Developing Arabic Event Extraction SystemUsing Rule Based Method

Razieh Baradaran, Department of information technology, university of Qom, Qom, Iran

R.baradaran@stu.qom.ac.ir

Behrouz Minaei-Bidgoli, Department of Computer Engineering, Iran University of Science and Technology, Tehran, Iran

B_minaei@iust.ac.ir

Name of the Presenter: Razeih Baradaran

Abstract

Event extraction is one of the most usefull and challenging information extraction tasks that is considered as foundation for many natural language processing applications. In this study we have presented a modular approach for event extraction from historical Arabic text with complex linguistic structure and completely unstructured format.

Presented approach has 4 main stages consist of preprocessing, event mention detection, gramatical relation detection and argument extraction. Argument extraction task has Independent subtasks for extracting each event argument.

Key words: Information Extraction, Event Extraction

1. Introduction

Nowadays, large volume of data is available and information extraction as abranch of science is faced with sinificant development.Information extraction is process of extracting prespecified information from raw texts. Events are One of the most popular and complex information which can be extracted from text. Event extraction as information extraction subtask,extracts events and their participants from text. These participants can be people, organization, Date, time and other related information.

According to ACE[1] standard the following concepts are defined:

Event: something that happens in the specific time and place like occurrence of a particular crime or incident.

Event trigger: the main word which most clearly expresses an event occurrence

[1]Automatic Content Extraction

Event argument: the mentions that are involved in an event (participants)

Event mention**:** a phrase or sentence within which an event is described, including triggers and arguments

For extract events from text, first step is determining event mention. Event mention (usually is a sentence) consist of event trigger and zero or more argument. In second step, we must extract event arguments such as event time, event date and other involved participants.

Information extraction history back at least past three decades while event extraction is a complex and relatively new research field. Event mention can be several sentences or phrases and this is one the complexity of event extraction. Also an event has arguments and attributes that must be identified in its event mention.

Automatic extraction of events from text has been started by Topic Detection and Tracking (TDT) project in 1998 by the National Institute of Standards and Technology (NIST).This project develops technologies to extract events from news stream and track the progression of these events over the time.

Then, information extraction was followed with Automatic Content Extraction (ACE) Program in 1999. This program develops technologies that can automatically process and extract desired information from natural language documents. This information was consisted of entities, relations and event that were gradually added to the ACE program (first entities, then relations and finally events). ACE program defines separate tasks to identify any of this information.

In ACE program, eVent Detection and Recognition (VDR) task, consists of detecting eight predefined different types of events that each of event type has subtypes. After detecting event type, event subtype must also be specified.

Each event has some arguments which plays specific semantic rule. For example, an attack event has arguments with agent, place, time, number of injuries, number deaths and other semantic rules.

It can be said ACE program is a way to standardize information extraction tasks, which is used to evaluate other systems.

In addition to two mentioned systems, more researches have been done in this area, which some use statistical and data oriented methods to event extraction such as Naughton (2008) that uses support vector machine classifier and language modeling approach to classify sentences as on-event or off-event. Some other researchers, duo to higher accuracy of rule base systems, use knowledge oriented approach and pattern matching techniques.

REES[2] is a knowledge based event and relation extraction system that is presented by Aone and Ramos-Santacruz (2000). REES extracts events by use of lexico-syntactic patterns. The result shows that f-measure standard for 26 event types is about 70 percent and for 37 event types is about 60 percent. The authors stated most fauilor system in event extraction is due to lack of event information in lexicon. Also this system doesn't able to detecting noun triggers and only detect verb triggers.

[2]Relation and Event Extraction System

In other research, Vargas-VeraandCeljuska (2004) are presented a news event extraction system. This system uses a set of lexico-semantic patterns to extract events and arguments from news reports. It compares each events related patterns to news reports, if match accured, the event information are extracted.the precision and recall are respectively, 68 percent and 52 percent. Using confidence value in rule selection, precision and recall are respectively changed into 90 percent and 14 percent. That low recall in second condition, reduces system validity.

