Named Entity Recognition of Persons' Names in Arabic Tweets

Omnia H. Zayed
Center of Informatics Science
Nile University
Giza, Egypt
omnia.zayed@gmail.com

Samhaa R. El-Beltagy
Center of Informatics Science
Nile University
Giza, Egypt
samhaa@computer.org

Abstract

The rise in Arabic usage within various social media platforms, and notably in Twitter, has led to a growing interest in building Arabic Natural Language Processing (NLP) applications capable of dealing with informal colloquial Arabic, as it is the most commonly used form of Arabic in social media. The unique characteristics of the Arabic language make the extraction of Arabic named entities a challenging task, to which, the nature of tweets adds new dimensions. The majority of previous research done on Arabic NER focused on extracting entities from the formal language, namely Modern Standard Arabic (MSA). However, the unstructured nature of the colloquial language used in tweets degrades the performance of NER systems developed to support formal MSA text. In this paper, we focus on the task of Arabic persons' names recognition. Specifically, we introduce an approach to extract Arabic persons' names from tweets without employing any morphological analysis or language-dependent features. The proposed approach adopts a rule-based model combined with a statistical one. This approach uses unsupervised learning of patterns and clustered dictionaries as constrains to identify a person's name and resolve its ambiguity. Our approach outperforms the best reported result in the literature on the same test set by an increase of 19.6% in the F-score.

1 Introduction

Named Entity Recognition (NER) is the task of identifying certain types of named expressions in unstructured text and classifying them into a predefined set of categories. These expressions can be personal and geographic named expressions, as well as temporal and numeric ones. NER is a crucial constituent of many Natural Language Processing (NLP) applications (Jurafsky and Martin, 2009). Examples of these applications include Machine Translation, Text Summarization, Opinion Mining, and Semantic Web Searching (Benajiba et al., 2009).

The advent of Twitter has offered people a significant new way of communication that enables them to share their ideas, thoughts, and real-time news, an example of which was the D.C. earthquake[1], which was reported on Twitter as it was unfolding. In addition, Twitter can be used by government services to reach large audiences in real time in order to send awareness messages to citizens. The sheer amount of regularly generated tweets and their ubiquitous nature are among the factors that have encouraged researchers in many fields to analyse such content automatically for event detection and opinion mining. The informal nature of messages exchanged within this platform poses new challenges for NLP applications, as their content tends to be short, noisy and to deviate from known grammatical rules (Zayed and El-Beltagy, 2015).

When it comes to automatic text analysis, the Arabic language is challenging not only due to its inflective nature but also due to its complex linguistic structure, its rich morphology, (Farghaly and Shaalan, 2009) as well as its inherent ambiguity. Ambiguity is in fact, one of the major challenges in detecting Arabic persons' names (Zayed and El-Beltagy, 2015).

Research in the area of Arabic NER is still in its early phases compared to that of English NER (Shaalan, 2014), with the focus of most of the research being done in this area being on MSA. The language being used on most social media platforms however is colloquial Arabic introducing a new set of complications with the multitude of dialects being employed. With the rapid increase in online social media usage by Arabic speakers, it is important to build Arabic NER

[1] http://socialmediasun.com/impact-of-social-media-on-society/

Proceedings of Recent Advances in Natural Language Processing, pages 731–738,
Hissar, Bulgaria, Sep 7–9 2015.

systems capable of dealing with both colloquial Arabic and MSA text.

The aim of this work is to extract Arabic persons' names, the most challenging Arabic named entity as discussed in Section 2, from tweets. Previous approaches that have tackled the problem of Arabic NER relied heavily on complex linguistic processing in terms of parsing and morphological analysis to solve the ambiguity problem. While these approaches are applicable to MSA text, they cannot handle colloquial Arabic with an acceptable precision. The unstructured nature of the colloquial language used in tweets degrades the performance of NER systems that are trained on the formal language style used in news contexts for example. This fact was proved by experimental results as presented in (Darwish, 2013) on Arabic tweets and in (Ritter et al., 2011) on English ones.

