sources. Information Extraction & Relation Extraction terms have emerged as a result of this need.

This paper's aim is to extract relations between the entities of the first sentences of English Wikipedia Biography pages and their information boxes. First sentences mainly consist of the birth date, birthplace and death date of a famous person. In order to train a learning model, Conditional Random Fields method is used and Natural Language Processing (NLP) techniques are utilized to prepare our dataset for the learning model. Below steps show the general flow of study that this thesis focuses on;

- Parsing of Wikipedia biography dumps by Stanford CoreNLP
- Research on sequential text processing methods

¹<http://www.wikizero.net/index.php?q=aHR0cHM6Ly9lbi53aWtpcGVkaWEub3JnL3dpa2kvSW50ZXJwZWRpYQ>

- Data preparation for relation extraction
- Relation extraction through conditional random fields methodology

To be able to run CRF on English Wikipedia dumps, parsed Wikipedia dumps of biographies by Stanford Core NLP [7] study are used. Data cleaning, sample selection, labeling and feature extraction of the data are performed based on this output. These steps require the most effort to prepare the required data for CRF model.

This research work analyzes the prepared data through a supervised classification methodology. Main parameters needed for this methodology to be able to train successful CRF model are labeling of target values and defining features in parallel with labeling process.

Key Words: Relation Extraction, Information Extraction, Natural Language Processing, Conditional Random Fields, Wikipedia, Content Extraction, Sequential Text Processing

ÖZET

İngilizce Wikipedia Makalelerinden Bilgi Kutuları için Anlamsal Çıkarım Modeli
Geliştirilmesi

Kaan Karabal

Tez Danışmanı: Asst. Prof. Şeniz Demir

Aralık, 2018, 30 sayfa

İnternet çağının başlangıcı olarak adlandırılabilen 90'lı yılların başlangıcında o dönemin en yaygın ve güvenilir bilgi kaynağı olarak kabul edilen Britannica Ansiklopedileri yerini internet ortamındaki kaynaklara bırakmaya başlamıştır. Özgür ansiklopediler olarak tanımlanan bu ansiklopedilere verilebilecek ilk örnek, 25 Ekim 1993 tarihinde Rick Gates tarafından yaratılan "Interpedia"dır. Interpedia'nın kuruluş felsefesi denebilecek ana fikir, günümüzün en çok kullanılan ücretsiz internet ansiklopedisi Wikipedia'nın da üzerinde kurgulandığı temeller ile de benzerlikler göstermektedir. Bu temeller genel manada tüm kullanıcıların içerik girebilmesine ve veri tabanında kayıt bırakabilmesine izin veren kurgudur.

İnternet üzerinde bulunan ücretsiz kaynaklarının sayısının hızla artması ile birlikte bu kaynaklar üzerinden elde edilen bilgilerin doğruluklarının teyit edilebilmesi için de farklı tekniklere ihtiyaç duyulmaya başlanmıştır. Son zamanlarda sıkça duyulan Bilgi Çıkarımı (Information Extraction), İlişki Çıkarımı (Relation Extraction) ve Şartlı Rastgele Alanlar (Conditional Random Fields) terimleri bu ihtiyaç sonucu oluşmuş kavramlardır.

Bu tezin amacı Bilgi Çıkarımı, İlişki Çıkarımı ve Şartlı Rastgele Alanlar İngilizce Wikipedia Biyografi makalelerinin ilk cümleleri içerisindeki varlıklar ile bilgi kutusu (information box) içerisinde etiketli olan varlıklar arasındaki ilişkiyi çıkarmaktır. Biyografilere ait makalelerdeki ilk cümleler genel olarak makalenin yazıldığı kişiye ait doğum tarihi, doğum yer ve ölüm tarihi gibi bilgileri içermektedir. Bu kişilerin sayfalarına ait bilgi kutularında da ilgili bilgilerin etiketlenmiş olarak olması beklenmektedir. Bu hipotez

üzerinden makale kapsamında Doğal Dil İşleme (Natural Language Processing) ve Şartlı Rastgele Alanlar teknikleri kullanılarak geliştirilecek gözetimli makine öğrenmesi metodu sayesinde ilgili alanlar arasındaki ilişkinin çıkarılması hedeflenmiştir. Bu kapsamda genel çalışma akışı aşağıdaki şekilde oluşturulmuştur;

- İngilizce Biyografi makaleleri içeriklerinin Stanford CoreNLP doğal dil işleme aracı kullanılarak ayrıştırılması
- Sıralı metin işleme yöntemlerinin araştırılması
- İlişki çıkarımına yönelik veri hazırlığının yapılması
- Şartlı rastgele alanlar metodu kullanılarak varlıklar arası ilişkinin çıkarılması

Şartlı rastgele alanlar metoduna ilişkin modelin geliştirilebilmesi için öncelikle ilgili Wikipedia alanları Stanford CoreNLP aracı kullanılarak ayrıştırılmıştır. Bu ayrıştırma sonucu elde edilen veriler kullanılarak araştırma için gerekli veri temizliği, örnek seçimi, veri temizliği ve etiketleme işlemleri yapılarak veri şartlı rastgele alanlar metodunda kullanılabilecek şekilde yeniden düzenlenmiştir. Bu hazırlıklar sonucu oluşturulan veri ile birlikte gözetimli makine öğrenmesi modelinin ihtiyaç duyacağı şekilde eğitim ve test veri setleri oluşturularak model tamamlanmış ve çıkan sonuçlar değerlendirilmiştir.

