Mühendislik Bilimleri ve Tasarım Dergisi
7(3), 539 – 551, 2019
e-ISSN: 1308-6693

Journal of Engineering Sciences and Design
DOI: 10.21923/jesd.448251

Araştırma Makalesi Research Article

A NAMED ENTITY RECOGNITION MODEL FOR TURKISH LECTURE NOTES IN HISTORY AND GEOGRAPHY DOMAINS

Önder Can SARI[1*], Özlem AKTAŞ[2]

[1]Dokuz Eylül University, Grad. School of Natural and Applied Sciences, Dept. of Computer Engineering, İzmir, Turkey
[2]Dokuz Eylül University, Engineering Faculty, Dept. of Computer Engineering, İzmir, Turkey

Keywords

Computational linguistics, Named entity recognition, Natural language processing, Information extraction, Educational technology.

Abstract

Named entity recognition (NER) is an information extraction (IE) task that is in the scope of natural language processing (NLP) and text mining. Its extent and methods may differ between studies, but basically, it aims to detect expressions that indicates a person, location, organization etc. In this study, a NER structure is developed for Turkish lecture notes (for history and geography courses). Separately, this structure is a project that is specialized for an information extraction task. Besides, it also has an educational value, as the projected outcome from its execution is meaningful words or word groups from the content of input lecture notes, which can be used to construct glossary of terms structures for individual courses or course subjects. With these glossary of terms structures, it is aimed to detect expressions in the content of a lecture note that can be used for questions and support a test preparation process. In this document, general information about NER task and its scope is given; previous studies on the field are mentioned; the system developed in line with this study is introduced; success of the system is evaluated through experiment results and some thoughts for enhancement are shared.

TARİH VE COĞRAFYA ALANINDAKİ TÜRKÇE DERS METİNLERİ İÇİN BİR VARLIK İSMİ TANIMA MODELİ

Anahtar Kelimeler

Bilişimsel dilbilim, Varlık ismi tanıma, Doğal dil işleme, Bilgi çıkarımı, Eğitimsel teknoloji.

Öz

Varlık ismi tanıma; doğal dil işleme ve metin madenciliği alanlarının kapsamında yer alan bir bilgi çıkarımı görevidir. Kapsam ve kullanılan metotlar açısından, çalışmalar arasında farklılıklar görülse de temel olarak, bir metin içerisindeki kişi, yer, kurum-kuruluş vb. belirten ifadelerin doğru şekilde tespit edilmesini hedefler. Bu çalışmada, Türkçe yazılmış ders metinleri (tarih ve coğrafya alanlarında) için bir varlık ismi tanıma yapısı geliştirilmiştir. Tek başına ele aldığımızda bu yapı, bir bilgi çıkarımı görevi doğrultusunda özelleştirilmiş bir projedir. Bunun yanı sıra çalışmanın eğitimsel bir değeri de vardır; çünkü sistemden beklenen sonuç, verilen ders metninin içeriğinden anlamlı kelime ya da kelime grupları bulunmasıdır ki; bu da farklı dersler ya da ders konuları için terimler sözlüğü yapıları oluşturmak için kullanılabilir. Oluşturulan sözlüklerin, bir ders metninin içeriğindeki soru değeri taşıyabilecek ifadelerin tespitine ve sınav hazırlama sürecine yardımcı olması hedeflenmektedir. Bu dokümanda, varlık ismi tanıma görevi ve görevin kapsamı hakkında genel bilgi verilmiş; alanda yapılmış önceki çalışmalardan bahsedilmiş; bu çalışma doğrultusunda geliştirilen sistem tanıtılmış; sistemin başarısı, yapılan deney sonuçları üzerinden değerlendirilmiş ve geliştirme-iyileştirme olanakları hakkında yorumlar paylaşılmıştır.

Alıntı / Cite

Sarı, Ö., Aktaş, Ö. (2019). A Named Entity Recognition Model For Turkish Lecture Notes In History and Geography Domains, Journal of Engineering Sciences and Design, 7(3), 539-551.

Yazar Kimliği / Author ID (ORCID Number)	Makale Süreci / Article Process	
Ö. Sarı, 0000-0003-4226-9633	**Başvuru Tarihi / Submission Date**	26.07.2018
Ö. Aktaş, 0000-0001-6415-0698	**Revizyon Tarihi / Revision Date**	06.01.2019
	Kabul Tarihi / Accepted Date	03.04.2019
	Yayım Tarihi / Published Date	15.09.2019

[*] İlgili yazar / Corresponding author: onder.sari@ceng.deu.edu.tr, +90-232-375-2317

1. Introduction

The term *named entity (NE)* is used to define anything that can be referred to with a proper name. The process of *named entity recognition* (NER), which is a subtask of information extraction, aims to locate and classify named entities in text into pre-defined categories. This is a combined task; as it must fulfill two requirements respectively: First task is to find bounds of text that constitute proper names; second one is to classify them according to their types correctly.

Generic news-oriented NER systems focus on detection of things like people, places and organizations, while specialized applications may be concerned with many other types of entities, including commercial products, works of art, proteins, genes and other biological entities (Jurafksy and Martin, 2009). In most NER systems, it is a common approach to extend the scope of a NE to include things that aren't proper names; but have characteristic meanings within the text. This generally leads the inclusion of temporal expressions like dates, times, named events and numerical expressions like measurements, counts, prices to the NE categories (also called as tags). The system that will be detailed on this paper is specialized for Turkish lecture notes within geography and history domains to detect named entities. Detected characteristic terms are the projected sources to build glossary of terms structures for geography and history domains.

Table 1. A list of NE types with the kinds of entities they refer to

Type	Tag	Sample Categories
People	PER	Individuals, fictional characters, small groups
Organization	ORG	Companies, agencies, political parties, sports teams
Location	LOC	Physical extents, mountains, lakes, seas
Geo-Political Entity	GPE	Countries, states, provinces, counties
Facility	FAC	Bridges, buildings, airports
Vehicles	VEH	Planes, trains, automobiles

NER systems mostly take an unannotated block of text as input and produce and annotated block of text that points the names of entities. For example, the projected output for the unannotated input text "Mustafa Kemal Atatürk 1881 yılında Selanik'te doğdu." *(Mustafa Kemal Atatürk was born in Thessaloniki in 1881.)* is "[Mustafa Kemal Atatürk]$_{\text{Person}}$ [1881]$_{\text{Date}}$ yılında [Selanik]$_{\text{Location}}$'te doğdu."

Word ambiguity is a major concern for NER systems, like most of the other natural language processing (NLP) tasks. For example, the word "Washington" might indicate a person, a location, an organization (a sports club) or a facility (a ship). Or the word occurrence "Ural" in Turkish text can refer to a location (a river) or a person. NER systems use different approaches to overcome issues like that and increase their success rates.