Our proposed approach adopts a rule-based model combined with a statistical model. The statistical model is based on association rules and is built by employing unsupervised learning of context patterns that indicate the presence of a person's name. This approach makes use of a limited set of dictionaries augmented with a name-clustering module, coupled with a set of rules to identify a person's name and resolve its ambiguity.

The main contributions of this work can be summarized as follows:

1. Introduces a "text style" independent approach to recognise persons' names that can be easily ported to other languages, text styles/genres and domains.
2. Overcomes the ambiguity problem of persons' names without using language-dependent resources such as parsers, taggers and/or morphological analysers. The only resource required by the system is a list of persons' names which can be easily obtained from publicly available resources such as Wikipedia[2].

The rest of the paper is organised as follows: Section 2 highlights some of the unique characteristics of the Arabic language with respect to the task of persons' names extraction. Section 3 reviews previous work done on Arabic NER with focus on Arabic NER from social media contexts. The proposed approach is discussed in Section 4. In Section 5, the conducted experiments to evaluate the system's performance are described. Finally, the conclusion and future work are presented in Section 6.

2 The Effect of Arabic Specific Challenges on NER

Arabic is a widely used language spoken by over 300 million people, and one of the official languages used at the United Nations (UN). The very special characteristics of the Arabic language step up the challenges faced by researchers when developing an Arabic NLP application (Farghaly and Shaalan, 2009). Among the challenging characteristics are a rich morphology, complex orthography, and the different levels of ambiguity. Additionally, tweets are usually written in colloquial Arabic, with dialects from all over the Arab World being represented, which complicates the problem of Arabic NER (Zayed and El-Beltagy, 2015). This problem is more formally referred to as diglossia. Many researchers, (Shaalan, 2014; Farghaly and Shaalan, 2009; Zayed et al., 2013; Zayed and El-Beltagy, 2015), examined the unique characteristics of Arabic extensively. The coming sub-sections will highlight these characteristics briefly and explain their effects on the extraction of persons' names, which is the focus of this paper.

2.1 Diglossia

One of the major linguistic features that characterise the Arabic language is its diglossia, which refers to the existence of two forms of the language: formal and informal. The formal language, namely Modern Standard Arabic (MSA), is used for most written and formal spoken purposes but is not used for daily communication, whereas the informal language, namely colloquial Arabic, is used for daily communication and may differ geographically (Farghaly and Shaalan, 2009). Colloquial Arabic is comprised of multiple spoken Arabic dialects used for daily communication in different Arab countries. It varies regionally from one Arabic speaking country to another. Colloquial Arabic is very commonly used within all social media platforms. There are significant differences between Arabic dialects regarding various linguistic features. These differences also exist between these dialects and MSA (Habash, 2010). As mentioned earlier, colloquial Arabic adds challenges to NER due to its unstructured and informal nature.

The usage of colloquial Arabic as a written language on social media platforms adds extra

[2] https://www.wikipedia.org/

complexity to an already difficult problem, as discussed in sub-section 2.4.

2.2 Complex Orthography

Arabic has no capital letters which is a distinctive feature when it comes to NER. Besides, it has no letters dedicated for short vowels. Special marks placed above or below the letters, namely diacritics, are used to compensate for the absence of short vowels. However, these diacritics are rarely used in contemporary writings; yet, it is possible for a native speaker to infer the missing diacritics (Farghaly and Shaalan, 2009).

The absence of diacritics causes structural and lexical ambiguity in which a word may belong to more than one part of speech with different meanings. For example, the word “يحيي” without diacritics can imply the male name “Yahya”, or the verb (greets) or the verb (gives life back) (Farghaly and Shaalan, 2009; Zayed and El-Beltagy, 2015; Zayed et al., 2013).

2.3 Rich Morphology

The Arabic Language has an agglutinative and inflective nature in which suffixes, infixes, and prefixes can be attached to the root of a word.