Xu and et al. (2006) use frequent seeds in train document, and tages these seeds in other texts by SProUT[3] system and then create extracting rules using MINIPR systems.

Abuleil (2007), extract events in Arabic news document using natural language techniques. He use 300 passages from Aljazeera.net for train and test dataset(150 passages for train and 150 passages for test). In this system first triggers are detected using searching a set of keywords in text. In This research each events are breaked into their elements and are analized to identify the role each plays in the event. Finally the precision and recall are respectively, 92 percent and 87 percent.

Weaknesses of this system are consists of that it cannot detect co-reference events and presence of all event components is necessary, while lack of some components in natural language documents so happen.

Most event extraction systems support English texts and and in many languages like Arabic, a research need to be felt. In adition Arabic language is an ancient language with many historical documents that have valuable information for researchers. So requires systems to extract information automaticly.

In this research, we implement a system for event extraction in historical Arabic texts. According to Specific linguistic characteristics and more complex linguistic rules of Arabic literature rather than others, event extraction in this language is more complex. In addition historical Arabic text has relatively different structure from modern Arabic text that increases the complexity.

We implementour event extraction system with a modular approach. So event information extracted independently and not affect on the other extraction process.

The outline of the paper is as follows: in section 2 describes Data which is used in our study.Section 3, introducesour event extraction system and their stag. Section 4 explains experiments and their results; and in the last section we conclude the paper and state future orientation.

2. Data and Material

We use two dataset in this research. First dataset is Tabari history of Islam book Vol.1 raw text, which is used for sentence classification and specifying on-event and off-event sentences. Then we use Computer Research Center of Islamic Sciences, NOOR co. dataset for argument extraction phase, which includes about 12000 die event paragraphs. This dataset contains Arabic historical books texts such as Muruj adh-dhahab, altabaghat alkobra, Ya'qubi, Ansab al-Ashraf and etcetera.

3. Research Methodology

[3]Shallow processing with Unification and Typed feature structures

In this section, we describe our system architecture and their components

1. System Architecture

System architecture as shown in figure 1 is consists of 4 main stage. Each stage is described in the following:

1.1 Preprocessing Step

Data preprocessing is a necessary step in most information extraction systems. In this step data is converted to appropriate format required for information extraction process. The first step in our system is also rawtext preprocessing.

Raw text preprocessing is include tokenizing, stemming, POS[4] tagging, noun and verb group detecting (Base Phrase Chunking) and name entity recognizing. However Base Phrase Chunking and name entity recognizing are only used in argument extraction and are not used in sentence classification as on-event or off-event.

1.2 Sentence Classification Step

In this step, we classify sentences as on-event or off-event sentences. We apply rule base approach and a set of trigger keywords which are provided by linguistic expert team. If a sentence contains one of keywords and matches with related morphological characteristics,it marks as on-event sentence.

In this system, an event trigger may be presented by more than one word. In this case represented words may be separated by other words within event mention. Also, it may be a verb or Noun trigger.

The step inputs are set of trigger keywords, input text with POS and stem tags. The output is tagged sentences as on-event or off-event.

1.3Shallow parsing step

The inputs of this step are input text, noun and verb group tags and NER tags. Shallow parsing module is a syntax analyzer and determine subject, verb and object in sentences. This module uses a rule base method and maps grammatical rules like Hammadi (2012) rules to extract the triple subject, verb and object in each instance.

1.4 Argument Extraction Step

Argument extraction step uses a rule base method and prior step output to extract events argument according to ACE standard definition. In ACE standard some arguments are general event attributes which apply to most events like event time and event place.And some arguments are specific event attributes and apply to a specific event like victim is a specific argument for die event.

In this system, each argument type extractsindependently and has specific rules, so lack of some information in common text sentences, don’teffect on followed extraction process. The process of each argument type extraction is as follows:

- Event Time Extraction

[4] Part Of Speech

For extracting time role in an on-event instance,we first specify time phrases using time keywords and pattern matching technique. Then we match event time extraction rules to extract time role.

