UNDERSTANDING PEOPLE TITLE PROPERTIES TO IMPROVE INFORMATION EXTRACTION PROCESS

Saleem Abuleil and Khalid Alsamara

MMIS Department, Chicago State University, Chicago, USA

ABSTRACT

In this paper, we introduce a new approach to tackle the process of extracting information about people mentioned in the Arabic text. When a person name is mentioned in the Arabic text usually it is combined with a title, in this paper the focus is on the properties of those titles. We have identified six properties for each title with respect to gender, type, class, status, format, and entity existence. We have studied each property, identified all attributes and values that belong to each one of them and classified them accordingly. Sometimes person title is attached to an entity; we have also identified some properties for these entities and we show how they work in a harmony with person title properties. We use graphs for the implementation, nodes to represents person title, person name, entity and their properties, where edges are used to present inherited properties from parent nodes to child nodes.

KEYWORDS

People Titles, Title Properties, NLP, Arabic Language

1. INTRODUCTION

Information Extraction (IE), as defined in the Message Understanding Conferences, has been traditionally defined as the extraction of information from a text in the form of text strings and processed text strings that are placed into slots labeled to indicate the kind of information that can fill them. The problem of extracting information from a large document collection can be approached using many different algorithms. The three classic models used in information extraction, (under which all these algorithms can be loosely grouped), are called Rule-based, Pattern Learning, and Supervised Learning. Most of the Arabic Named Entity Recognition (NER) systems use keywords such as titles to tag proper name phrases in the text, once they tag proper name phrase they use either rule-based systems or statistical approach to tag proper names and extract information about them. Using titles to tag proper names in the Arabic text is an important technique that has been used widely, but titles have been used as keywords for the purpose of identifying proper name phrases and tag proper names without studying and exploring their properties. Our technique in this paper is to identify and use title properties and attributes to enhance the result of extracting information about people names in the Arabic text.

David C. Wyld et al. (Eds) : AIAPP, CRIS, CoSIT, SIGL, SEC - 2018
pp. 37–46, 2018. © CS & IT-CSCP 2018 DOI : 10.5121/csit.2018.80403

2. LECTURER

Al-Kouz [1] presented a framework designed for mining the explicit and implicit lexical semantic information impeded in the structure and the content of Aljazeera.net. Furthermore, it provides an efficient and structured access to the resulted semantic graph, the authors also claim in their paper that Aljazeera.net is professionally edited and has rich semantic structure and it establishes an asset, an impediment and a challenge for research in Arabic Natural Language Processing. Abdallah [2] proposed a simple method for integrating machine learning with rule-based systems and implement this proposal using the state-of-the-art rule-based system for NERA, the experimental evaluation shows that their integrated approach increases the F-measure by 8 to 14% when compared to the original (pure) rule-based system and the (pure) machine learning approach, and the improvement is statistically significant for different datasets, more importantly, their system outperforms the state-of-the-art machine-learning system in NERA over a benchmark dataset. Abdul Hamid [3] introduced simplified yet effective features that can robustly identify named entities in Arabic text without the need for morphological or syntactic analysis or gazetteers, a CRF sequence labeling model is trained on features that primarily use character n-gram of leading and trailing letters in words and word n-grams, the proposed features help overcome some of the morphological and orthographic complexities of Arabic.

Abuleil [4] presented a new technique to extract names from the text by building a database and graphs to represent the words that might form a name and the relationships between them. First, they mark the phrases that might include names, second they build graphs to represent the words in these phrases and the relationships between them, and third, they apply rules to find the names. Benajiba [6] investigated the impact of using different sets of features in three discriminative machine learning frameworks, namely, support vector machines, maximum entropy and conditional random fields for the task of named entity recognition ,they explore lexical, contextual and morphological features and nine data-sets of different genres and annotations; they measure the impact of the different features in isolation and incrementally combine them in order to evaluate the robustness to noise of each approach.