This aspect creates semantic ambiguity in which one word could imply different meanings. A lot of examples can be found frequently in tweets such as, the word “مني” which may imply the colloquial phrase (from me), or the female name “Mona”. This problem will be complicated by adding a conjunction such as (and) at the beginning of the word to have a new word “ومني” which may imply (and from me) or (and Mona). The attachment of clitics such as conjunctions, particles and invocation letters to any given word only serves to complicate the task of extracting Arabic persons’ names. This problem is not confined to the example above, but extends to cases where invocation particles attach directly (without a white space separation) to Arabic named entities due to the limited number of characters allowed in twitter messages. For example, the invocation particle "يا" (O) can be found frequently in tweets attached directly to a name such as in “يامني” (O Mona) (Zayed et al., 2013; Zayed and El-Beltagy, 2015).

2.4 Ambiguity

The different levels of ambiguity in Arabic text is among the major challenges in detecting Arabic persons’ names (Zayed et al., 2013; Shaalan, 2014; Farghaly and Shaalan, 2009; Zayed and El-Beltagy, 2015). Many persons’ names are either derived from adjectives or can be confused with other nouns sharing the same surface form. Moreover, some Arabic persons’ names match with verbs or prepositions. In addition, some foreign persons’ names transliterated to Arabic may be confused with prepositions or pronouns. Examples of some ambiguous names are [Ahlam, Al-Asad, Tourk, Ann, Lee] which may confused with [dreams, the lion, he left, that, me/mine]. Some colloquial words may match with foreign persons’ names such as [Wayen/Wein, Mo, and Abby] which are polysemies of [Where, Not, and I want] in the Algerian/Tunisian, Saudi and Kuwaiti dialects, respectively. A variety of other examples can be found in (Zayed et al., 2013; Zayed and El-Beltagy, 2015).

Because of these factors, Arabic persons’ names are the most challenging Arabic named entities to be extracted without any morphological processing. Ignoring name ambiguity and employing a rule-based system that depends on straightforward matching using dictionaries, will result in an NER system that performs poorly (Shihadeh and Neumann, 2012; Darwish, 2013). On the other hand, the nature of colloquial Arabic will not allow the application of parsers and morphological analysers (the traditional solution for these challenges), as these tools have yet to perform at an acceptable level of accuracy on colloquial text (Zayed and El-Beltagy, 2015; Zayed et al., 2013). In this paper, the ambiguity of Arabic persons’ names is resolved by using scored patterns, learned in an unsupervised manner, and clustered dictionaries, as will be explained in detail in section 4.

3 Related Work

Shaalan (2014) surveyed the work done on Arabic NER. The majority of the previous work pertained to the formal MSA language style used in the news domain. A list of numerous works are reviewed extensively in this survey.

In this section, we will focus on the work done to extract Arabic named entities from social media contexts.

An attempt to extract Arabic named entities from tweets is introduced in (Darwish and Gao, 2014). In this work, a Conditional Random Field (CRF) classifier was utilized to extract persons’, locations’, and organizations’ names depending on “language-independent” features. The authors used a set of tweets that was collected and annotated in previous work by the authors, (Darwish, 2013), as a test set. The overall system achieved an F-score, on this test set, of 65.2%.

Prior to this work, Darwish (2013) applied a system which was trained on news to extract named entities from Arabic tweets. The system utilized cross-lingual features and knowledge bases (KBs) from English using cross-lingual links to train a CRF classifier. The system obtained an overall F-score of 39.9% on the tweets set used to test it. As mentioned previously, this same test set was used for the evaluation of the system presented in (Darwish and Gao, 2014).

A recent attempt to extract Arabic persons' names from tweets is presented in (Zayed and El-Beltagy, 2015). In this work, the authors present a hybrid approach that exploits context bigrams to train a Naïve Bayes classifier, which in turn, is plugged into a rule based model. The system performance was tested on a set of tweets used in Darwish (2013) and (Darwish and Gao, 2014). The F-score of this system on this set was: 59.59%. This same set of tweets was used to evaluate the proposed approach and the result is presented in Section 5.