- Event Place Extraction

Place is also general argument like time that must be extracted in most event types. We use NER tags and event place extraction rules to specify event place. Also as an addition way we use determinant particle like "فی" ("in" in English) that followed by places(في "in" البصره + "Basra") to specify places and increase extraction accuracy.

- Event Participants Extraction

We use syntactic labels that create in prior step and apply related rules to extract other event roles like victim and agent. In case of failure of this module duo to prior step errors, we use NER and POS tags to specify sentence entities and map another rules to extract related roles.

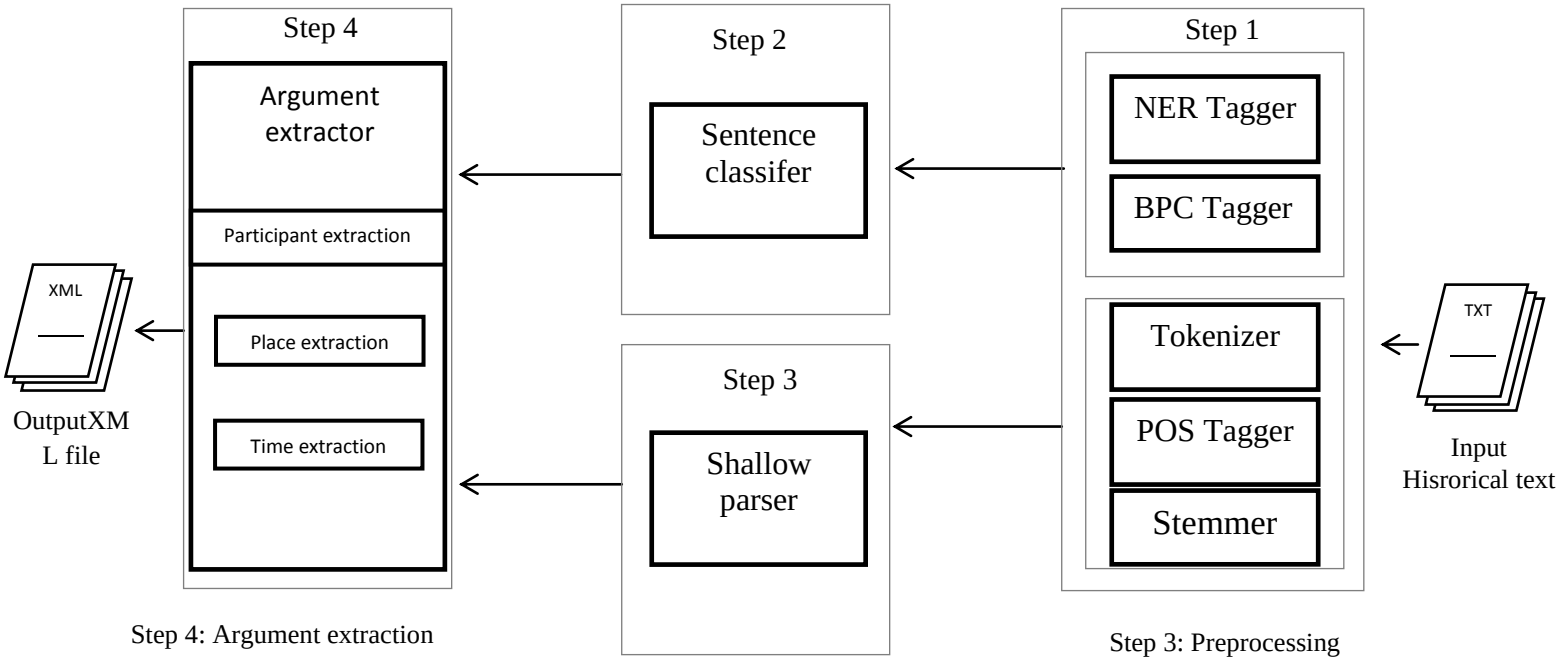

Fig 1: Event Extraction System Architecture

2 Results and Analysis

We use AMIRA tool for tokenizing, POS tagging and Base Phrase Chunking.also KHOJA stemmer is used to specify words stems.Noor ANER system that implemented by

Asgari Bidhendi (2012) is usedfor tagging name entities in text with some setting for tagging places.