Chen [8] described their system for the CoNLL-2012 shared the task, which seeks to model co-reference in Onto Notes for English, Chinese, and Arabic; they adopt a hybrid approach to co-reference resolution, which combines the strengths of rule-based methods and learning-based methods, they official combined score over all three languages is 56.35. In particular, their score on the Chinese test set is the best among the participating teams. Habash [9] made an argument that is the many differences between Dialectal Arabic and Modern Standard Arabic (MSA) pose a challenge to the majority of Arabic natural language processing tools, which are designed for MSA, so in their paper retarget an existing state-of-the-art MSA morphological tagger to Egyptian Arabic (ARZ), their evaluation demonstrates that their ARZ morphology tagger outperforms its MSA variant on ARZ input in terms of accuracy in part-of-speech tagging, diacritization, lemmatization and tokenization; and in terms of utility for ARZ-to English statistical machine translation. Pasha [10] presented MADAMIRA, a system for morphological analysis and disambiguation of Arabic that combines some of the best aspects of two previously commonly used systems for Arabic processing, MADA (Habash and Rambow, 2005; Habash et al., 2009; Habash et al., 2013) and AMIRA (Diab et al., 2007). MADAMIRA improves upon the two systems with a more streamlined Java implementation that is more robust, portable, extensible and is faster than its ancestors by more than an order of magnitude.

3. PEOPLE TITLE PROPERTIES

When a person name is mentioned in the Arabic text usually it is attached to a title. We have studied these titles and identified different properties for each one of them; we also identified some attributes and values for each property. We have identified six properties for each title with respect to gender, type, class, status, format, and existence of an entity. In this section, we explain each one of them and we show some examples in table 1.

Gender: in the Arabiclanguage, there are two values for this property masculine and feminine. In Arabic language to form a feminine title from the masculine, you simply add "taa' marbuta" which looks like (ة, ـة) to the end of the title, for exampleوزيرWazer (he) Minister is a masculine and to form the feminine from it we add "taa' marbuta" " ة" to the end of the title وزيرةWazertn (she) Minister

Type: We have classified title into three Types:

- Occupational title that indicates a position or job of the person like Manager مدير Minister وزيرPresident رئيسand Consultant مستشار
- A social title like Mr. سيد, Ms. انسه,and Mrs. سيدة.
- Professional title that refers to a certain profession like engineerمهندس, physician طبيب, attorney محامي, and teacher استاذ

A person might hold two titles at the same time such as(Mr. and consultant), and(engineer and manager), etc. some titles could be used for two classes like الشيخ Sheikh, could be used for social or occupational.

Class: based on job field that people they hold we have classified titles into different classes: politics, religion, education, sport, media, industry, military, etc. some titles could be used for two types like الشيخ Sheikh, could be used for religion or politics, some titles like president and manager could be classified into different classes politics, education, sports, industry, etc. to identify the class for these cases we use the entities as we discuss later. Some titles do not have a class such as Mr. and Mrs.

Status: person title could be simple or compound, simple title has one word such as سيد Mr. and وزير Minister, compound title has two words like ولي العهد Crown Prince and الناطق الرسمي Official Speaker

Format: there are two formats of the title either defined or indefinite. Arabic word starts with ال (the) to define it. Al- (ال) is the definite article in the Arabic language. For example, the word وزير wazer "Minister" can be made definite by prefixing it with al-, resulting in الوزير al-wazer "the Minister". Consequently, al- is typically translated as The in English. A defined title that starts with الand the word followed is not a verb, adjective, nationality, or particle then the noun is most likely is a person name.

Entity Existence: Sometimes an entity comes between person title and person name, entity existence property has two values either Yes or No. Most likely if a person title is defined no entity to follow. More details about entities are discussed in the next section

4. ENTITIES AND THEIR PROPERTIES

Sometimes entity or nationality comes between person title and person name. We have studied hundreds of cases and based on our study we identified four properties for each one of them with respect to type, class, gender, and status. See table 2. In this paper we handle the following three most common scenarios:

- Person Title + **Entity (both title and name are mentioned)** + Person Name

  Example: رئيس**دولة تركيا**اردغان President of **Turkey State**Erdoğan

  In this example entity title (دولة State) and entity name (اردغانErdoğan) are mentioned

- Person Title + **Entity (title is omitted and name is mentioned**) + Person Name

  Example: مراسل **الجزيرة** إلياس كرام Report of **Aljazeera** Alysa Karram

  In this example entity tile (اخباريهقناة News Channel) is omitted and entity name (**الجزيرة Aljazeera)** is mentioned