A system introduced in (Zirikly and Diab, 2014) utilized morphological analysis and gazetteers among other lexical and contextual features to train a CRF classifier in order to extract persons' and locations' names from micro-blogs. The system was tested on a manually annotated portion of an Egyptian dialect corpus collected and provided by the LDC[3] from web blogs. The system obtained an F-score of 49.18% for the task of persons' names recognition. A performance comparison between our system and this system is not possible as the dataset used for evaluation the former, is not publically available.

In (Zayed et al., 2013), a similar system to the one discussed in this paper is presented. However, the presented system was applied to formal MSA text. In this paper, we extend the work carried out in (Zayed et al., 2013) to extract persons' names from Arabic tweets.

4 Overview of the Proposed Approach

In this work, we introduce a novel approach to extract persons' names and resolve their ambiguity from Arabic tweets. In this approach, a rule-based model combined with a statistical model, is adopted. The approach is suitable for both MSA, as proved previously in (Zayed et al., 2013), and colloquial Arabic as illustrated in this paper. Our approach tries to overcome two of the major shortcomings of using rule-based techniques which are the difficulty of modifying a rule-based approach for new domains and the necessity of using huge sets of gazetteers. The approach depends on unsupervised learning of patterns and clustered dictionaries as constrains to identify a person's name and resolve its ambiguity. Moreover, the approach does not require complex linguistic pre-processing or language-dependent features.

4.1 General Architecture

The presented approach makes use of unsupervised learning of patterns and clustered dictionaries as combinatory constraints plugged into a rule-based model to extract persons' names and resolve their ambiguity.

This idea was initially introduced by Zayed et al. (2013). The authors' experiments, in the context of formal MSA used in news articles, proved that this approach can be used to overcome the ambiguity problem of Arabic persons' names without using morphological analysis. In this paper, we apply the same methodology to extract persons' names and resolve their ambiguity from Arabic tweets.

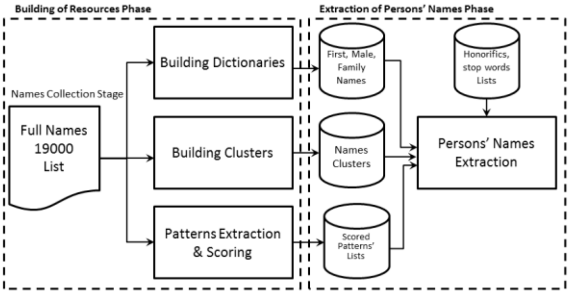
Building of Resources Phase
Extraction of Persons' Names Phase
Names Collection Stage
Full Names 19000 List
Building Dictionaries
Building Clusters
Patterns Extraction & Scoring
First, Male, Family Names
Names Clusters
Scored Patterns' Lists
Honorifics, stop words Lists
Persons' Names Extraction

Figure 1: System's General Architecture

The approach is composed of two phases, as shown in Figure 1. In the first phase, "The building of resources phase", persons' names are clustered, in addition, 'name' indicating patterns are extracted. In the second phase, "Extraction of persons' names phase", name patterns and clusters are used to extract persons' names from input text. Both of these phases are described in depth, in the following sub-sections.

4.2 The Building of Resources Phase

This phase utilizes a list of persons' full names gathered from publicly available resources. This collected list is processed to build dictionaries of first, middle and family persons' names as well as to create clusters of persons' names. Middle names are further used to build a list of male

[3] Linguistic Data Consortium (LDC2012T09: GALE Arabic-Dialect/English Parallel Text)

names. Finally, the list is used to build a statistical model of name indicating patterns. This process was previously introduced in (Zayed et al., 2013). We revisit this process in brief in the following sub-sections.