System implementation and rule mapping was done using c# programming.

As mentioned in previous section we use two dataset for evaluating our system performance.For testing sentence classification module we use Tabari history of Islam book Vol.1 raw text. We evaluate our experiment using standard performance measures metrics precision, recall and f-measurePerformance results are shown in table 1. As the results show precision and recall are respectively 91.5 and 86.1for detecting on-event sentences which are good result for an event extraction system.

	Class on-event	Class off-event
Precision	91.5	98.2
Recall	86.1	98.2
F-measure	88.72	98.59

Table1. Performance results in sentence classifiction step

For illustration some of instance that system specify them as on-event are represented in following. Underlined words are detected as triggers:

Example 1:

"قال ابن إسحاق: و استشهد يوم أجنادين ممن حفظ عنه الحديث نعيم بن عبد الله النحام العدوي، و هشام بن العاص بن وائل السهمي."

"Ibn Ishaq said: Naim bin Abdullah al-Noham al-Adwiwho was saved him narrative, and Hisham ibn al-Aas bin Wael al-Sahmi killed on Ajnadeen day."

In this example, "استشهد" word means "martyrdom" that represent die event and our program tag this paragraph as on-event, because according applied rules, a verb with "شهد" stem in this linguistic format, represent a die event.

Example 2:

"و حدثني عبّاس بن هشام الكلبي عن أبيه عن عوانة و غيره قالوا: لمّا حضرت معاوية الوفاة و ذلك في سنة ٧٦٥ ستّين كان يزيد غائبا، فدعا معاوية الضحّاك بن قيس الفهري، و كان على شرطه، و مسلم بن عقبة بن رباح بن أسعد المرّي، فأوصى إليهما."

"Abbas bin Hisham al-Kalbi told me from his father from Awana and other that they said: When Muawiya death was arrived in the year 765 Yazid was absent, Muawiya called Dahhaak bin Qais al-Fihri, that he was police, and Muslim bin Uqbah ibn Rabah ibn As'ad al-Marri, then he advised them."

In this example, "حضرت الوفاة"(although they are separated) represent die event; because when a verb with "حضر" stem is placed in sentences and after it with limited distance words such as "موت" or "وفاة" is placed, represents die event.

Example 3:

"قال: و توفي محمد بن الفضيل بالكوفة سنة خمس و تسعين و مائة و شهد جنازته وكيع بن الجراح.

"Said: Mohammad Ibn al-fazil died in Kufa in one hundred and ninety five year and saw his funeral Wakia bin al-jarah"

In above example, there are two triggered words, although if one of them was been occurred, was enough for tagging on event paragraph.

As you have seen in the last two examples, this program don't relies on one word triggered or verb triggered and phrases such as "حضرت الوفاة" (dying) is also considered as well as noun words such as "جنازة (his funeral).

For extracting argument we useComputer Research Center of Islamic Sciences, NOOR co. dataset which is large volume dataset contain about 12000 die event paragraph. Several events may be occured in one paragraph. Performance results for argument extraction module in 420 paragraphs are shown in table 2.This module extract argument withoverall precision 80.51 and recall 71.29for. However A large numbers of error are related to prior step errors like POS tagger, stemmer and NER tagger that have affected on our results.