Word segmentation (tokenization) is a common starting point for NER systems. If statistical techniques are preferred, sequence labeling is the next step. In this approach, classifiers are trained to label the tokens with tags to indicate presence of particular kinds of named entities. IOB-format (inside-outside-beginning), which tries to distinguish the beginning of named entities (B), words inside a NE (I) and unrelated words that are outside NEs (O), is a widely used tagging format. IO-format is a more generalized approach as it classifies tokens as inside or outside a NE. Table 2 shows the resulted tag sequences when these two encoding formats are applied on an example sentence. B-PER indicates token to be beginning of a person typed NE, I-PER indicates token to be inside a NE. PER expression used on IO encoding indicates that token is part of a person typed NE and doesn't provide additional information about beginnings. As the example shows, IO encoding might implicate erroneously merged named entities like "Sue Edvard Munch", while IOB encoding correctly implicates two separate named entities as "Sue" and "Edvard Munch".

Table 2. Difference between IO and IOB encoding

Token	IO encoding	IOB encoding
Fred	PER	B-PER
showed	O	O
Sue	PER	B-PER
Edvard	PER	B-PER
Munch	PER	I-PER
's	O	O
last	O	O
painting	O	O

Selection of a set of features is also significant for NER success. Features are derived information about tokens, which are used to make predictions more accurate. Multiple features can be used for training NER systems. *Shape* is a common feature that holds information about character-level forms of tokens like lower-cased, capitalized, all capital, mixed case, contains hyphen etc. Capitalization and punctuation marks are important clues to detect named entities on structured texts. Most NER systems benefit from *gazetteers* that are large lists of place names. Similar lists for corporation names, biological terms also exist. Lists with *predictive words* like honorifics, titles can also be used to find clues. List usages are used to derive information about tokens, for features like *exists in gazetteer* or *predictive token*. "Part-of-speech (POS) tag, term after stemming, bag of words, syntactic chunk label" are some of the other utilizable features.

NER algorithms are mainly divided into three models: Statistical, rule-based and hybrid approaches. Main paradigm of statistical models is to automatically learn

rules and patterns of named-entities through a pre-annotated training data. Additionally, training data has to be labeled to provide information about selected features (if used). Most common statistical models are HMM (Hidden Markov Models), ME (Maximum Entropy) and CRF (Conditional Random Fields). Rule-based models rely on orthographical, morphological and lexical information derived from feature sets. Syllabication, tokenization, morphological analysis or lexicon lookups are the main operations to assign feature values. Using lexicons to store person, location and organization names that imply a NE existence is a common approach. Pre-defined grammatical rules or character transformation conditions about the source language are also beneficial, especially for agglutinative languages which require intensive suffix usage. Lexicon structures are generally stored in databases to provide robustness, while defining grammatical rules as built-in resources is also preferable to allow easier modification if necessary. Hybrid models aim to exploit advantages of both statistical and rule-based approaches with a combined structure. Although it is a serviceable approach to minimize the effects of domain changes, storage requirements and possible system overhead should not be neglected.

A NER system with high-success rate might be serviceable for many applications and use case scenarios in today's world; like classifying content for news providers, recommender systems, customer support, media analysis, sentiment detection, email scanning, more accurate literature search or educational purposes which the proposed model is developed for.

2. Scientific Literature Review

Message Understanding Conferences (MUC) are designed to promote and evaluate research in information extraction (IE). These conferences were initiated by Navy Operational Support Center (NOSC) to assess research on the automated analysis of military messages containing textual information. Two primary evaluation metrics *precision* and *recall* are detailed and used for IE tasks in MUC-2. NER for English is one of the tasks of MUC-6 which is organized in 1996. Training corpus is generated by annotating Wall Street Journal articles. ENAMEX (for people, organization, location) and NUMEX (time, currency, percentage) tags are introduced in this conference. 15 participants enrolled for the NER task. Most successful system reached 97% precision and 96% recall values. (Grishman and Sundheim, 1996)

Cucerzan and Yarowksy (1999) is the first published NER research that includes Turkish. System is language independent and depends on bootstrapping algorithm with iterative learning on a character-based tree structure. System is built after the acceptance that words strongly tend to exhibit only one sense in a document. It uses a small NE list about the source language as training seeds and morphological and contextual patterns as features. For example, "-escu" is stated as an almost perfect indicator for a last name in Romanian. This study reports 60% precision, 47% recall and 53% f-measure for Turkish.

Alfonseca and Manandhar (2002) built a general named entity recognition (GNER) system to find the most accurate generalization (hypernym) for an unknown concept or instance, by using WordNet ontology (lexical database). To classify an unknown instance, system runs queries on search engines to derive similarity scores for candidate words. Used notion here is that words semantically related must co-occur with the same kinds of words.

Tür et al. (2003) developed a NER system based on n-gram language models embedded in Hidden Markov Models. The study consists of four models: Lexical model uses boundary flags between word tokens to indicate name entity borders with *yes, no* and *mid* flags. Contextual model is used to capture information from surrounding context of word tokens. Morphological model uses case information (initial-upper, all-lower, all-upper, mixed etc.) alongside with a proper name database that stores common Turkish person, location and organization names. Tag model is only concerned with trigram possibilities for name entity tag (person, location, organization, else) and boundary flag (yes, no, mid) combinations. Newspaper articles are used for experiments. When all models combined, system has a success rate with 90.4% NE text accuracy, 92.7% NE type accuracy and 91.5% f-measure.

Table 3. Example usage of contextual model for unknown words (Tür et al. [2003])

Output Sequence	Probability
Dr./*else* boundary/*yes* unk/*person*	0.990119
Dr./*else* boundary/*yes* unk/*location*	0.000690
Dr./*else* boundary/*yes* unk/*organization*	0.000880
Dr./*else* boundary/*yes* unk/*else*	0002688

Like MUC, CoNLL events give shared tasks about computational linguistics to participants. Task in CoNLL-2003 is to build a language independent NER (English and German are the test languages); with a special challenge which is to include unannotated data to the training phase of the system. Participants are provided with different features (pos tag, chunk tags, affix information, gazetteers etc.) and given freedom to decide among them. It is observed that, instead of using unannotated data to find out additional gazetteer terms, using them to obtain capitalization information seemed to have much positive effect on results. (Sang and Meulder, 2003)

Wentland et al. (2008) built a multilingual NE resource called HeiNER. Wikipedia is used as the main resource, as it contains a large amount of NEs compared to other

commonly used lexical resources like WordNet. Redirect pages and disambiguation pages of Wikipedia are used to build a disambiguation dictionary. Another advantage of using Wikipedia articles is that, there is a high probability for an article heading to describe a NE. This surpasses some of the common NER problems like NE boundary detection or necessity of morphological normalization.

Küçük and Yazıcı (2009a) built a rule-based NER system for Turkish and tested its success on different domains (news articles, child stories, history texts). System uses lexical resources like dictionary of Turkish person names, list of well-known political people, list of well-known organizations and pre-defined pattern bases to detect possible NEs. Resulted f-measure is 78% for news articles domain; but it drops down to 69% for child stories and 55% for historical texts. Existence of foreign person names in child stories and absence of historical person and organization names in lexical resources are determined to be leading causes for performance drops Results are in line with the general opinion that performance decrease is possible when rule-based NER systems are ported to other domains.