Anahtar Kelimeler: İlişki Çıkarımı, Bilgi Çıkarımı, Doğal Dil İşleme, Şartlı Rastgele Alanlar, Wikipedia, İçerik Çıkarımı, Sıralı Metin İşleme

TABLE OF CONTENTS

Academic Honesty Pledge	5
EXECUTIVE SUMMARY	6
ÖZET	8
TABLE OF CONTENTS	10
1. INTRODUCTION	11
1.1. Literature Review & Hypothesis	12
2. ABOUT THE DATA	14
2.1. General Aspects of the Wikipedia Biography Data	14
2.2. Details of the Data	15
2.3. Initial Data Examination	16
3. PROJECT DEFINITION	20
3.1. Project Objective	20
3.2. Project Scope	20
4. METHODOLOGY	21
4.1. Parsing of Wikipedia Dumps	22
4.2. Labeling & Feature Extraction	23
4.3. Sequential Text Processing Methods (CRF)	24
4.4. Evaluation of Methodologies	25
5. Conclusion	27
APPENDIX A	28
Alphabetical list of part-of-speech tags used in the Penn Treebank Project:	28
REFERENCES	29

1. INTRODUCTION

This thesis objective is to build models that will enable to flag the words in the sentences based on the specified labels & features by extracting the relations between the Wikipedia articles' first sentences and their information boxes.

In a general perspective, Relation Extraction is a subfield of Natural Language Processing (NLP) research which helps to analyze huge amounts of data which can't be analyzed or processed manually without any automated process support. As mentioned before there are two important sub-concepts of NLP that are Information Extraction and Relation Extraction. Relation extraction was our main guide throughout this thesis to find out semantic relationships within articles.

Relation extraction consists of two main processes of Information Extraction. Named entity recognition and coreference resolution processes work consequently to find meaningful relations that will reflect the meaning of that specific NE (Name Entity). Details of the methodologies to be used are explained in Section 4.

Based on the concepts above, aim in this work is to find out relations between the entities in the first sentence and the information boxes of Wikipedia article through the birth date, death date and birth place information's and then to store these relations for the further use. In Wikipedia pages, it has been expected that the most important & informative data for that article exist in the very first paragraph to support information box. Based on this assumption, CRF model that is built for this work tries to classify birth date, death date, and birth place information.

To develop such a model, data preparation steps needed to be fulfilled from parsing of the Wikipedia articles through Wikipedia database to labeling & feature assignment of each word in the first sentences of articles. Details of these steps can be found in Section 2.

Parallel to data preparation stage, an appropriate CRF model is selected for Python language which is called as Python-CRFSuite. This model requires a set of three inputs for each word in the model. These are;

- Word itself
- Part of Speech Tag (POS) (Appendix A)
- Named Entity Label (NER Label)

Word itself and part of speech tags are determined inputs, however, feature of words input are defined and assigned within the scope of this work. The main logic is the labeling first word of a birthday as “B-BDATE”, birth place as “B-LOC” and death date as “B-DDATE” while the following words will be labeled as “I-BDATE”, “I-LOC” and “I-DDATE”. Figure 1 shows labels and part of speech tags of all words in a sample sentence.

	Word_Itself	Part_of_Spc_Tags	Label_of_Words
378	roger	NN	O
379	staub	NN	O
380	-LRB-	-LRB-	O
381	1	CD	B-BDATE
382	july	NNP	I-BDATE
383	1936	CD	I-BDATE
384	--	:	O
385	30	CD	B-DDATE
386	june	NNP	I-DDATE
387	1974	CD	I-DDATE
388	-RRB-	-RRB-	O
389	was	VBD	O
390	an	DT	O
391	alpine	JJ	O
392	ski	NN	O
393	racer	NN	O
394	and	CC	O
395	olympic	JJ	O
396	gold	NN	O
397	medalist	NN	O
398	from	IN	O
399	switzerland	NN	B-LOC
400	.	.	O

Figure 1. Sample model input example of a sentence

1.1. Literature Review & Hypothesis

It is aimed to extract relations between entities in relation extraction applications. Rule-based approaches require a clear pattern to be inferred. In this case, the performance rates in non-official documents are falling. Also, most rule-based approaches today are used for document types tailored to a specific domain. When a system developed in this way is desired to be applied to documents belonging to another domain, the defined rules may be insufficient.

That is why instead of a rule-based approach, a supervised machine learning methodology is a better fit to the problem of this study. The main reason behind this is, our data can't provide effective information for all possible scenarios and the data isn't as structured as a rule-based approach will require. Instead of a rule-based approach, with data on hand, it is possible to label our data which will solve one of the most important requirement of a supervised machine learning method. Training dataset will be created based on these labeled datasets and the model created will be able to build necessary rules to be able to classify based on the inputs and the requirements provided.