- Person Title + **Entity (title is mentioned and name/nationality is omitted)** + Person Name

  Example: الناطق الرسمي باسم **الحكومه**الدكتور محمد المومني

  Official speaker of the **government** Dr. Muhammad Almumani

  In this example entity title (الحكوكة government) is mentioned and nationality (الاردنيه Jordanian) is omitted

In some cases it is difficult to identify the class property of person title like President رئيس and Managerمدير where each one of them can be classified into different classes such as politic, sport, industry, but when an entity is maintained, and by using the value of the class of that entity, it helps to identify the class of the person title, for example when a company name is attached to person title رئيس President, the title is classified as industry category and when university name is attached to the same person tile رئيس president, the title this time is classified as education category.

In this paper, we identify nationality as an adjective and we identify four properties for it with respect to gender, type, format and country. Gender could be either masculine or feminine, the type has the value nationality, and the country has the value which country the nationality belongs to. When nationality comes directly after theرئيس president, we add a class property with a value politics.

Table 1.Title Properties

Gender		Format	Type	Class	Entity Existence
Masculine	Feminine				
وزير Minister	وزيرة	Indefinite	Occupational	Politics	Yes
الوزير The Minister	الوزيرة	Defined	Social	Politics	No
لاعب Player	لاعبة	Indefinite	Occupational	Sport	Yes
الاعب The Player	الاعبة	Defined	Occupational	Sport	No
ولي العهد Crown Prince	وليه العهد	Indefinite	Occupational	Politics	No
الناطق الرسمي Official Speaker	الناطقة الرسمية	Defined	Occupational	Follows entityclass	Yes
مدير Manager	مديره	Indefinite	Occupationally	Follows entityclass	Yes
المدير The Manager	المديرة	Defined	Social	Follows entityclass	No
المحامي The Attorney	المحاميه	Defined	Professional	Law	No
الشيخ The Sheikh	الشيخه	Defined	Social Occupational	Politic or Religion	No
السيد Mr.	السيدة Mrs.	Defined	Social	N/A	No
المدرس The Teacher	المدرسة	Indefinite	Professional	Education	No
مراسل Reporter	مراسلة	Indefinite	Media	Media	Yes

Table 2.Entity Properties

Title	Gender	Class	Type
دولة state	Feminine	Politics	Location
حكومة	Feminine	Politics	Location
مملكة kingdom	Feminine	Politics	Location
وزارة ministry	Feminine	Politics	Organization
جمعية society	Feminine	Social	Organization
نادي club	Masculine	Social / Sport	Organization
جامعة university	Feminine	Education	Organization
كلية college	Feminine	Education	Organization
مدرسة school	Feminine	Education	Organization
ملعب stadium	Masculine	Sport	Location
قناه اخبارية News Channel	Feminine	Media	Organization
جامع mosque	Masculine	Religion	Organization
كنيسة church	Feminine	Religion	Organization

5. ANALYSIS

We use graphs to implement the concepts in this paper; we use nodes to represents person title, person name, nationality, entity and their properties, and we use edges to present inherited properties from parent node to child node, see fig 1. We tag person name phrases in the text, each phrase starts with a title and terminated with a person name, next we tag the elements in the tagged phrases with respect to titles, entities, and nationalities and then apply the concepts of this paper to identify the properties of each one of them. Properties are inherited from one node to another (parent to child) and once a child gets inherited property from the parent they also forward this information to next node in the graph. When reach the last node in the graph we process all information from all nodes and produce the results. Harmony of inherited information between nodes is also validated with respect to the properties of the titles, entities, and nationalities. There is should be a match between two titles belong to the same person, same gender and format values between adjective (nationality) and the element presided to it either a title or an entity

In figure 2 we illustrate different scenarios about the same person and we show how the person inherits all properties from all parent nodes, the person name is Albright, Albright is a female, she is a politician, and her occupation is the minister of the USA foreign ministry. Example 1 contains two titles and one adjective (nationality), the first title and the adjective both have the same values of format property and gender property, both titles mentioned for the same person and they have the same value of gender property. Example 3contains one entity which is الخارجيـــه foreign and one adjective الامريكيـــه American, since both of them have the same value of format property "defined", the title وزيـرة minister has a different format property value "indefinite", then the nationality refers to the organization and not to the person title, the entity title وزارة ministry is omitted, both titles and the nationality of this example belong to the same person they have the same value of gender property.