Küçük and Yazıcı (2009b) also tested their system on transcription test derived from video texts. 16 news videos from Turkish Radio and Television Company (TRT) archive are selected for experiments. Videos are manually transcribed as no automatic speech recognizer exists for Turkish back then. Evaluation resulted in a precision of 73%, recall of 77% and f-measure of 75%.

Tatar and Çiçekli (2011) described an automatic rule learning method using supervised learning. System starts with a set of named entities collected from a training dataset and generates rules from them. Main goal here is to get through domain adaptability problems, which is common for rule-based systems. System utilized from orthographical, contextual, lexical and morphological features. 2-level gazetteer structures are used in lexical model. For example, *location* is a higher level, more general categorization while location.country, location.city are secondary level, more specific classification. System is tested on Turkish news articles (TurkIE dataset) and resulted in a precision of 91.7%, recall of 90% and f-measure of 91%.

Küçük and Yazıcı (2012) moved through their rule-based model and developed a hybrid system. 2 statistical features *n* (denotes the number of occurrences of an entity text) and *p* (denoted the number of occurrences which happen to be annotated) are defined and *p/n* is used as a confidence value for each entity. In training phase, entities with high confidence values are extracted and added to the resources of recognizer. Significant performance improvement over rule-based system is observed with f-measure values of 85.9% on news data set, 85% on child stories and 66.9% on historical texts.

Şeker and Eryiğit (2012) used conditional random fields (CRF) as their statistical model. Alongside with gazetteers, they used generator gazetteers (22 person, 44 location, 60 organization) that holds tokens that could come after or before regular words and construct NEs. 14 features are defined in 3 categories (morphological, lexical, gazetteer lookup). Windows width for CRF features is defined as {-3,+3} where 0 is current token, +1 is next, -1 is previous token etc. Features are tested by including them one by one to the system. Experiments showed that all features but SS (start of sentence) had improved performance of the system. When all features included, system had reached 94.6% final f-measure in MUC metrics and 91.9% final f-measure in CoNLL metrics.

Küçük et al. (2014) performed NER experiments on Turkish tweets. 2320 tweets are collected to form data set. Besides seven basic types (person, location, organization [these three are also called as PLO], date, time, money, percent), a misc type (product names, tv shows, music bands etc.) is also used for annotation. Hashtag usage is also suggestive as it is common to have NEs in hashtags. Two lists for person and organization names, which are detected to be used as single tokens in news articles (at least 30 times in Europe Media Monitor database) are built and used in system. Results show that 25% of PLO initial letters are not properly capitalized, only 32% of person names are composed of first name-surname pairs and %10 of PLO text has affected from normalization of Turkish characters. Another problem is the multiword NE tokens in hashtags that are written without whitespace. System reached 66% precision, 31.5% recall and 42.6% f-measure values.

Küçük and Arıcı (2016) composed and shared a dataset comprising news articles in Turkish with named entities annotated, for general use of NER studies. 10 news articles from METU Turkish Corpus are selected and final annotation document consists of 1425 named entities (398 person, 567 location, 460 organization).

Şeker and Eryiğit (2016) moved through their study in 2012 and added TIMEX and NUMEX entity types. They also worked on a new dataset (Web2.0 domain) with user generated content (UGC). Additional features like numeric value, percentage sign etc. are defined and used for new entity types. A lexicon named Auto Capitalization Gazetteer (CAP) is constructed, which contains gazetteer terms that are unlikely to be used as common noun. Unlike their previous study, this time feature performances are tested by removing them from the complete model one by one. This way SS (start of sentence) feature is determined have 2.11% positive effect on performance. Experiments on UGC data set resulted with 67.9% success on best

model. When CAP feature is removed it causes more than 20% performance loss.

Ertopçu et al. (2017) developed different methods to test various parameters and find out most successful results. Best results are reached when multilayer perceptron is used as classifier algorithm (learning rate as 0.1) with window size set as 1 and 7 features (IsCapital, IsDate, IsFraction, IsTime, MainPos, RootForm, SurfaceForm) are selected, with 7.54% error rate.

3. Material and Method

Proposed NER model is an information extraction software developed for educational purposes. It is specialized for Turkish lecture notes within geography and history domains. Primary goal of the research is to detect named entities from the context of input text documents with high accuracy. Building basis for steady and satisfying glossary of terms structures (for history and geography terms) using qualified named entities among the detected is defined as the next step. Utilizing these structures to support a test preparation process is the long-term goal of the research.

This study is a software project that is specialized for an information extraction task. In this direction, a NER model is developed for Turkish lecture notes. History and geography courses are selected as the domain. The model is developed to build a base for a glossary of terms structures (for history and geography terms) which can be used for educational purposes (to support a test preparation process is the projected aim) by finding out named entities in course documents accurately.

3.1. Proposed Framework

Implemented NER structure uses a rule-based model. It takes a text document as input and returns detected named entities with their types as output.

System is developed to work on sentences; so first, the sentence boundary detection (SBD) module is executed on input text file. This module takes an input text file, pre-processes it (removes symbols, other irrelevant characters and whitespaces, connects itemized textual parts to each other etc.), detects headings and sentence boundaries, finally returns a list of sentences and a list of headings. SBD module is provided with rules for sentence boundary conditions, which are translated into regular expressions on back-end side. To minimize erroneous detections arising from abbreviations, a list of Turkish abbreviations (consists of 204 elements) is used to apply an abbreviation check operation.

Table 4 shows example pre-defined sentence boundary rules that are used in SBD module execution. (LC → lower-case character, UC → upper-case character, WS → whitespace, D → digit; true → indicates a sentence boundary condition, false → indicates a not sentence boundary condition)

Table 4. Example SBD rules

Condition	Output
LC . UC	True
LC . LC	False
LC . D	True
UC . LC	False
LC . WS . UC	True
LC . WS . D	True

After execution of SBD module, sentences of input text are made available for NER system. Therefore, the success of NER model also depends on the success of SBD module. Each sentence is handed to NER respectively and processed with the tokenizer, lexical model and contextual model. These three models prepare given sentence by providing informative labels. Finally, the recognizer model is executed and sentence with labeled tokens is analyzed to detect named entities. Figure 1 shows a representation of the proposed framework.