Based on the definition of the problem in this study, a set of specific methods and models has been chosen. First, in order to parse raw Wikipedia data the Stanford CoreNLP[13] toolkit have been used which logic defined detailly by the work of Manning et. al. [13]. This tool has been used in two different steps within this thesis. First one is during the parsing of Wikipedia dumps which outputs the data in the format that can be seen in Figure 3 (First paragraphs) and Figure 5 (Information box) for each biography. Nguyen, Matsuo, and Ishizuka [3] highlighted a point in the nature of Wikipedia which is in most of the sentences in an article discuss its principal entity besides they claim that it is likely that the relation between a pair of secondary entities may also be discussed in the same article. That is why during the preparation of training dataset, labels (NER) and part-of-speech (POS) tags of both sequential entities will be considered. With support of this assumption, definition of Conditional random fields made by Wallach (2004) [10] have a great match on the basis of needs and solution expected within this work.

In this study, it is aimed to find out a relation between the first sentences of Wikipedia biography articles and the information boxes based on the birth date, birth place and death date of biography owner. As a starting point, it was aimed to get the labels of the birth place, birth date, and death date information from the information box of the Wikipedia pages and then map these labels with the entities in the sentences. This formed the foundation of training dataset for the model. It was preferred to use Conditional Random Fields in the study which is developed for extracting relations from Wikipedia texts. By labeling the entity names in the information box, the required training set for Conditional Random Fields was created and the system was trained based on these train and test datasets.

Logics which supports the reason of why CRF has been chosen in this study explained explicitly by Lafferty, McCallum and Pereira [6]. For example, CRF's avoid the bias problem that can be caused by Maximum Entropy Markov Models (MEMMs). This functionality of CRF prevents the possible bias among the labels of entities. In this study, POS tags of the entities have also been considered as pro-active action to prevent bias in the model output.

The main assumption of this paper relies on the fact that for Wikipedia biographies, there is a high possibility of first sentences to cover the birth date, birth place and death date of biography owner.

2. ABOUT THE DATA

Wikipedia, as it can be understood from its motto of “free encyclopedia that anyone can edit”, enables most of all its users to update the contents. This right allows for continued discussion of the authenticity and reliability of the information in Wikipedia, but research on the validity of the information in Wikipedia reveals that the information accuracy is at the level of Ana Britannica[12].

English Wikipedia dumps are the main data source of this paper. English Wikipedia consists of 5.674.792 free articles in various languages and the number of articles within increases every day with the support of 33.940.030 enthusiastic users of it.

Wikipedia like many other topics that require manual activity sometimes suffers from wrong data entry caused by users. This issue creates the basis of our problem that is going to be described detailly under project definition section (Section 3).

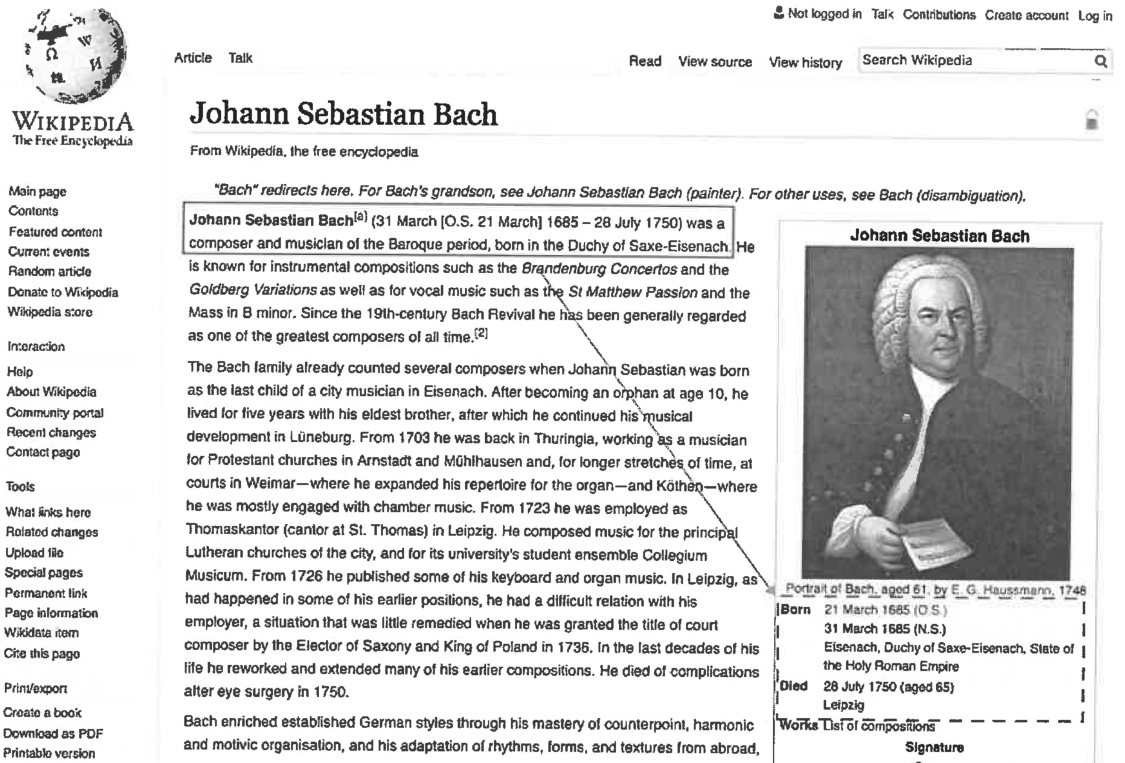
In this paper, a dataset of total 173.288 English Wikipedia biographies has been examined. This dataset has been separated into two subsets as shown in Table 1.