<table>
<tr><td colspan="4">Example 1</td></tr>
<tr><td>الوزيرة The Minister</td><td>Title</td><td rowspan="2">Defined</td><td rowspan="2">Feminine*</td></tr>
<tr><td>الامريكية American</td><td>Adjective</td></tr>
<tr><td colspan="4"></td></tr>
<tr><td>السيدة The Mrs.</td><td>Title</td><td colspan="2">Feminine*</td></tr>
</table>

<table>
<tr><td colspan="4">Example 3</td></tr>
<tr><td>وزيرة Minister</td><td>Title</td><td>Indefinite</td><td rowspan="3">Feminine*</td></tr>
<tr><td>الخارجية Foreign</td><td>Entity</td><td rowspan="2">Defined</td></tr>
<tr><td>الامريكية American</td><td>Adjective</td></tr>
<tr><td colspan="4"></td></tr>
<tr><td>السيدة The Mrs.</td><td>Title</td><td colspan="2">Feminine*</td></tr>
</table>

In figure 3 we show different scenarios where the value of the class property of person title is uncertain and we are going to use the value of the class property of the entity to identify it. In example 1, the value of the class property of the entity is education, and then we use it for the value of the class property of the title "رئيس". Example 2, the value of the class property of the entity is politics, and then the value of the class property of the title "رئيس" is also politics, the second title is Sheik "شيخ" and could have two possibilities either politics or religion but since the first title has the class property value is politics we select the value politics as well for the second

title. In example 3, the value of the class property of the entity is industry, and then the value of the class property of the title "رئيس" is also industry.

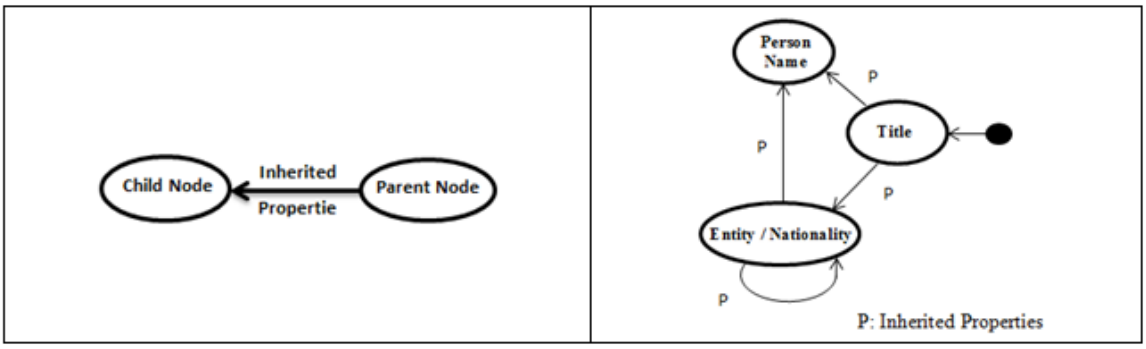
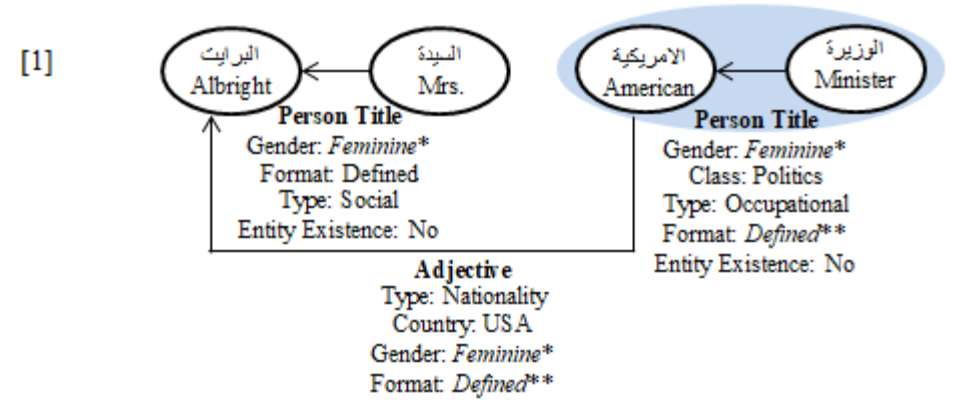
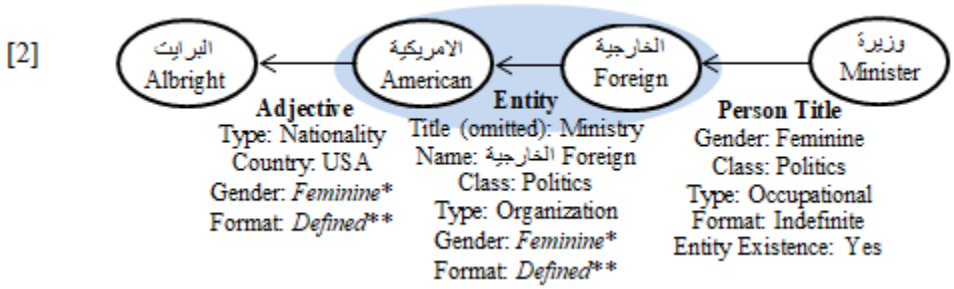