4.2.1 Names Gathering and Building of Dictionaries

The system depends on persons' names dictionaries that were collected by Zayed et al. (2013) and which are available online[4]. The authors employed both Wikipedia's people category[5] and Kooora[6] Arabic sports website to collect a list of nearly 19K full persons' names ("full_names_19000_list"). This list, was then processed and refined to build lists of first, male/middle and family persons' names automatically. These lists were necessary, since the aim of this work is not just to recognize names of famous people, but instead to identify the name of any person even if it does not appear in the collected lists. The technique followed in gathering the names and building these lists is described in (Zayed et al., 2013).

4.2.2 Building of Names' Clusters

The inherent ambiguity of the Arabic language degrades the performance of a system based on straightforward matching using dictionaries to extract previously unseen person's name as shown by experimentation on news articles (Shihadeh and Neumann, 2012) and on tweets (Darwish, 2013).

One of the common problems when extracting names, is the possibility of incorrectly extracting a name that is a combination of an Arabic name and a foreign name. For example, given the tweet "عوده تشافي وراكتيش في الوسط..." (the return of Xavi and Rakitić in the middle…), using a simple matching approach would result in the extraction of the full name "عوده تشافي" "Ouda Xavi", which is wrong. The problem could be encountered in various contexts, which all have a common factor: one part of the name is Arabic and the other part of the name matches with a transliteration of a foreign name. To overcome the incorrect extraction of entities like this, the observation that it is highly unlikely that an Arabic person's name will appear beside a foreign person's name can be utilized. However, name lists do not contain information regarding the origin of a name.

A workaround this lack of information was presented in (Zayed et al., 2013) in the form of name clusters. In this solution, name clusters were constructed by considering each single name a node. Since a full name, is made up of multiple names, names in a full name, are connected via links. Names are then clustered into communities using a graph clustering criterion (Blondel et al., 2008). As illustrated in Figure 2, culturally similar names are grouped together.

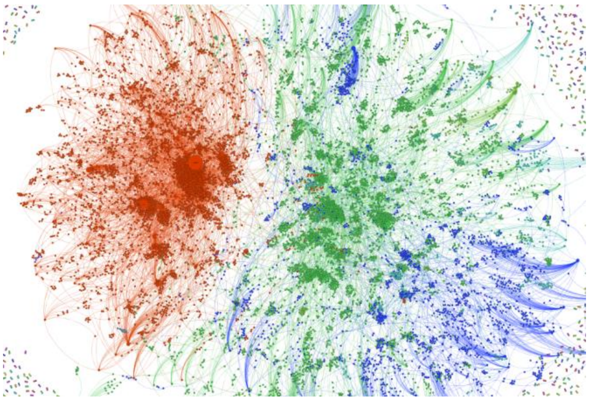

Figure 2: visualization of graph clustering of 19K persons' names

To overcome the above mentioned problem, only names in the same cluster can be combined together to form a name.

4.2.3 Extracting Scored Patterns

The goal of this phase is to build lists of patterns indicating the occurrence of a person's name in an unsupervised way. These patterns are scored using the support score to build a statistical model. After that, the statistical model is integrated with a set of rules, dictionaries and clusters to extract Arabic persons' names and resolve their ambiguity. This procedure is divided into 4 steps, as shown in Figure 3.

The initial two steps are carried out to create and pre-process the dataset used for learning the scored patterns. Since our target is to identify persons' names from Arabic tweets, we had to create our own dataset of tweets. To our knowledge, no similar dataset is currently available for NER research.