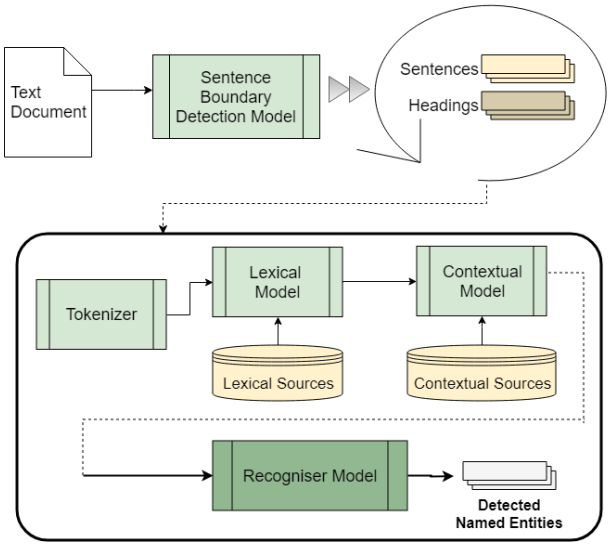

Figure 1. Proposed framework for the NER structure

3.2. Tokenizer and Tokens

Derived sentences of the input text file are first processed by tokenizer. Tokenizer scans through the input sentence and detects word boundaries and punctuation marks to get the list of tokens. A token can be a complete word, a punctuation mark or a morpheme after a punctuation mark. Tokens of a sentence are stored in a double linked list structure as a Token class object holds the information of previous and next tokens. A token object also holds a list of boolean variables that indicate states (labels). Labeling a token provides useful background information to be used while detecting named entities. Tokenizer applies the initial labeling on collected tokens. Considering the system requirements on

further stages, 15 tokenizer labels within four different categories are defined. As shown on Table 5, case, numeric, punctuation and location information are provided with labeling on this stage.

Table 5. Categorized tokenizer labels

Case Information	Numeric Information	Punctuation Information	Location Information
SW_CAPITAL	NUM	PUNCT_ APOSTR	BEFORE_ APOST
ALL_CAPITAL	ROMAN_NUM	PUNCT_ OTHER_MID	
EW_DOT	ORD_NUM	PUNCT_ OTHER_END	AFTER_ APOST
	DAY_NUM	PERCT	
	MONTH_NUM		
	YEAR_NUM		

- SW_CAPITAL: Indicates whether the token text starts with a capital letter or not.
- ALL_CAPITAL: Indicates whether all characters of the token text are capitalized or not.
- EW_DOT: Indicates whether the last character of the token text is a dot or not.
- NUM: If set to true, indicates that the token text denotes a numeric value.
- ROMAN_NUM: If set to true, indicates that the token text denotes a roman number.
- ORD_NUM: If set to true, indicates that the token text denotes an ordinal number.
- DAY_NUM: If set to true, indicates that the token holds a numeric value in [1,31] range.
- MONTH_NUM: If set to true, indicates that the token holds a numeric value in [1,12] range.
- YEAR_NUM: If set to true, indicates that the token holds a numeric value in [100,5500] range.
- PUNCT_APOSTR: Indicates whether the token text is an apostrophe character or not.
- PUNCT_OTHER_MID: Indicates whether the token text holds a punctuation mark used in the middle of a sentence; like comma, semi colon, parenthesis etc.
- PUNCT_OTHER_END: Indicates whether the token text holds a sentence ending punctuation mark (except dot) or not.
- PERCT: Indicates whether the token text is a percentage sign or not.
- BEFORE_APOST: If set to true, points that next token of the current token is an apostrophe.
- AFTER_APOST: If set to true, points that previous token of the current token is an apostrophe.

3.3. Lexical Model Sources

Lexical and contextual models are used to label tokens with additional information using lexicon structures. Lexicons used by lexical model indicates possible proper names (of a person or a location-region) except the auxiliary list which contains conjunctions.

- **TR_FirstNames:** Stores Turkish first names based on a database that contains Turkish Language Association (TDK) person names dictionary terms. Initial list holds 9699 elements; but the number is reduced to 9619 after some elimination, which is detailed in Section 3.3.1.
- **TR_CommonSurnames:** Stores a comprehensive list of Turkish surnames which are extracted from Wikipedia lists for Turkish actors-actresses, Turkish politicians (from 20th and 21st century), Turkish writers and Turkish commanders in Turkish War of Independence. Multiple occurrences of the same person (for example a politician who has served in both 20th and 21st century) and the duplicates of frequent surnames are eliminated. Final list contains 3039 elements.
- **FRGN_FirstNames:** Stores a list of foreign first names, derived from "the most influential people of all time" list published on *ranker.com*. This list consists of 2762 scientists, politicians, artists, athletes, philosophers etc. from different countries. Data is extracted as an XML file, then normalized to get plain lists of first names, surnames and mid names. Normalization phase includes the removal of prepositions or articles like "of, the", ordinal numbers, roman numbers and words that indicates a title or a nickname (like "St, Holy, Crazy, King, Queen, Baron, Prince, Princess). Duplicate occurrences of a name are also excluded. Final list contains 1489 elements.
- **FRGN_CommonSurnames:** Stores a list of foreign surnames. Foreign last names and mid names are also derived from the source list from *ranker.com*. Final list contains 1864 elements.
- **FRGN_MidNames:** Stores a list of foreign mid names like "de, von, bin" or shortened forms which is an initial upper-case letter trailed by a dot. Final list contains 34 elements.
- **Countries:** Stores the names of 193 member states of United Nations (UN), states consisting in these members (like England, Wales, Scotland, Northern Ireland) and self-governing states (like Puerto Rico, Virgin Islands, New Caledonia). Palestine, Taiwan and TRNC (Turkish Republic of Northern Cyprus) are the other states included. Additionally, some former country names that are likely to occur in historical texts (like Yugoslavia, USSR) are also included. Final list contains 257 elements.
- **TR_Cities:** Stores the names of 81 cities of Turkey and common different usages for them (like Afyon for Afyonkarahisar). Final list contains 86 elements.
- **TR_Districts:** Stores the names of districts of Turkey. Initial list holds 984 elements; after elimination of districts with same names and central districts named after their inclusive city, final list contains 897 elements.
- **FRGN_StatesCities:** Stores the names of capital cities of all countries and states-cities with high population or historical, touristic significance. Cities that are named after their countries are excluded and the final list contains 380 elements.

- **GeographicRegions:** Stores the names of continents or important geographic regions. The list contains 22 elements.
- **Conjunctions:** Stores conjunctions used in Turkish language. This auxiliary list is used to detect conjunction usage at the beginning of a sentence to avoid misleading NE detections.

3.3.1. Final Exclusions from Lexical Sources

Initial list taken from *ranker.com* includes some Turkish people like Mustafa Kemal Atatürk, Halide Edip Adıvar, Orhan Veli Kanık, Yunus Emre. This led some intersection between Turkish name lists and foreign name lists. 29 mutual words are detected between TR_FirstNames and FRGN_FirstNames lists; while 13 mutual words are detected between TR_CommonSurnames and FRGN_CommonSurnames lists. Leaving some of them on both lists are considered appropriate but some of them are excluded from one of the lists.

- Words like "Abdullah, Selma, Selman, Zakir" etc. are left on both lists.
- Words like "Edip, Evliya, Halide, Hamdi, Kemal, Mustafa, Orhan, Yunus, Ziya" etc. are excluded from FRGN_FirstNames list.
- Words like "Adam, Alan, Boy, Sun, San" etc. are excluded from TR_FirstNames list.
- Words like "Adıvar, Çelebi, Emre, Kanık, Pamuk, Atatürk, Tanpınar" etc. are excluded from FRGN_CommonSurnames list.
- Words like "Bradley, Reynaud, Spence" are excluded from TR_CommonSurnames list. These elements came from the names of Turkish people of foreign origin or married to a foreign person.
- In lexical sources, there also exists some overlap with contextual model sources. These overlapping words are excluded from lexical sources to give them their final forms.