Argument	Precision (%)	Recall (%)	F-measure (%)
agent	88.64	70.27	78.89
victim	81.57	64.81	72.23
place	57.3	42.5	48.8
time	81.57	84.12	82.82

Table2. Performance results in argument extraction step for die event

	Precision (%)	Recall (%)	F-measure (%)
Average results	77.27	65.42	70.85
Overall results	80.51	71.29	75.62

Table3. Average and overal performance results in argument extraction step

In following some output instances of argument extraction module are shown. Underlined words are extracted information:

Example 4:

ولدأبانبنصالحسنةستينوماتبعسقلانسنةبضععشرةومائةوهوابنخمسوخمسينسنة.

"Aban bin SalehBorned in sixtieth year and died in middle of one hundered and ten year and he was fifty five years old."

Trigger: مات, Victim:أبانبنصالح, Agent:, Place:عسقلان, Time:سنةبضععشرةومائةوهوابنخمسوخمسينسنة

Example 5:

قالمحمدبنعمر: توفيأبوبكربنمحمدبنعمروبنحزمبالمدينةسنةعشرينومائةفيخلافةهشامبنعبدالملك.

"Mohammad bin Omar said: Abu Bakr bin Muhammad bin Amr bin Hazm Died in Medina in Hundred and twentieth year in succession Hisham bin Abdul Malik."

Trigger: توفي Victim:أبوبكربنمحمدبنعمروبنحزم , Agent:, Place:المدينة, Time:سنةعشرينومائةفيخلافةهشامبنعبدالملك

Example 6:

حضرتالأرقم بن أبي الأرقمالوفاة و ذلك سنة خمس و خمسين

"Al-Argham bin Abi al-Argham'sdeath was arrived in Fifty-fifth year."

Triggerحضرتالوفاة: Victim:الأرقمبنأبيالأرقم, Agent:, Place:, Time:سنةخمسوخمسين

Example 7:

السلوليمولىلهم. ماتسنةخمسومائتينبالكوفةفيخلافةالمأمون

"Al-Salouly was their sire.HeDied in two hundred and fifth year in Kufa in succession Ma`moun"

Trigger: مات Victim:السلولي , Agent:, Place: الكوفة , Time:سنةخمسومائتينفيخلافةالمأمون

In this example, time argument consist of two parts which seperated by anotherwords.

Example 8:

وفيهاماتسليمانبنعليبنعبداللهبنعباس،وأبانبنتغلبوسعدبنسعيدابنقيسأخويحيىبنسعيد.

Trigger: مات Victim:سليمانبنعليبنعبداللهبنعباس،وأبانبنتغلبوسعدبنسعيدابنقيسأخويحيىبنسعيد, Agent:, Place:, Time:

"And Sulaiman bin Ali bin Abdullah bin Abbas, and Aban bin Taghlab and Saad bin Saeed Ibn Qais brother of Yahya bin Saiddied"

In this example, one argument such as victim can consist of several instances. There is only victim role and another roles dont appear in sentence.

Example 9:

أبوأسيرةبنالحارثبنعلقمة،منبنىمبذولبنعمروبنغنمبنمازن،قتلهخالدبنالوليد

"Abu Asirah bin al-Harith bin Alqamah, ofth tribe of Mabdhul bin Amr binGhanam bin Mazen, was killed byKhalid bin al-Walid"

Trigger: قتله, Victim:(أبوأسيرةبنالحارثبنعلقمة)مرجعضميرمتصلبهقتله, Agent:خالدبنالوليد, Place:, Time:

Our system can able to detect pronoun reference such as victim in above example.

Example 10:

حدثنا علي بن عاصم عن حصين عن عمرو بن جاوان عن الأحنف قال: لما انحاز الزبيرقتلهعمرو بن جرموز بوادي السباع.

“Ali ibn Asim told usof Hasin of Amr bin Jawan of Ahnaf that he said: When Zubairdefeated, Amr ibn Jermozkilled himin wide animal place“

Trigger: قتله Victim:(الزبير) مرجع ضمير متصل به قتله,Agent:عمرو بن جرموز,Place:, Time:,

In this example place argument is not specified. This error is caused by NER error in tagging this place as place entity and tokenizer error in separating some preposition like “ب”in above example.

