

Automatically Constructing Dictionaries for Extracting Meaningful Crime Information from Arabic Text

Meshrif Alruily, Aladdin Ayesh, Hussein Zedan¹

Abstract. Arabic is widely spoken language but very few text mining tools have been developed for Arabic language. This paper presents an algorithm for automatically building dictionaries. The target domain is crime profiling. The corpus is mined for crime type, location and nationality. This work is then validated through three experiments, the results of which show that the algorithm developed here is promising.

1 INTRODUCTION

Information extraction aims to extract specific, predefined entities from text. Feldman and Sanger [3] have stated that entities, such as peoples names, organizations names, locations, and crime type can all be extracted. According to Toral and Munoz [5], there are two types of evidence that help in identifying entities. Internal evidence can be deduced from the sentence that contains the entity by noticing the sequence of words. On the other hand, external evidence is gained from the context. This paper is an extension of our previous works in terms of endeavouring to find the best method for recognizing and extracting patterns of interest within the crime domain. Consequently, our goal is to develop a system that is able to recognize valuable information in a given document, such as types of crime, locations and nationalities. Therefore, the system here automatically generates three dictionaries, i.e. for the crime type, location and nationality, in domain of crime events. These automated dictionaries are then used to extract information from text instead of using manual dictionaries. As well-known the approaches that are used to identify named entities are hand-craft rules, machine learning approaches and hybrid. Alruily et al. [1] developed a software to extract crime information from texts based on a dictionary created manually. Riloff [4] developed a program called AutoSlog that automatically constructs a domain-specific dictionary for information extraction. Also, Chau et al. [2] used named entity extraction techniques to identify meaningful entities from police narrative reports. The rest of the paper is organized as follows: section 2 describes the domain analysis; section 3 presents Emergent and Adaptive Dictionaries (EAD) method; section 4 provides the results of the experiments and an evaluation of their performance; and finally, the conclusion is presented in section 5.

2 DOMAIN ANALYSIS

In this research, the crime domain is studied in the Arabic language. Most Arabic newspaper reports have the same structure, with respect to writing style, in the crime domain. Most journalists start with a

sentence containing the name of the police station that has investigated the crime, followed by a description of that crime. The crime description is about the type of crime committed and the type of criminal. Following these, details of any victims and other information are described. Fig. 1 shows an example from Alkhaleej newspaper. So the crime location can be deduced from this pattern "شرطة الشارقة"

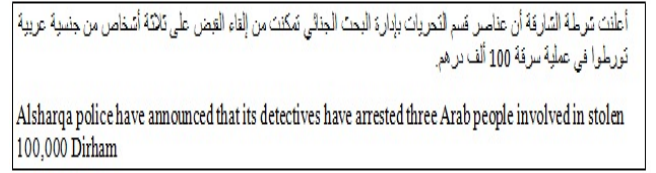


Figure 1. Crime Description

/ shurtat alsharqa". Furthermore, other related crime information can be extracted from the text, such as the nationalities of the people involved in the crime (e.g. "جنسية عربية" / jnsyt arabiat) type of crime can also be extracted from this pattern: "سرقه / tawaratwo fi sareekat". It can be noticed that the type of crime appeared in connection with prepositional phrases, as in the above example. The Arabic language has approximately fourteen prepositions. For example; the preposition "ب : bi" which has different meanings, so based on a whole sentence, its meaning can be inferred. In this research context, it means "because". Also, the preposition "في : fi" represents the preposition "in" in the English language. With regard to the preposition "ل : li", it has many meanings but in the current domain being studied it again means 'because'. Sometimes the preposition "إلى : ela" takes the place of preposition "ل : li". Additionally, they can work as a link between some verbs and nouns. Moreover, the meanings of sentences that contain some specific verbs cannot be identified without prepositional phrases. Consequently, it can be said there is a dependency relationship between some particular verbs "vrb" and prepositional phrases "preph". For example, the verb "تعرض" means "expose" in English, and always relates with the preposition "إلى : ela" or "ل : li". Also, the synonyms of the words "تورط" and "تخصّص" in English are "involved" and "specialized", respectively. Both of these two words come with preposition "في : fi". Moreover, the verb "قام" means "performed" and it correlates with the preposition "ب : bi". The noun after the preposition is considered to be the type of crime. With regard to crime location, it is usually attached

¹ Software Technology Research Laboratory, De Montfort University, email: meshrif, aayesh, hzedang@dmu.ac.uk

to these words: "منطقة / mantqt / region", "مدينة / madinat / town", "محافظة / mohafdat / Province", "شمال / shmal / north", "جنوب / jnwb / south", "شرق / sharq / east", "غرب / grb / west" and "وسط / wst / centre". By exploiting this relationship, a dictionary for location can be automatically constructed. Moreover, regarding the nationalities of the people involved in crime events. Nationality usually appears next to the word "الجنسية / aljnsyt".