3.4. Contextual Model Sources

Source lists used by contextual model indicates possible neighbor expressions for proper names. These expressions might or might not be in the NE text; their case information is mostly the criteria looked for this decision.

- **Before Person lists:** Stores words or word groups that might come before a person name. Four lists are used for this purpose. Lists include profession titles like "*Lord, Gazi" (Lord, Veteran)*, honorifics like *"Bay, Bayan" (Mister, Missis)*, abbreviations like "*Dr., Prof."* and mid-expressions like *"komutanı, padişahı" (commander of, sultan of)*.
- **After Person:** Stores profession titles in Turkish like *"Efendi, Hatun, Han, Paşa"* that possibly come after a person name.
- **After State or Country lists:** Stores words or word groups that might come after a state or country name. Two lists are used for this purpose. One list includes ending expressions like *"Krallığı, Cumhuriyeti" (Kingdom, Republic)*; other includes mid-expressions like *"başbakanı, imparatoru" (prime minister of, emperor of)*.
- **After Location:** Stores words or words groups that might come after a location name other than a state or country. The list includes expressions like *"belediye başkanı, Bölgesi, valisi" (mayor of, Region, governor of)*.
- **After Organization:** The list includes expressions like *"Derneği, Meclisi, Kurumu" (Association, Council, Institution)*.
- **After Geographical Formations:** The list includes terms like *"Gölü, Dağı, Irmağı" (Lake, Mountain, River)*. There also exists a list which holds possible expressions that a geographical formation might ends with in Turkish like *"ırmak, dağlar"* etc.
- **After Geographical Events:** The list includes terms like *"Depremi, Yangını" (Earthquake, Fire)*.
- **After Historic Events:** The list includes terms like *"Savaşı, Devrimi, İsyanı" (War, Revolution, Riot)*.
- **After Historic Buildings:** The list includes terms like *"Sarayı, Köprüsü" (Palace, Bridge)*.
- **Months:** Holds the names of the months.

3.5. Labeling by Lexical and Contextual Models

Tokenizer parses a sentence, generates tokens and initially labels them. Unlike tokenizer, lexical and contextual models should not label tokens one by one, as some lexicon terms might contain multiple words. So, tokens are passed to these models with n-grams. Initial token window width is defined as 4 and it decreases on every iteration until it reaches to zero. Multi-word lexicon terms are not missed and labeled correctly this way.

Table 6. Search patterns of a 7-token sentence for n-gram lexicon lookups.

N Value	Search Patterns
4	1234 – 2345 – 3456 – 4567
3	123 – 234 – 345 – 456 – 567
2	12 – 23 – 34 – 45 – 56 – 67
1	1 – 2 – 3 – 4 – 5 – 6 – 7

Tokens are labeled via n-gram lexicon lookups in lexical and contextual models to get their final forms before the execution of recognizer model. Table 7 shows the labels used in lexical and contextual models.

Table 7. Lexical (L) and Contextual (C) model labels

Model	Label Name	Description
L	LEX_TR_FN	Lexical term, Turkish first name
L	LEX_TR_LN	Lexical term, Turkish last name
L	LEX_FRGN_FN	Lexical term, foreign first name
L	LEX_FRGN_MN	Lexical term, foreign mid-name
L	LEX_FRGN_LN	Lexical term, foreign last name
L	LEX_CTRY	Lexical term, country name
L	LEX_TR_CITY	Lexical term, Turkish city name
L	LEX_TR_DIST	Lexical term, Turkish district name
L	LEX_FRGN_CITY	Lexical term, foreign city name
L	CONJ_SWC	Conjunction that starts with capital
L	NOT_LEX_SWC	Not a lexical term but starts with capital
C	B_PERSON	Before person expression
C	A_PERSON	After person expression
C	A_LOC_CTRY	After location-country expression
C	A_LOC_OTH	After location (other) expression
C	A_ORG	After organization expression
C	A_HIST_BLDG	After historic building expression
C	A_HIST_EVNT	After historic event expression
C	A_GEO_FORM	After geographic formation expression
C	A_GEO_EVNT	After geographic event expression
C	EW_GEO_FORM	Indicates a possible geographic formation with its ending
C	MONTH_NAME	Indicates a month name

Figure 2 shows a use case example of tokenization and token labeling with three different models on the sentence *"Dünya'da 23 Eylül günü, Türkiye Cumhuriyeti'nde ve tüm Kuzey Yarım Küre'de sonbahar başlar." (On the day of 23 September in the world, it is the beginning of autumn in Turkey and the whole Northern Hemisphere.)*. Token labels from different models are shown with different colors.

3.6. Named Entities and Recogniser Model

As developed NER system is specialized for lecture notes in the scope of history and geography courses, extent of a NE is adjusted to meet the requirements. 13 NE types are defined, which are explained on Table 8.

After token derivation and labeling is completed, recognizer is executed to find out named entities. System can both be tested on a single sentence or a complete text document. Figure 3 shows a use case example where the system is tested with the input sentence *"Bornova Anadolu Lisesi ve İzmir Atatürk Lisesi öğrencileri, Cumhuriyet Bayramı'nı kutlamak için Gündoğdu Meydanı'nda toplandı." (Students of Bornova Anatolian High School and İzmir Atatürk High School are gathered in Gündoğdu Square to celebrate Republic Day.)*. Execution resulted in four NE detections. Tokens *"Bornova, İzmir, Atatürk, Gündoğdu"* are all lexicon terms and might be named entities on their own in different sentences. On the example sentence though, these terms are correctly found to be parts of longer named entities. System is designed to consider the container named entities instead of single lexicon terms in such circumstances.

Table 8. Defined NE types (13 tags)

Named Entity Type	Description
Person_Turkish	Indicates a Turkish person
Person_Foreign	Indicates a foreign person name
Location_State_Country	Indicates a country, state, continent or geographic region
Location_Other	Indicates a city or district
Historic_Term_Building	Indicates a historic building or structure
Historic_Term_Event	Indicates a historical event
Geographic_Term_Formation	Indicates a specific geographical formation
Geographic_Term_Event	Indicates a specific geographical event such as a natural disaster
Organization	Indicates an organization within a wide range of fields (politics, education, military, media, law, medical etc.)
Percentage	Indicates a percentage or fraction expression
Date	Indicates a single date expression in multiple formats or a date range expression.
Date_or_Number	Indicates a clock expression or a numeric value below 1200 or above 2000.
Other	Indicates a detected NE which is not classified as one of the distinctive types.

4. Research Findings

4.1. Experimental Results

Success of the system is tested via experiments on actual lecture notes. 30 history and 30 geography documents are selected for this task. Precision and recall metrics for TEXT (to correctly detect borders of the NE) and TYPE (to correctly detect type of the NE) attributes are used for evaluation. Experiments for geography domain and history domain are separated to allow comparisons; conclusive results are calculated by combining these two experiment sets. Detected NE types are also counted among correctly guessed type values to compare distributions between different domains.