3 Conclusions

Event extraction is one of the most useful and challenging information extraction tasks and there are high research needs in this area.

In this research we implement an event detection and extraction system in historical Arabic texts.

The linguistic complexity of the Arabic language and different structure of historical texts are the research challenges.

Our system uses rule based approach and extract information based on lexico-syntactic rules. System architecture has a modular structure and eventinformation is extracted independently, so extraction failure in one element does not effect on another one.

As future work, can creating rules automatically using rule induction methods and few primary rules. In this case system requirement to expert linguistic knowledge will be less.

.

References

Abuleil, S. (2007). *Using NLP Techniques For Tagging Events in Arabic Text*, Proceedings of the 19th IEEE International Conference on Tools with Artificial Intelligence, vol. 2, 440-443.

Ahn, D. (2006). *Stage of Event Extraction*, Proceedings of the Workshop on Annotating and Reasoning about Time and Events, 1–8.

Ananiadou, S.,Pyysalo, S., Tsujii, J. and B. Kell, D. (2010*). Event extraction for systems biology by text mining the literature*, Trends in Biotechnology, vol. 28, 381-390

Aone, C. and Ramos-Santacruz, M. (2000). *REES: A Large-Scale Relation and Event Extraction System*, Proceedings of the sixth conference on Applied natural language processing, pages 76-83.

Asgari Bidhendi, M & Minaei-Bidgoli, B & Jouzi, H. (2012). *Extracting person names from ancient Islamic Arabic texts*, LRE-Rel Workshop. 1-6.

Hammadi, O.I, (2012), *Grammatical Relation Extraction in Arabic Language*, Journal of Computer Science 8 (6): 891-898

Hogenboom, F., Frasincar, F., Kaymak, U. and de Jong, F. (2011). *An Overview of Event Extraction from Text*, Available from http://ceur-ws.org/Vol 779/derive2011_submission_1.pdf.

Lei, B., Sheng, B. (2012). *Methods of Customer Requirements Feature Extraction on Product Reviews*, Publisher Journal of Information & Computational Science, **Vol.** 9, pages 2429- 2439.

Naughton, M.,Stokes, N. and Carthy, J. (2008). *Investigating Statistical Techniques for Sentence-Level Event Classification*,Proceedings of the 22nd International Conference on Computational Linguistics, pages 617–624.

Piskorski, J., Tanev, H. and Wennerberg, P.O. (2007). *Extracting Violent Events From On Line News for Ontology Population*, Publisher Springer-Verlag Berlin Heidelberg, pages 287–300.

Sangeetha s, R.S. Takur and Michael Arock. (2010). *Event Detection using Lexical Chain*, Publisher Springer-Verlag Berlin Heidelberg, LNAI 6233, pages 314-316.

Sangeetha S, R.S. Takur and Michael Arock. (2010). *Domain Independent Event Extraction System Using Text Meaning Representation Adopted for Semantic Web*, Publisher International Journal of Computer Information Systems and Industrial Management Applications (IJCISI.M), ISSN: 2150-7988 Vol.2, pages 252-261.

Vargas-Vera, M. andCeljuska, D. (2004). *Event Recognition on News Stories and Semi-Automatic Population of an Ontology*, Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence, 2004, pages 615-618.

Xu, F., Uszkoreit, H. and Li, H. (2006). "*Automatic Event and Relation Detection with Seeds of Varying Complexity*", Proceedings of the AAAI 2006 Workshop Event Extraction and Synthesis.

http://projects.ldc.upenn.edu/ace/intro.html, August, 2012

http://www.csi.ucd.ie/staff/jcarthy/home/TDT.htmlAugust, 2012

AMIRA Tool:http://www.elda.org/medar-conference/pdf/56.pdf,August, 2012

KHOJA Tool. http://zeus.cs.pacificu.edu/shereen/research.htm,August, 2012