2.1. General Aspects of the Wikipedia Biography Data

As expected, the content of biography articles in the Wikipedia generally starts with a short brief of the article owner which includes birth & death details (time & place) and their area of expertise in the first sentence. The main point of differentiation of information box from the introduction of the article is, information is in a more structured shape as it can be seen in Figure 2.

Information boxes consist of several structured data about the article owner, these data's in the information table formed the label data that has been used to label related words in the first sentence of the articles. For this purpose, this paper used only three main areas from information boxes for labeling of the first sentences. These three areas are;

- Birth Date
- Birth Place
- Death Date



The screenshot shows the Wikipedia article for Johann Sebastian Bach. The first sentence is highlighted: "Johann Sebastian Bach^[a] (31 March [O.S. 21 March] 1685 – 28 July 1750) was a composer and musician of the Baroque period, born in the Duchy of Saxe-Eisenach. He is known for instrumental compositions such as the *Brandenburg Concertos* and the *Goldberg Variations* as well as for vocal music such as the *St Matthew Passion* and the Mass in B minor. Since the 19th-century Bach Revival he has been generally regarded as one of the greatest composers of all time.^[2]"

The information box on the right contains the following data:

Portrait of Bach, aged 61, by E. G. Haussmann, 1748	
Born	21 March 1685 (O.S.) 31 March 1685 (N.S.) Eisenach, Duchy of Saxe-Eisenach, State of the Holy Roman Empire
Died	28 July 1750 (aged 65) Leipzig
Works List of compositions	
Signature	

Figure 2. Sample Information Box & First Sentence

2.2. Details of the Data

Dataset consists of 173.288 biographies from Wikipedia as a result of Wikiproject Biography study [8]. The dataset provides the following information in 6 different text files:

- SET.id contains the list of Wikipedia ids, one article per line.
- SET.url contains the URL of the Wikipedia articles, one article per line.

- SET.box contains the infobox data, one article per line.
- SET.nb contains the number of sentences in the first paragraph of each article, one sentence per line.
- SET.sent contains the sentences of the first paragraphs, one sentence per line.
- SET.title contains the title of the Wikipedia article, one per line.

Within these files “.sent”, “.nb” and “.box” files under train & test subdirectories conduct the base datasets of this study.

2.3. Initial Data Examination

In order to start data cleaning process, we first explored the data type and data format in these text files. We then defined actions to prepare input data for model development.

In Figure 3 below, a sample of sentences in first paragraphs, “.sent” datasets, is shown. This dataset contains every sentence in the first paragraphs of each article in one line. The very first action taken for this dataset is the extraction of only *first sentences* of each article to create a new text file that consists of only *first sentences* in each line and call this new dataset “.sent_v2”. To achieve this, “.nb” dataset has been used as a reference point to be able to find and extract first sentences.

leonard shenoff randle -lr- born february 12 , 1949 -rr- is a former major league baseball player .
 he was the first-round pick of the washington senators in the secondary phase of the june 1970 major league baseball draft , tenth overall .
 philippe adnot -lr- born 25 august 1945 in rhèges -rr- is a member of the senate of france .
 he was first elected in 1989 , and represents the aube department .
 a farmer by profession , he serves as an independent , and also serves as the head of the general council of aube , to which he was elected to represent the canton of méry-sur-seine in 1988 .
 in 1996 and 2008 , he was re-elected to the senate in the first round , avoiding the need for a run-off vote .
 having contributed to the creation of the university of technology of troyes , in 1998 he was made the first vice president of the university board , of which he is currently the president .
 he is a member of the senate 's committee on the laws relating to the freedoms and responsibilities of universities .
 as of 2009 , he serves as the delegate from the administrative meeting for senators not on the list of another group he is decorated as a chevalier of the ordre national de mérite agricole .
 miroslav popov -lr- born 14 june 1995 in dvůr králové nad labem -rr- is a czech grand prix motorcycle racer .
 he currently races in the fiv cev moto2 championship for montaze broz racing team aboard a suter .
 john " jack " reynolds -lr- 21 february 1869 -- 12 march 1917 -rr- was a footballer who played for , among others , west bromwich albion , aston villa and celtic .
 as an international he played five times for ireland before it emerged that he was actually english and he subsequently played eight times for england .
 he is the only player , barring own goals , to score for and against england and is the only player to play for both ireland and england .

Figure 3. Sample view of “.sent” dataset

The second dataset that has been examined is “.box” dataset which consists of the “information box” inputs of each article as shown in Figure 4.