Precision values are calculated by dividing number of correct guesses to number of all detections; recall values are calculated by dividing number of correct guesses to number of actual named entities. Evaluation metrics used on experiments are formulated below on Equation 1,2,3 and 4.

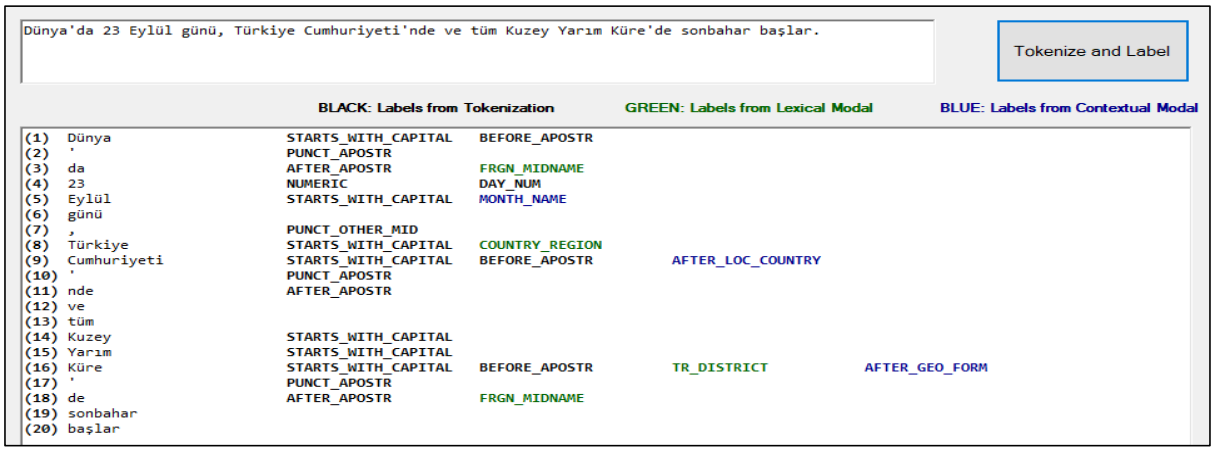

Figure 2. Example system usage to show tokenization and token labeling applied on an input sentence.

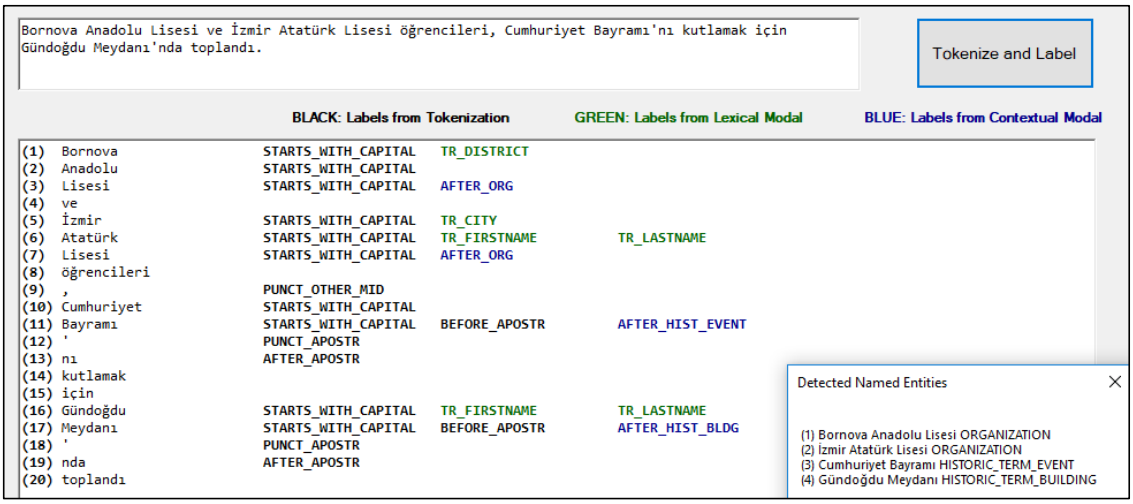

Figure 3. Example system usage to show NE detections on an input sentence.

$$\text{Precision TEXT (\%)} = \frac{100\ (\#\ \text{of Correct TEXT})}{\#\ \text{of Detected NE}} \quad (1)$$

$$\text{Precision TYPE (\%)} = \frac{100\ (\#\ \text{of Correct TYPE})}{\#\ \text{of Detected NE}} \quad (2)$$

$$\text{Recall TEXT (\%)} = \frac{100\ (\#\ \text{of Correct TEXT})}{\#\ \text{of Actual NE}} \quad (3)$$

$$\text{Recall TYPE (\%)} = \frac{100\ (\#\ \text{of Correct TYPE})}{\#\ \text{of Actual NE}} \quad (4)$$

Table 9 and 10 shows the experiment results for history and geography course text files and Table 11 shows the combined results.

Actual number of named entities are determined before performing the experiments. 30 history documents contain 1654, 30 geography documents contain 991 named entities, which makes a grand total of 2645 named entities on 60 documents. Average number of named entities per document is calculated as 55.13 for history domain, 33.03 for geography domain and 44.08 for the combined dataset.

NE type distributions on the test documents are also determined before experimentation. On 30 history documents, there exist 133 Person (Turkish), 48 Person (Foreign), 273 Location (State/Country), 126 Location (Other), 101 Organization, 9 Historic Term (Building), 127 Historic Term (Event), 39 Geographic Term (Formation), 221 Date, 26 Date or Number, 5 Percentage and 546 Other tagged named entities. It is observed that no NE with Geographic Term (Event) tag exists on these documents.

On 30 geography documents, there exist 8 Person (Foreign), 225 Location (State/Country), 200 Location (Other), 4 Organization, 3 Historic Term (Building), 3 Historic Term (Event), 209 Geographic Term (Formation), 27 Geographic Term (Event), 47 Date, 62 Date or Number, 20 Percentage and 183 Other tagged named entities. It is observed that no NE with Person Name (Turkish) exists on these documents.

Experiments on history course text files resulted in 96.06% precision for TEXT, 92.67% precision for TYPE, 95.83% recall for TEXT and 92.44% recall for TYPE. Experiments on geography course text files resulted in 96.59% precision for TEXT, 93.37% precision for TYPE, 97.07% recall for TEXT and 93.84% recall for TYPE.