[4] http://bit.ly/NileAPgazet

[5] http://ar.wikipedia.org/wiki/تصنيف:تراجم

[6] http://www.kooora.com/default.aspx?showplayers=true

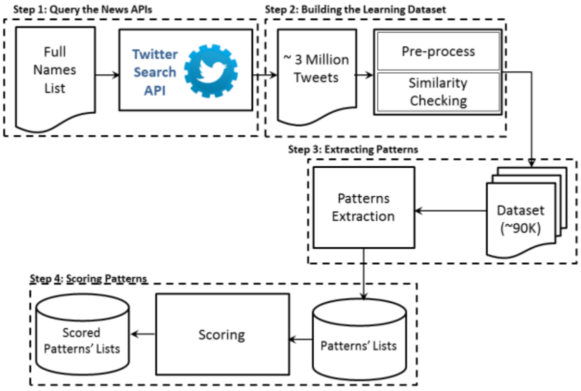

Figure 3: Building lists of patterns with score form Twitter context

The Twitter Search API was utilized to download Arabic tweets by using a random set of name selected from the aforementioned list of persons' names, as query terms. Since we are interested in getting tweets written in Egyptian Colloquial the queries were restricted to using the geo-code parameter "*30.0500, 31.2333, 500km*". This geo-code specifies the location of the retrieved tweets to be Cairo with a radius of 500km. Using this geo-code allows us to get the majority of tweets from Egypt and a small amount from Saudi Arabia, Jordan and Palestine. The language parameter was also set to Arabic ("lang: ar").

After tweets retrieval, using the Twitter Search API as mention earlier, normalization and preprocessing steps are carried out to omit unwanted features such as diacritics, hyperlinks and English words. It was also necessary to eliminate redundant tweets, due to re-tweets. To carry out this step, a similarity check was performed by employing the cosine similarity technique (Singhal, 2001) with a threshold value of 0.72. The final dataset consisted of around 100 thousand unique tweets.

Following these steps, unigram patterns around each name are extracted to form three lists of patterns. A list to keep unigram patterns before a name, and another one to keep unigram patterns after a name. Finally, a complete pattern list is created to keep set of complete patterns around the name. An example of a tweet that appears in the learning dataset is " لما استاذ ابراهيم عبد المجيد بيعملي فيفوريت" (when Mr. Ibrahim Abd El-Meguid is tagging me). The unigram patterns around the person's name "ابراهيم عبد المجيد" "Ibrahim Abd El-Meguid" are extracted as follows: the word "لما" (when) is added in the "before" list, the word "بيعملي" (is tagging) is added in the "after" list, and finally the set <when><name><is tagging> is added in the "complete pattern" list.

The final step is to score these patterns according to their significance in indicating the occurrence of a person's name. Therefore, each pattern in the three lists is scored using association rules support measure (Agrawal et al., 1993). Support is calculated as the ratio of the count of a pattern followed by a name over the total count of all patterns followed by a name. The three newly created lists of scored patterns are saved descendingly according to the value of the score.

4.3 Extraction of Persons' Names Phase

In this phase, the name extractor is created and used. The name extractor is composed of the scored patterns which are combined with rules to extract persons' names from tweets and resolve their ambiguity. Clustered dictionaries are used within the rules to ensure that all candidate portions of a name fall in the same cluster. Thus, the aforementioned problems of straight forward matching of names using dictionaries are avoided.

The baseline rule assumes that any full name consists of a first name followed by one or more male names followed by zero or one family name. Unigram patterns, honorifics, punctuations and titles that appear before and after a person's name are used to detect the name boundaries.

Examples of one of the employed rules include:

وفاه الطفل ماجد مدحت **برصاص...**
The death of the child Maged Medhat, who was shot by…

The use of scored patterns is crucial to avoid straight forward matching mistakes such as the extraction of "يمن سعيد" (Youmn Saied) which means here (Happy Yemen) in the phrase below.

باذن الله ترجع يمن سعيد
God willing, happy Yemen will return

Additionally, the use of clusters as a combinatory constraint eliminates false positives such as the extraction of "بشر بان" (Beshr Ban) which means here (bode that), as the Arabic name (Beshr) which means here (bode) and the transliterated foreign name (Ban) which means here (that) are not in the same cluster.