"name_1:jackname_2:reynolds image_1:jack image_2:reynolds.jpg
 image_size_1:200px caption:<none> fullname_1:john
 fullname_2:reynolds birth_date_1:21 birth_date_2:februarybirth_date_3:1869
 birth_place_1:blackburn birth_place_2:, birth_place_3:england
 height_1:5 height_2:4 death_date_1:12 death_date_2:march
 death_date_3:1917 death_place_1:sheffield death_place_2:,
 death_place_3:england position_1:midfielder/forward youthyears_1:1884
 youthyears_2:-- youthyears_3:1885 youthyears_4:1886
 youthclubs_1:witton youthclubs_2:blackburn youthclubs_3:rovers
 youthclubs_4:blackburn youthclubs_5:park youthclubs_6:road
 years_1:1884 years_2:-- years_3:1885 years_4:1886 years_5:1891 years_6:--
 years_7:1892 years_8:1892 years_9:1892 years_10:-- years_11:1893
 years_12:1893years_13:-- years_14:1897years_15:1897years_16:1898
 years_17:1898years_18:-- years_19:1899years_20:1899years_21:--
 years_22:1902years_23:1902years_24:-- years_25:1903years_26:1903
 years_27:1904years_28:-- years_29:1905clubs_1:wittonclubs_2:blackburn
 clubs_3:roversclubs_4:blackburn clubs_5:park clubs_6:road clubs_7:west
 clubs_8:bromwich clubs_9:albionclubs_10:droitwich clubs_11:town
 clubs_12:west clubs_13:bromwich clubs_14:albion clubs_15:aston
 clubs_16:villa clubs_17:celtic clubs_18:southampton clubs_19:bristol
 clubs_20:st clubs_21:george clubs_22:royston clubs_23:f.c. clubs_24:-
 lrb- clubs_25:yorkshire clubs_26:-rrb-clubs_27:grafton clubs_28:f.c. clubs_29:-
 lrb- clubs_30:new clubs_31:zealand clubs_32:-rrb-clubs_33:stockport
 clubs_34:county clubs_35:willesden clubs_36:towncaps_1:17 caps_2:20
 caps_3:96 caps_4:4 caps_5:2 caps_6:1 goals_1:2 goals_2:1
 goals_3:17 goals_4:1 goals_5:0 goals_6:0 nationalyears_1:1890
 nationalyears_2:-- nationalyears_3:1891nationalyears_4:1892nationalyears_5:--
 nationalyears_6:1897nationalyears_7:189xnationalyears_8:--
 nationalyears_9:189x nationalteam_1:ireland nationalteam_2:england
 nationalteam_3:english nationalteam_4:league nationalteam_5:xi
 nationalcaps_1:5 nationalcaps_2:8 nationalcaps_3:4 nationalgoals_1:1
 nationalgoals_2:3 article_title_1:jack article_title_2:reynolds

*article_title_3:-lrb- article_title_4:footballer article_title_5:,
article_title_6:born article_title_7:1869 article_title_8:-rrb-“.*

This expression may seem confusing in first look however because of each value has a structured label, we were able to use related values as target values while labeling *birth_date*, *birth_place* and *death_date* informations. As mentioned, values in the “.box” correspond to our target values to label the related information in the first sentences of articles, that’s why it is critical to extract values in a format that can be compared with the values in the first sentences. When the structure of birth date formats in the first sentences extracted, several different types of formats encountered such as below. Birth place formats vary on a much wider scale which makes it almost impossible to standardize. Therefore during data preparations, we will look for a one-to-one match between the selected target values in “.box” dataset and values that are going to be labeled in the first sentences of articles.

Birth date format samples in the articles:

- 25 april 1900
- born december 9 , 1985
- november 19 , 1802
- born 8 august 1974
- born august 20

Another important point to highlight is about the null values in *birth_date*, *birth_place* and *death_date* informations. The main assumption for the missing *birth_date* and *death_date* informations is, if the value is missing we did not try to fill these values because the model that is going to be built should be able to differentiate whether there is any required information in the first sentence or not.

Before going into details of data preparation, it is important to understand CRF model and its requirements detailly in terms of model developments. Related data preparations which are going to be done based on model requirements are explained detailly in Section 4.

3. PROJECT DEFINITION

Wikipedia like many other topics that require manual activity sometimes suffer from wrong data entry caused by users. These information boxes that this study focuses on get data from users manually based on specific categorizations. Although there are some specific categories or directions to fill these areas, wrong entries are still possible. Importance of these information boxes is, they provide a generic but important information of the topic and can provide fast answers for the users of Wikipedia.

3.1. Project Objective

The objective of this project is to find out meaningful semantic relationships between information in the first sentence of the first paragraph of articles and their information boxes through the labels mentioned in Section 2.1..

To achieve the project objective, a supervised machine learning model was trained based on the POS and NER tags of words. The reason of using this method is to both understand the quality of information box inputs while on the other hand developing a classification & prediction model which is able to find out whether words with required labels are in the sentence or not. If they are, this model is able to identify these words.

3.2. Project Scope

Based on the defined objective of the project in this thesis being able to develop consistent supervised classification model, feature and labeling of the data have a great importance. That is why during data preparation steps, different labeling logics have been tried & reviewed and as a labeling method with a sequential logic have been selected and embedded in the project. The reason behind the selection of sequential labeling logic is strongly correlated with the way of CRF model is working. The way how CRF works and why it matches the objective of this study explained briefly by Hanna M. Wallach (2004) [10] as a model which provides a conditional probability $p(Y/x)$ where X and Y are both random variables, over label sequences give a particular observation sequence x , rather than a joint distribution over both label and observation sequences. This kind of approach enables CRF to act as a probabilistic framework for labeling and segmenting sequential data [10].