Fig 1. Graph

6. CONCLUSION

In this paper we have studied person titles, entities and nationalities attached to them, identify the properties for each one of them and used this information to extract information about people mentioned in the Arabic text we also validated the inherited property values between the nodes in the graph. Our source of data is Ajazeera.net, further analysis to be done in the future.





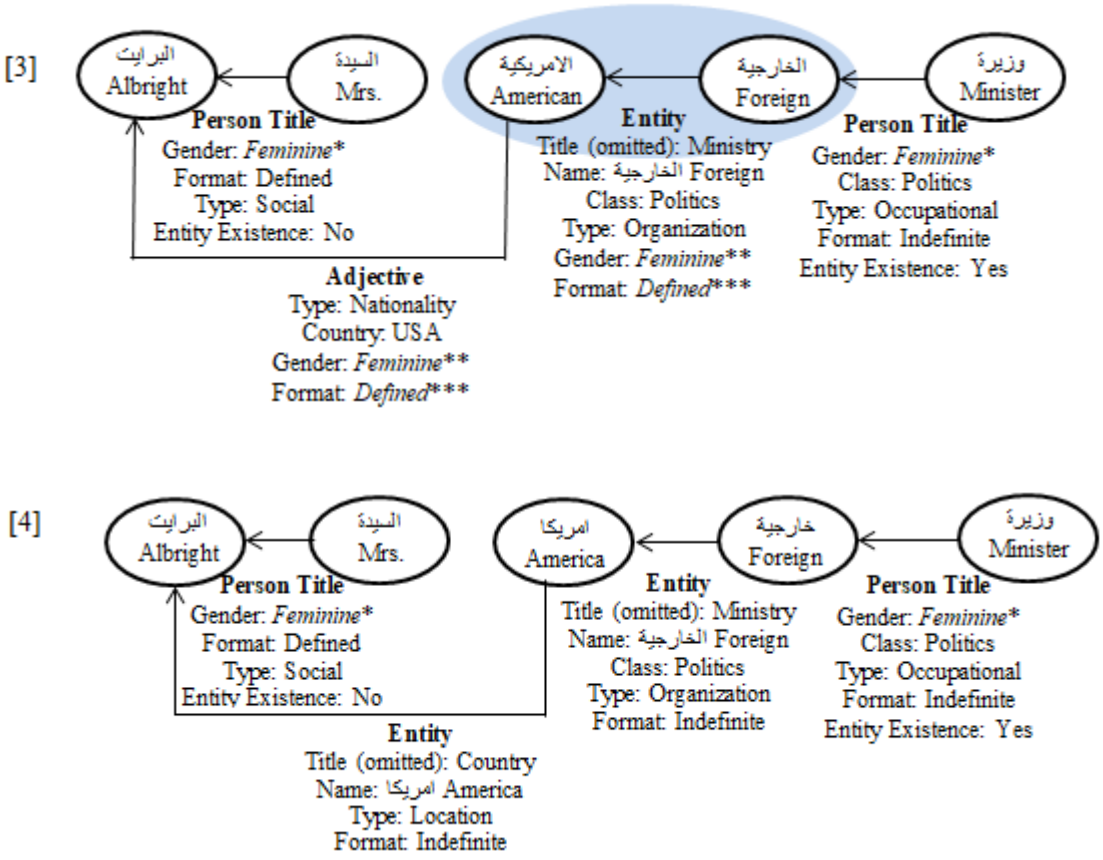
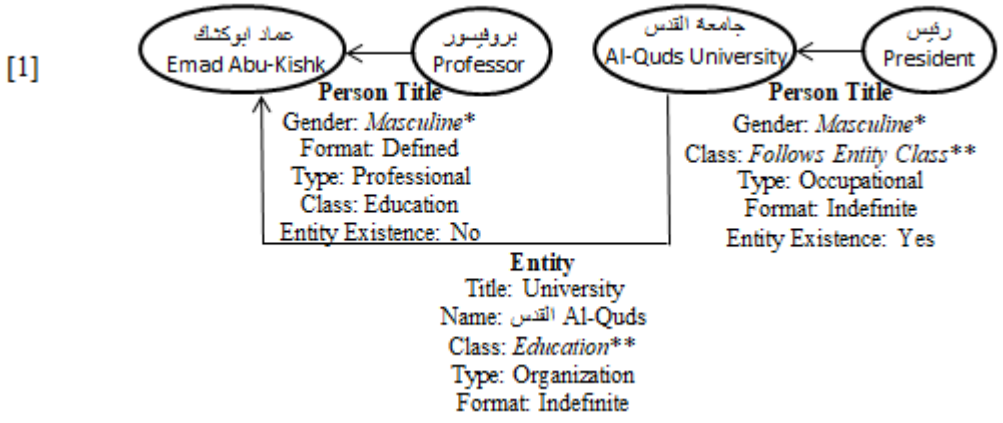