Table 9. Experiment results for history course text files (Particular results of first 5 documents are included)

DOC NAME	# of Actual NE	# of Detected NE	# of Correct TEXT	# of Correct TYPE	# of Missed NE	Precision TEXT (%)	Precision TYPE (%)	Recall TEXT (%)	Recall TYPE (%)
1. Bayezid Dönemi	71	72	70	68	1	97,22	94,44	98,59	95,77
1. Dünya Savaşı Öncesi Gelişmeler	69	66	64	63	5	96,97	95,45	92,75	91,30
1. Dünya Savaşı	50	48	47	46	3	97,92	95,83	94,00	92,00
1. Meşrutiyet	34	34	33	31	1	97,06	91,18	97,06	91,18
2. Dünya Savaşı'nın Nedenleri, Gelişimi	47	48	46	45	1	95,83	93,75	97,87	95,74
TOTAL	**1654**	**1650**	**1585**	**1529**	**69**	**96,06**	**92,67**	**95,83**	**92,44**
AVG	**55,13**	**55,00**	**52,83**	**50,97**	**2,30**				

Table 10. Experiment results for geography course text files (Particular results of first 5 documents are included)

DOC NAME	# of Actual NE	# of Detected NE	# of Correct TEXT	# of Correct TYPE	# of Missed NE	Precision TEXT (%)	Precision TYPE (%)	Recall TEXT (%)	Recall TYPE (%)
Akarsu Havzalarımız	34	33	32	31	2	96,97	93,94	94,12	91,18
Aktif Nüfusun Ekonomik Faaliyet Gruplarına Göre Dağılımı	14	14	14	14	0	100,00	100,00	100,00	100,00
Basınç Çeşitleri ve Özellikleri	36	33	33	33	3	100,00	100,00	91,67	91,67
Başlıca Kıyı Tipleri	29	29	28	27	1	96,55	93,10	96,55	931,10
Bölgeler Coğrafyası – Akdeniz Bölgesi	21	22	20	20	1	90,91	90,91	95,24	95,24
TOTAL	**991**	**996**	**962**	**930**	**25**	**96,59**	**93,37**	**97,07**	**94,84**
AVG	**33,03**	**33,20**	**32,07**	**31,00**	**0,83**				

Table 11. Combined experiment results for 60 course text files.

DOCUMENTS	# of Actual NE	# of Detected NE	# of Correct TEXT	# of Correct TYPE	# of Missed NE	Precision TEXT (%)	Precision TYPE (%)	Recall TEXT (%)	Recall TYPE (%)
HISTORY Documents (30)	1654	1650	1585	1529	69	96,06	92,67	95,83	92,44
GEOGRAPHY Documents (30)	991	996	962	930	25	96,59	93,37	97,07	93,84
TOTAL	**2645**	**2646**	**2547**	**2459**	**94**	**96,26**	**92,93**	**96,29**	**92,97**
AVG	**44,08**	**44,10**	**42,45**	**40,98**	**1,57**				

Combined results are **96.26%** precision for TEXT, **92.93%** precision for TYPE, **96.29%** recall for TEXT and **92.97%** recall for TYPE.

Results show that success rate for geography domain is slightly better than history domain. But the fact that average number of NEs in a history document is way higher than average number of NEs in a geography document (more than 22) should not be avoided. In both domains, accuracy on TEXT resulted to be higher than accuracy on TYPE, for both precision and recall metrics. Main reason for this is, when the boundaries of a NE is not correctly distinguished, predicting the type of this incorrect text turns out the be an unfeasible task. Ambiguous lexicon terms and person names that can also be used as common nouns are two other issues that cause erroneous detections.

An analysis to detect success rate of the model for individual NE types is also made on experiment results. Table 12 compares number of correctly detected NEs for each type with the actual number in history and geography domains, also in the combined test set with 60 documents. For each NE type, average numbers of detected and actual NEs in 60 documents are also included. Accuracy (Acc) value for each NE type *t*, which is formulated on Equation 5 is used for evaluation. System success at detecting NEs with Percentage, Date, Location (State/Country), Historic Term (Event) and Other types reached highest accuracy values with 100%, 98.88%, 96.79%, 93.85% and 92.87% respectively. Lowest accuracy value among 13 NE types is observed on Geographic Term (Event) with 88.89% (24 out of 27).

$$\text{Acc}_t(\%) = \frac{100\ (\#\text{of NEs Correctly Detected as } t)}{\#\text{ of Actual } t \text{ typed NEs}} \qquad (5)$$

Distribution of correctly detected NE types for both domains is also shown on Table 12. Other, Location (State/Country), Date, Person (Turkish) and Historic Term (Event) are the five most occurred NE types for history documents. Location (State/Country), Location (Other), Geographic Term (Formation), Other and Date or Number are five most occurred

Table 12. Experiment results for individual NE types.

DOCUMENTS	Detected / Actual	Person Turkish	Person Foreign	Location State/Country	Location Other	Organization	Historic Term Building
HISTORY Docs (30)	Detected	121	43	259	114	90	8
	Actual	133	48	273	126	101	9
GEOGRAPHY Docs (30)	Detected	0	7	223	185	4	3
	Actual	0	8	225	200	4	3
TOTAL (60 docs)	**Detected**	**121**	**50**	**482**	**299**	**94**	**11**
	Actual	**133**	**56**	**498**	**326**	**105**	**12**
AVG	**Detected**	**2,02**	**0,83**	**8,03**	**4,98**	**1,57**	**0,18**
	Actual	**2,21**	**0,93**	**8,30**	**5,43**	**1,75**	**0,20**
Accuracy (%)		**90,98**	**89,28**	**96,79**	**91.71**	**89,52**	**91,67**

DOCUMENTS	Detected / Actual	Historic Term Event	Geo Term Formation	Geo Term Event	Date	Date or Number	Percentage	Other
HISTORY Docs (30)	Detected	119	37	0	218	25	5	502
	Actual	127	39	0	221	26	5	546
GEOGRAPHY Docs (30)	Detected	3	185	24	47	55	20	175
	Actual	3	209	27	47	62	20	183
TOTAL (60 docs)	**Detected**	**122**	**222**	**24**	**265**	**80**	**25**	**677**
	Actual	**130**	**248**	**27**	**268**	**88**	**25**	**729**
AVG	**Detected**	**2,03**	**3,70**	**0,40**	**4,42**	**1,33**	**0,42**	**11,28**
	Actual	**2,16**	**4,13**	**0,45**	**4,46**	**1,46**	**0,42**	**12,15**
Accuracy (%)		**93,85**	**89,52**	**88,89**	**98,88**	**90,91**	**100**	**92,87**

types for geography documents. Absence of any Person (Turkish) tagged NE in geography domain and absence of any Geographic Term (Event) tagged NE in history domain are remarkable results. Location (State/Country) appears to be the most homogenously spread tag among the complete experiment set.

4.2. Encountered Challenges

Problems and restrictions, mostly in connection with Turkish language or common violations in input documents are encountered during the development process.

Using a wide Turkish first name lexicon provides a high recall in detecting person names; but it is possible to lead decreases in precision. This is because of the nature of Turkish, as some of the person name words might also indicate common nouns that are frequently used in lecture notes like *"Savaş (War), Barış (Peace), Nehir (River), Irmak (River)"*. Neighbor token controls mostly avoid erroneous detections when these terms are in the beginning of a sentence, controls for neighbor tokens. In some conditions, these controls aren't single-handedly enough. For example, CONJ_SWC lexicon is also beneficial when the first word of a sentence is a conjunction and followed by a NE.

Some expressions like *"Sultan, Şah" (Sultan, Shah)* in contextual model might occur both before or after a person name; it is also possible for two conditions to occur at the same time, for example *"Kanuni Sultan Süleyman" (Suleiman the Magnificent)*. System used to detect two different named entities in these situations (as *"Kanuni Sultan"* and *"Sultan Süleyman"*); then this is corrected and detected partial expressions are merged to reach the correct NE.