3 EMERGENT AND ADAPTIVE DICTIONARY (EAD)

Fig. 2 shows our proposed algorithm for creating dictionaries based on linguistic analysis, which is presented in section 2. Three dictionary types are built. Crime type dictionary is built automatically by extracting crime types from a given text. The system here is focused on prepositional phrases. The advantage of this method is that there is no need for an annotated corpus. In other words, the system has no linguistic components, such as PoS taggers or shallow parsers. Instead, lists of keywords and their prepositions are provided to the system in order to extract the desired patterns. The aspect of most significance here is that each keyword and its accompanying preposition is followed by a noun, and it is this noun that is considered to be a type of crime. To sum up, in the processing phase, the system looks for words in a text that match the words in the trigger list. When a match occurs, the next step is to search for a preposition that always comes with the keyword. After that, the word that immediately follows the preposition is extracted and added to the dictionary. For creating location dictionary. There is strong correlation between the name of the city and the keywords that are stated in section 2. So in the processing phase, each word in a text is matched with each word in the keyword list. Once a match occurs, the word that follows the keyword is extracted and added to the dictionary. Finally, for constructing nationality dictionary the keyword "الجنسية / aljnsyt" is used. However, patterns for nationality sometimes appear after or before the keyword. To solve this problem, the article "ال : al" is exploited. So, if the word after the keyword has the prefix "al", then the pattern is extracted as a nationality pattern, otherwise the word before the keyword is the candidate for the nationality pattern.

4 EXPERIMENT AND EVALUATION

The system could automatically built crime type dictionary with more than 400 tokens, but after removing the affixes and performing frequency analysis, the number was reduced to 93 words. Moreover, a threshold was applied, i.e. words with a high frequency were chosen for the dictionary and the rest were discarded. As a result, the number of words that formed the crime type dictionary is 12 words. On the other hand, More than 1,000 tokens were initially extracted to form location dictionary. Also, a frequency analysis process and a threshold were also applied here. Consequently, our dictionary contains only 73 different location names. With respect to nationality dictionary, around 200 tokens were extracted. Because of use the previous three processes the number of nationalities in this dictionary is 32. In order to reveal the extent of the effectiveness, three experiments were performed. The effectiveness of the system has been evaluated using standard precision, recall and F-measure in each experiment. The first experiment is for evaluating the ability of the type of crime dictionary for extracting a type of crime from a given document. The system was able to extract 236. The number of entities that

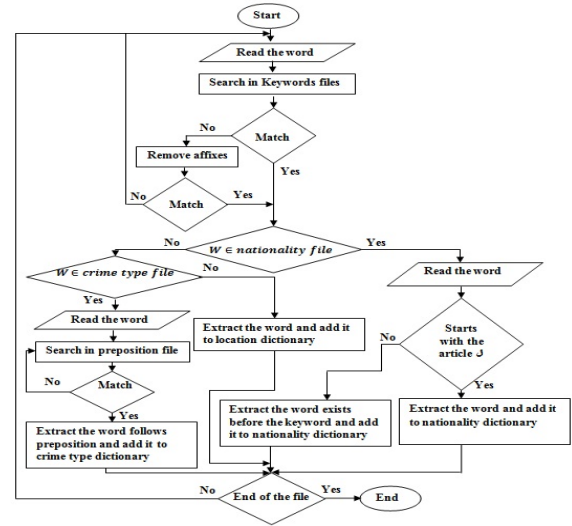


Figure 2. Dictionaries Builder (DB).

were correctly recognized is 213. Values of precision, recall and F-measure for this experiment are 90.25, 50.47 and 64.74, respectively. In the second experiment location dictionary is evaluated. 278 entities were correctly extracted out of 351. The averages of precision, recall and F-measure are 70.91, 79.20 and 74.83. Based on that the accuracy of this dictionary seems high, and therefore, it can be said that the performance of the location dictionary is good. Finally, in the third experiment the nationality dictionary was assessed; the system was able to correctly recognize 118 out of 123. The rates for precision, recall and F-measure are 81.19, 95.93, and 88.83 respectively, so very good results have been achieved.

5 CONCLUSION

This paper presents an algorithm for automated adaptive dictionary building to be used in crime text analytics within the domain of Arabic language. Three empirical tests have been performed to evaluate the performance of the algorithm applied into building three dictionary types. The outcomes of the experiments are positively encouraging. In conclusion, the approach shows efficiency and effectiveness in providing automated adaptive dictionary building algorithm.

REFERENCES

- [1] M. Alruily, A. Ayesh, and H. Zedan, 'Crime type document classification from arabic corpus', in *Second International Conference on Developments in eSystems Engineering*, pp. 153–159, Los Alamitos, CA, USA, (2009). IEEE Computer Society.
- [2] M. Chau, J. J. Xu, and H. Chen, 'Extracting meaningful entities from police narrative reports', in *dg.o '02: Proceedings of the 2002 annual national conference on Digital government research*, pp. 1–5. Digital Government Society of North America, (2002).
- [3] R. Feldman and J. Sanger, 'Information extraction', in *The Text Mining Handbook: Advanced Approches in Analyzing Unstructured Data: Cambridge university*, (2006).
- [4] E. Riloff, 'Automatic constructing a dictionary for information extraction tasks', in *The Eleventh National Conference on Artificial Intelligence*, pp. 811–816, (1993).
- [5] A. Toral and R. Munoz, 'A proposal to automatically build and maintain gazetteers for named entity recognition by using wikipedia', (2006).