Fig 2.Illustration A



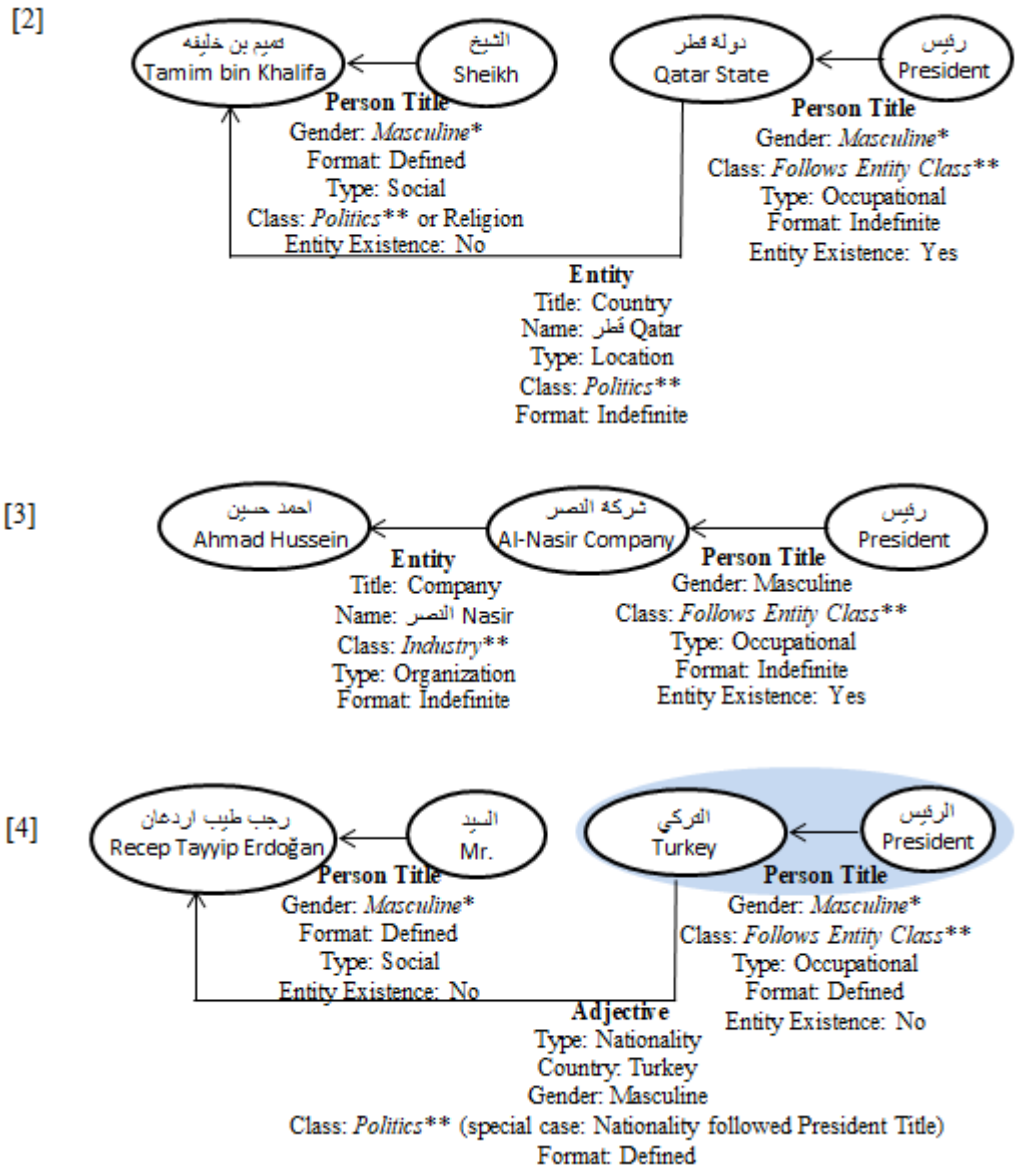


Fig 3. Illustration B

REFERENCES

[1] Al-Kouz, A., Awajan, A., Jeet, M., Al-Zaqqa, A.: Extracting Arabic semantic graph from Aljazeera. net. In: 2013 IEEE Jordan Conference on Applied Electrical Engineering and Computing Technologies (AEECT), (pp. 1–6). IEEE, Dec 2013

[2] Abdallah, S., Shaalan, K.,Shoaib, M.: Integrating rule-based system with classification for Arabic named entity recognition. In: Gelbukh, A. (ed.) CICLing 2012. LNCS, vol. 7181, pp. 311–322. Springer, Heidelberg (2012)

[3] Abdul Hamid, A., Darwish, K.: Simplified feature set for Arabic named entity recognition. In: Proceedings of the 2010 Named Entities Workshop, pp. 110–115. Association for Computational Linguistics, Uppsala (2010)

[4] Abuleil, S.: Extracting names from Arabic text for question-answering systems. In: Pro-ceedings of Coupling approaches, coupling media and coupling languages for information retrieval, RIAO 2004,pp. 638-647. Avignon(2004)

[5] Abuleil, S.: Hybrid system for extracting and classifying Arabic proper names. In: Pro-ceedings of the fifth WSEAS International Conference on Artificial Intelligence, Knowledge Engineering and Data Bases, AIKED 2006, pp. 205-210. Madrid (2006)

[6] Benajiba, Y., Diab, M., Rosso, P.: Arabic named entity recognition: A feature-driven study. IEEE Trans. Audio Speech Lang. Process. 17(5), 926–934 (2009)CrossRef Google Scholar

[7] Benajiba, Y., Diab, M., Rosso, P.: Arabic named entity recognition: An svm-based ap-proach. In: The International Arab Conference on Information Technology, ACIT 2008 (2008)

[8] Chen, C., Ng, V.: Combining the best of two worlds: a hybrid approach to multilingual coreference resolution. In: Joint Conference on EMNLP and CoNLL-Shared Task, pp. 56–63. Association for Computational Linguistics, July 2012Google Scholar

[9] Habash, N., Roth, R., Rambow, O., Eskander, R., Tomeh, N.: Morphological analysis and disambiguation for dialectal Arabic. In: HLT-NAACL, pp. 426–432 (2013)Google Scholar

[10] Pasha, A., Al-Badrashiny, M., Diab, M., El Kholy, A., Eskander, R., Habash, N., Roth, R.M.: (2014). Madamira: a fast, comprehensive tool for morphological analysis and disambiguation of Arabic. In: Proceedings of the Language Resources and Evaluation Conference (LREC). Reykjavik, Iceland Google Scholar