4. METHODOLOGY

Based on the data types in our dataset, information boxes provide a confidence to label the target birth date, birth place and death date information in the first sentence. For this task steps shown below were followed.

- 4.1 Parsing of Wikipedia dumps
- 4.2 Labeling & Feature Extraction
- 4.3 Sequential text processing methods (CRF)
- 4.4 Evaluation of methodologies

Culotta, McCallum, and Betz in their paper named *Integrating Probabilistic Extraction Models and Data Mining to Discover Relations and Patterns in Text* 2006 [16], they explain CRF (Conditional Random Fields) method as more flexible than the other supervised methodologies that can be used for text processing such as logistic regression & kernel.

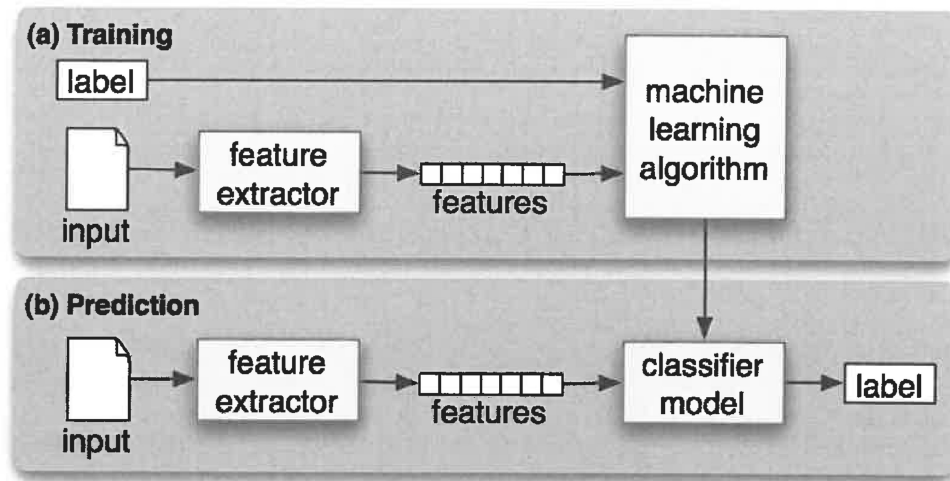


Figure 6. Generic Representation of Supervised Classification Methodology for Text Classification

The main idea of CRF is to train a model based on the given labels and extracted features of the words. As a supportive parameter to improve the consistency of model is to use part of speech tags (Appendix A) of words within the model in addition to features.

Since our focus here is to predict/classify birth date, birth place and death date of articles based on the relations between information boxes and first sentences, below labels were used;

- B-BDATE for the first part of birth date
- I-BDATE for following parts of birth date
- B-LOC for the first part of birth place
- I-LOC for following parts of birth place
- B-DDATE for the first parts of death date
- I-DDATE for following parts of death date

Next sections explain the details of data preparation and model development for CRF model.

4.1. Parsing of Wikipedia Dumps

Project data of this study have been processed by Stanford Core NLP toolkit under two steps. First of all, the sentences in the first paragraph, the number of sentences in each paragraph and the inputs of the information boxes of articles have been parsed. These are the main raw data of the project and required an intense preparation period.

In correlation with the goal of this work, first sentences of paragraphs have been extracted. During this process, “*nb*” has been used as a reference point to be able to point out the first sentences of each paragraph.

As a second step, to be able to convert these sentences into their tokens and determine their part-of-speech tags The Stanford CoreNLP Toolkit [13] have been used. This toolkit provided three main outputs for each sentence which are;

- Tokenized text
- Part-of-speech tags
- Named entity recognition (person, location, organization, etc)

The strong side of Stanford CoreNLP is that it can work with any type of character encoding which provided more of project data can be used in model development. However,

there were still sentences with non-ascii characters in the file which need to be eliminated from data.

At this stage of project **173.288** sentences have been considered. As a data cleaning action thanks to the nature of our data there isn't any null values but instead, there are sentences which contain non-ascii characters that needed to be eliminated. As a result of this elimination **26.653** sentences that contain non-ascii characters have been deleted from dataset and total of **146.635** sentences were remained for model development.

	Sentence	Words & Entities
Train	102.644	2.588.567
Test	43.991	1.107.272

Table 1. Train & Test dataset quantities

4.2. Labeling & Feature Extraction

Part-of-speech (POS) tags and named entity recognition tags are key elements for the dataset since they were used to develop a supervised CRF model.

Part-of-speech (POS) (Appendix A) is mainly a process of marking up a word in a text as corresponding to a particular part-of-speech, based on its definition, context, relationship with dependent / adjacent words and etc [14]. POS leads us to differentiate words not only through their meanings but also according to their connected words within a sentence or a paragraph. POS tags have been created automatically by Stanford CoreNLP toolkit. The main benefit of POS tags to the model is, POS tags will enable CRF model to differentiate different uses of same word which will help to decrease false positive results of CRF model. This will lead to decrease frequency of type 2 errors in models prediction.