او بشر بان المسلمين...
Or bode that Muslims…

5 System Evaluation

System		**Precision**	**Recall**	**F-score**
Cross-Lingual Resources approach trained on news presented in (Darwish, 2013)		40.5%	39.2%	39.8%
Supervised ML approach presented in (Darwish and Gao, 2014)		67.1%	47.8%	55.8%
E1: (mistakes + English Entities)	Hybrid (Zayed and El-Beltagy, 2015)	67.20%	53.53%	59.59%
	Our Proposed Approach	***81.93%***	***56.32%***	***66.75%***
E2: (**no** mistakes + English Entities)	Hybrid (Zayed and El-Beltagy, 2015)	71.24%	57.24%	63.47%
	Our Proposed Approach	**85.36%**	**59.31%**	**69.99%**
E3: (mistakes + **no** English Entities)	Hybrid (Zayed and El-Beltagy, 2015)	66.49%	58.74%	62.38%
	Our Proposed Approach	**81.99%**	**61.54%**	**70.31%**
E4: (**no** mistakes + **no** English Entities)	Hybrid (Zayed and El-Beltagy, 2015)	69.92%	64.15%	66.91%
	Our Proposed Approach	**85.40%**	**65.17%**	**73.92%**

Table 1: Evaluation results of our approach in comparison to other systems' for illustration

5.1 Evaluation setups

Evaluating the performance of the proposed approach was done using CoNLL's standard evaluation script[7]. CoNLL's evaluation methods are aggressive methods, which means that no partial credit will be assigned for a partially extracted named entity (Shaalan, 2014). The results are given in terms of the standard measures for NER evaluation (De Sitter et al., 2004) which are precision, recall and F-score for each NER class; in our case, there is only a single class, which is "persons' names".

Evaluation was conducted on a test dataset of 1,423 tweets with nearly 26k tokens, used by the authors of (Darwish and Gao, 2014; Darwish, 2013). Arabic and English named entities are both tagged in this test set. This test set is referred to here as Darwish's test set. Details on Darwish's test set are provided in (Darwish and Gao, 2014). Statistical analysis of the test set can be found in (Zayed and El-Beltagy, 2015). It is worth noting that this dataset contains tweets written in Egyptian, Levantine, and Gulf Arabic dialects.

5.2 Experiments

The purpose of this evaluation was to determine the ability of the proposed approach to deal with colloquial Arabic text used on Twitter.

Similar to (Zayed and El-Beltagy, 2015), we carried out four different experiments to test the performance of our system. The first experiment was done using the dataset without any preprocessing or modification. The next experiment was done after fixing some annotation mistakes discovered in the dataset. Two final experiments were conducted to test the effect of removing English entities as a part of our pre-processing steps with and without the correction of the annotation mistakes. Since our system does not address the extraction of English entities, it is not entirely fair to include those when evaluating it.

The results obtained by our system are presented in Table 1. The table also compares the result of our proposed approach to the most recent hybrid approach proposed in (Zayed and El-Beltagy, 2015), in addition to the results obtained from the supervised Machine Learning (ML) systems presented in (Darwish and Gao, 2014; Darwish, 2013) which are used to extract named entities from tweets. We are not sure if (Darwish and Gao, 2014; Darwish, 2013) followed the same aggressive evaluation methodology as we and (Zayed and El-Beltagy, 2015) did.

It can be seen from the results that even without addressing annotation mistakes or the removal of English entities the presented approach achieves an increase of 12.01% in F-score over the one presented in (Zayed and El-Beltagy, 2015), and an increase of 19.6% over the work of (Darwish and Gao, 2014). Moreover, the F-score of our approach shows an increase of 67.7% over the one presented in (Darwish, 2013). Fixing the annotation mistakes improved the results by around 4.85%. Excluding the English entities improved the recall by 5.89%.