Heading texts are handled with additional controls, as traditionally all heading words (except conjunctions) starts with a capital; even it doesn't indicate a proper noun. This caused to limit the usage of "Other" tag for a named-entity and raised the significance of apostrophe controls.

Separating a commonly used "Person" NE type into two (Person_Turkish and Person_Foreign) seems to cause TYPE mistakes in some occasions (which wouldn't happen if two types are merged as a single Person type). Especially because some first names used in Turkish like *"Musa, Enver, Zeynel, Süleyman"* are also common in Arab countries. Experiments show the performance drops are acceptable though; as differentiating Turkish and foreign person names is an important property for the further usage.

Absence of required punctuation marks (most frequently apostrophe and comma) and spelling errors on input text documents also has negative impacts on system success. It also decreases the quality of detected named entities and leads to an increased number of "Other" tagged named entities. Applying a spell check operation on the document before submitting it as an input is highly recommended.

5. Results and Evaluation

In this study a rule-based NER model is developed for Turkish lecture notes in the scope of history and geography courses. System is designed to take a text document as input and derive named entities within the context of document as output. While the study is in the scope of natural language processing, information extraction and text mining and computational linguistics; it can also be considered as a computer aided education software.

The model is developed as a Windows Forms Application in Microsoft .NET Visual Studio 2017 environment by using .NET framework 4.6.1 and C# programming language.

Success rate of the system is tested via experiments on actual lecture notes. 30 history and 30 geography documents are selected for this task. Precision and recall metrics for TEXT and TYPE attributes are used for evaluation. Final results are calculated as **96.26%** precision for TEXT, **92.93%** precision for TYPE, **96.29%** recall for TEXT and **92.97%** recall for TYPE.

System success for individual NE types is also observed. Based on experiment results, **90.98%** for Person (Turkish), **89.28%** for Person (Foreign), **96.79%** for Location (State/Country), **91.71%** for Location (Other), **89.52%** for Organization, **91.67%** for Historic Term (Building), **93.85%** for Historic Term (Event), **89.52%** for Geographic Term (Formation), **88.89%** for Geographic Term (Event), **98.88%** for Date, **90.91%** for Date or Number, **100%** for Percentage and **92.87%** for Other are the calculated accuracy values.

5.1. Educational Value

Considering the primary goals defined, experiment results are mostly satisfactory and developed NER model is proved to be a suitable auxiliary tool for the long-term educational goal, constructing steady glossary of terms for history and geography domains. As 13 types are defined for a NE, the model also proposes a suggestive taxonomy for detected terms, instead of a broad classification as "geographic term" or "historic term". Constructed glossary of terms structures are projected to support a test generation process, as each stored term spans an information content and suitable for question texts.

5.2. Future Enhancement

Decreasing the number of named entities with "Other" tag should be considered by additional NE types. For example, a large portion of these kind of named entities in history documents have a "nation, nationality" meaning; which can be encapsulated with a different tag usage. Lexicons can also be extended with ancient age location and person names. A spell-checker module can be integrated to the system to minimize negative effects of the absence of punctuation marks.

Acknowledgements

This research paper is prepared within the project that is supported by Dokuz Eylül University Department of Scientific Research Projects (DEÜBAP) with number 2018.KB.FEN.015.

Conflict of Interest

No conflict of interest was declared by the authors.

References

Alfonseca, E., Manandhar S. (2002). "An unsupervised method for general named entity recognition and automated concept discovery". In 1st International Conference on General WordNet.

Cucerzan, S., Yarowsky, D. (1999). "Language independent named entity recognition combining morphological and contextual evidence". Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora. New Brunswick, NJ: Association for Computational Linguistics.

Ertopçu, B., Kanburoğlu, A., Topsakal, O., Açıkgöz, O., Gürkan, A., Özenç, B., Çam, İ., Avar, B., Ercan, G., Yıldız, O. (2017). "A new approach for named entity recognition". In: International Conference on Computer Science and Engineering (UBMK), Antalya, Turkey, 2017.

Grishman, A., Sundheim, B. (1996). "Message Understanding Conference-6: a brief history". In Proceedings of the 16th conference on Computational linguistics - Volume 1 (COLING '96), Vol. 1. Association for Computational Linguistics, Stroudsburg, PA, USA, 466-471.

Jurafsky, D., Martin, J.H. (2009). "*Speech and language processing (2nd Edition)*". Prentice-Hall, Inc., Upper Saddle River, NJ, USA.

Küçük, D., Jacquet, G., Steinberger, R. (2014). "Named entity recognition on Turkish tweets". In: Language Resources and Evaluation Conference, 2014.

Küçük, D., Küçük, D., Arıcı, N. (2016). "A named entity recognition dataset for Turkish". In: 24th Signal Processing and Communications Applications Conference (SIU), Zonguldak, Turkey, 2016.

Küçük, D., Yazıcı, A. (2009). "Named entity recognition experiments on Turkish texts". In Proceedings of the 8th International Conference on Flexible Query Answering Systems, FQAS '09, pages 524–535, Berlin, Heidelberg. Springer-Verlag.

Küçük, D., Yazıcı, A. (2009). "Rule-based named entity recognition from Turkish texts". In Proceedings of the International Symposium on Innovations in Intelligent Systems and Applications, Trabzon, Turkey. pages 456–460.

Küçük, D., Yazıcı, A. (2012). "A hybrid named entity recognizer for Turkish with applications to different text genres". In: Gelenbe E., Lent R., Sakellari G., Sacan A., Toroslu H., Yazici A. (eds) Computer and Information Sciences. Lecture Notes in Electrical Engineering, vol 62. Springer, Dordrecht.

Sang, E., Meulder F. (2003). "Introduction to the CoNLL-2003 shared task: language-independent named entity recognition". In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003 - Volume 4* (CONLL '03), Vol. 4. Association for Computational Linguistics, Stroudsburg, PA, USA, 142-147

Şeker, G. A., Eryiğit, G. (2012). "Initial explorations on using CRFs for Turkish named entity recognition". In Proceedings of COLING 2012, Mumbai, India, 8-15 December.

Şeker, G., Eryiğit, G. (2016). "State of the art in Turkish named entity recognition". May 2018. Retrieved from https://pdfs.semanticscholar.org/7e7f/ed9d21a3e3a36c4eb3c7df1ee8116e8ec2ce.pdf

Tatar, S., Çiçekli, İ. (2011). "Automatic rule learning exploiting morphological features for named entity recognition in Turkish". *Journal of Information Science,* 37 (2), April 2011, 137-151.

Tür, G., Hakkani-Tür G., Oflazer K. (2003). "A statistical information extraction system for Turkish". Natural Language Engineering, vol. 9 (2), pp. 181-210.

Wentland, W., Knopp, J., Silberer, C., Hartung, M. (2008). "Building a multilingual lexical resource for named entity disambiguation, translation and transliteration". in Proceedings of the 6th International Conference on Language Resources and Evaluation, Marrakech, Morocco, 26 May–1 June 2008.