Feature Extraction & Labeling (NER), in the original output of Stanford CoreNLP Toolkit, words have been labeled with several different features such as date, name, location, other and with many different features. These features are the main reference point for a supervised model during the training step. At first, labels in information boxes are matched with the words in sentences. Because of only birth date, birth place and death date informations are going to be considered, only below labels are assigned as features of the dataset,

- B-BDATE for first part of birth date
- I-BDATE for following parts of birth date

- B-LOC for first part of birth place
- I-LOC for following parts of birth place
- B-DDATE for first part of death date
- I-DDATE for following parts of death date

```
[['bias',
  'word.lower=1985',
  'word[-3:]=985',
  'word[-2:]=85',
  'word.isupper=False',
  'word.istitle=False',
  'word.isdigit=True',
  'postag=CD',
  'postag[:2]=CD',
  'BOS',
  '+1:word.lower=extra',
  '+1:word.istitle=False',
  '+1:word.isupper=False',
  '+1:postag=JJ',
  '+1:postag[:2]=JJ'],
 ['bias',
  'word.lower=1985',
  'word[-3:]=985',
  'word[-2:]=85',
  'word.isupper=False',
  'word.istitle=False',
  'word.isdigit=True',
  'postag=CD',
  'postag[:2]=CD',
  'BOS',
  '+1:word.lower=extra',
  '+1:word.istitle=False',
  '+1:word.isupper=False',
  '+1:postag=JJ',
  '+1:postag[:2]=JJ'] ]
```

Figure 7. Features of words

4.3. Sequential Text Processing Methods (CRF)

When the possible variations and probabilities of target values that this study seeking in the dataset have tried to be identified, the term of entropy comes up as a term need to be considered.

Each type of information has different variation of sequences. Date information in the sentences varies mainly amongst below possibilities,

- 25 April 1900
- Born/died December 9, 1985

- November 19, 1802
- Born/died 8 august 1974
- Born/died august 20

General expectation of information's in the sentences are in the sequence of birth date, birth place and death date. Of course, because of not all sentences have three of these informations at the same time these sequences may vary.

When all these different possibilities considered a distribution of related target values results in a uniform distribution. For datasets with maximum entropy representing the information with a defined set of features such as those defined in this work may face with a label bias problem [6]. What label bias problem represents has been defined by Lafferty, McCallum and Pereira as "If one of the two words is slightly more common in the training set, the transitions out of the start state will slightly prefer its corresponding transition, and that word's state sequence will always win" (June, 2001, p.2) [6]. What CRF does best to overcome these problems is to develop a model based on a probabilistic approach while considering global maximum likelihood convergence.

4.4. Evaluation of Methodologies

Based on the knowledge gained about how CRF model should be trained and run to get classification results all project train & test data have been formed accordingly as mentioned in Section 4.2.

Model success is mainly depending on the classification quality of selected labels. Therefore, the words with labels of "O" have been ignored.

In Figure 7 first set of results were obtained with model parameters being set as;

- L1 penalty = 1.0
- L2 penalty = 1e-3
- Max iterations = 100
- Feature.possible_transitions = True

	precision	recall	f1-score	support
B-BDATE	0.92	0.97	0.95	69588
I-BDATE	0.81	0.96	0.88	115266
B-DDATE	0.92	0.88	0.90	24340
I-DDATE	0.79	0.55	0.65	45330
B-LOC	0.65	0.50	0.57	24606
I-LOC	0.67	0.50	0.57	17860
avg / total	0.82	0.83	0.82	296990

Figure 8. Classification results

Based on these precision, recall and f1-score results, model has a good performance to classify birth date & death date information's from the sentences however the performance of the birth place is below average with same evaluation metrics.

To better understand these results, transitions between the labels are evaluated as shown in Figure 8. Based on these results, as expected, most likely results are the B-BDATE & B-DDATE labels followed by I-BDATE and I-DDATE. Positive weights in the table mean that model thinks that birth date information is followed by birth date information in many scenarios B-BDATE -> B-BDATE for example. Same applies for I-BDATE -> I-BDATE and I-DDATE -> I-DDATE versions as well.

```

Top likely transitions:
B-BDATE -> B-BDATE 10.120944
O -> O 7.787368
I-BDATE -> I-BDATE 6.980748
I-DDATE -> I-DDATE 5.465836
B-DDATE -> B-DDATE 5.402862
B-LOC -> B-LOC 4.134237
I-LOC -> I-LOC 3.271496
B-BDATE -> I-LOC -0.557935
I-LOC -> B-BDATE -0.557935
B-BDATE -> B-LOC -0.854191
B-LOC -> B-BDATE -0.854191
B-BDATE -> I-BDATE -0.983034
I-BDATE -> B-BDATE -0.983034
B-BDATE -> I-DDATE -1.108143
I-DDATE -> B-BDATE -1.108143

```

```

Top unlikely transitions:
I-LOC -> I-BDATE -3.226522
B-DDATE -> I-DDATE -3.380249
I-DDATE -> B-DDATE -3.380249
O -> I-DDATE -3.491745
I-DDATE -> O -3.491745
I-LOC -> I-DDATE -3.569848
I-DDATE -> I-LOC -3.569848
I-BDATE -> I-DDATE -4.685138
I-DDATE -> I-BDATE -4.685138
O -> B-LOC -5.201798
B-LOC -> O -5.201798
O -> I-LOC -5.285091
I-LOC -> O -5.285091
B-LOC -> I-LOC -5.604038
I-LOC -> B-LOC -5.604038

```

Figure 9. Transitions between labels

5. Conclusion

To sum up, during the initiation of this study main aim was to find out a relation between the entities in the first sentences of Wikipedia articles and their information boxes. With the support of Information Extraction, Natural Language Processing, Sequential Text Processing and Conditional Random Fields techniques a supervised CRF model has been trained. Results of this model supported the main assumption of this study which is about the direct relationship between the information boxes and first sentences of articles in terms of birth date, birth place and death date information.

From this point, a solution can be studied as a further study to the lack of entities problem in the information boxes. As mentioned, information boxes are tables filled by the writers of that article. Because of all these processes depends on human effort, there is a great possibility for the existence of wrong inputs or null values in information boxes. With the support of this study a validation model may be developed to check the validity of the entities in the information boxes or if there are null values these boxes may be filled with the expected value extracted from the article itself.

APPENDIX A

Alphabetical list of part-of-speech tags used in the Penn Treebank Project:

Number	Tag	Description
1.	CC	Coordinating conjunction
2.	CD	Cardinal number
3.	DT	Determiner
4.	EX	Existential <i>there</i>
5.	FW	Foreign word
6.	IN	Preposition or subordinating conjunction
7.	JJ	Adjective
8.	JJR	Adjective, comparative
9.	JJS	Adjective, superlative
10.	LS	List item marker
11.	MD	Modal
12.	NN	Noun, singular or mass
13.	NNS	Noun, plural
14.	NNP	Proper noun, singular
15.	NNP S	Proper noun, plural
16.	PDT	Predeterminer
17.	POS	Possessive ending
18.	PRP	Personal pronoun
19.	PRP\$	Possessive pronoun
20.	RB	Adverb
21.	RBR	Adverb, comparative
22.	RBS	Adverb, superlative
23.	RP	Particle
24.	SYM	Symbol
25.	TO	<i>to</i>
26.	UH	Interjection
27.	VB	Verb, base form
28.	VBD	Verb, past tense
29.	VBG	Verb, gerund or present participle
30.	VBN	Verb, past participle
31.	VBP	Verb, non-3rd person singular present
32.	VBZ	Verb, 3rd person singular present
33.	WDT	Wh-determiner
34.	WP	Wh-pronoun
35.	WP\$	Possessive wh-pronoun
36.	WRB	Wh-adverb

REFERENCES

Use APA styling for citations

<http://www.bibme.org/apa>

- [1] Lange, D., Böhm, C., Naumann, F., (2010). "Extracting Structured Information from Wikipedia Articles to Populate Infoboxes", Intl. Conf. of Information and Knowledge Management.
- [2] Weld, F. Wu and D. S. (2007). "Autonomously Semantifying Wikipedia", Conf. on Information and Knowledge Management.
- [3] Nguyen, D. P. T., Matsuo, Y., Ishizuka, M., (2007). "Exploiting Syntactic and Semantic Information for Relation Extraction from Wikipedia", IJCAI 2007 Workshop on Text-Mining & Link-Analysis.
- [4] Cucerzan, S., ve Yarowsky, . (1999). "Language Independent Named Entity Recognition Combining Morphological and Contextual Evidence", SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora.
- [5] Manning, C., & Schütze, H., (1999). "Foundations of Statistical Natural Language Processing.", MIT Press.
- [6] Lafferty, J., McCallum, A. ve Pereira, F., (2001). "Conditional random fields: probabilistic models for segmenting and labeling sequence data", International Conference on Machine Learning.
- [7] CoreNLP Tool developed by Stanford University
<http://stanfordnlp.github.io/CoreNLP/>
- [8] Project to create, develop and organize Wikipedia articles about persons & biographies.
https://en.wikipedia.org/wiki/Wikipedia:WikiProject_Biography
- [9] Canan Girgin, (2010) "Semantic Relation Extraction by Conditional Random Fields From Turkish Wikipedia Pages"
- [10] Hanna M. Wallach (2004) "Conditional Random Fields: An Introduction"
- [11] Steven Bird, Ewan Klein and Edward Loper
- [12] BBC News, Wikipedia survives research test,
<http://news.bbc.co.uk/2/hi/technology/4530930.stm> , 05.08.2018
- [13] Jenny Finkel, Christopher D. Manning, Mihai Surdeanu, John Bauer, Steven J. Bethard, David Mc. Closky (2014) "The Stanford CoreNLP Natural Language Processing Toolkit"
- [14] Part-of-speech tagging, Wikipedia,
<http://www.wikizeroo.net/index.php?q=aHR0cHM6Ly9lbi53aWtpcGVkaWEub3JnL3dp a2kvUGFydC1vZi1zcGVyY2hfdGFuZ2luZw> , 16.12.2018
- [15] A. McCallum, D. Freitag and F. Pereira. Maximum Entropy Markov models for

- information extraction and segmentation. In *International Conference on Machine Learning*, 2000
- [16] Aron Culotta, Andrew McCallum, Jonathan Betz (2006) *Integrating Probabilistic Extraction Models and Data Mining to Discover Relations and Patterns*