6 Conclusion and Future Work

This paper presented an approach for extracting Arabic persons' names and resolving their ambiguity. Our main intention while developing this approach is to attempt to resolve the inherent

[7] http://www.cnts.ua.ac.be/conll2000/chunking/conlleval.txt

ambiguity of Arabic persons' names without using "language-dependent" resources or depending on extensive lexical resources. The main goal is to be able to port the system to other domains, languages and text genres. This approach integrated name dictionaries and name clusters with a statistical model for extracting context unigram patterns in an unsupervised way, which are used to indicate the occurrence of persons' names. The main idea of this approach is to learn combinatory constraints via clustering of names and scored patterns. The approach exploited a list of full names, gathered from publicly available resources. Evaluation of the presented approach shows that it outperforms all recent attempts to extract Arabic named entities from tweets.

For the future, we plan to extend this approach to extract other named entities such as locations and organizations.

References

Rakesh Agrawal, Tomasz Imieliński, and Arun Swami. 1993. Mining association rules between sets of items in large databases. In Proceedings of the 1993 ACM SIGMOD international conference on Management of data, SIGMOD 1993, number May, pages 207–216, New York, USA. ACM Press.

Yassine Benajiba, Mona Diab, and Paolo Rosso. 2009. Arabic Named Entity Recognition: A Feature-Driven Study. IEEE Transactions on Audio, Speech, and Language Processing, 17(5):926–934, July.

Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. 2008. Fast unfolding of communities in large networks. Journal of Statistical Mechanics: Theory and Experiment, 2008(10):P10008, October.

Kareem Darwish. 2013. Named Entity Recognition using Cross-lingual Resources: Arabic as an Example. In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, volume 1, pages 1558–1567, Sofia, Bulgaria. Association for Computational Linguistics.

Kareem Darwish and Wei Gao. 2014. Simple Effective Microblog Named Entity Recognition: Arabic as an Example. In Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2014), pages 2513–2517, Reykjavik, Iceland, May. European Language Resources Association (ELRA).

A De Sitter, T Calders, and Walter Daelemans. 2004. A formal framework for evaluation of information extraction. Technical report, Antwerp.

A Farghaly and K Shaalan. 2009. Arabic natural language processing: Challenges and solutions. ACM Transactions on Asian Language Information Processing (TALIP), 8(4):1–22.

Nizar Y. Habash. 2010. Introduction to Arabic Natural Language Processing. Mogran & Claypool.

Daniel Jurafsky and James H. Martin. 2009. Information Extraction. In Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition., chapter 22, pages 725–743. Prentice Hall, 2nd edition.

Alan Ritter, Sam Clark, and Oren Etzioni. 2011. Named Entity Recognition in Tweets : An Experimental Study. In Conference on Empirical Methods in Natural Language Processing, pages 1524–1534.

Khaled Shaalan. 2014. A survey of arabic named entity recognition and classification. Computational Linguistics, 40(2):469–510, June.

Carolin Shihadeh and Günter Neumann. 2012. ARNE: A tool for namend entity recognition from Arabic text. In Fourth Workshop on Computational Approaches to Arabic Script-based Languages (CAASL4), San Diego, CA, USA.

Amit Singhal. 2001. Modern Information Retrieval: A Brief Overview. Bulletin of the IEEE Computer Society Technical Committee on Data Engineering, 24(4):35–43.

Omnia Zayed, Samhaa El-Beltagy, and Osama Haggag. 2013. An approach for extracting and disambiguating arabic persons' names using clustered dictionaries and scored patterns. In Elisabeth Métais, Farid Meziane, Mohamad Saraee, Vijayan Sugumaran, and Sunil Vadera, editors, NLDB 2013, LNCS, volume 7934, pages 201–212. Springer Berlin Heidelberg.

Omnia H. Zayed and Samhaa R. El-Beltagy. 2015. A Hybrid Approach for Extracting Arabic Persons' Names and Resolving their Ambiguity from Twitter. In 20th International Conference on Application of Natural Language to Information Systems (NLDB 2015), Passau, Germany, June. Springer.

Ayah Zirikly and Mona Diab. 2014. Named Entity Recognition System for Dialectal Arabic. In Proceedings of the EMNLP 2014 Workshop on Arabic Natural Language Processing (ANLP), pages 78–86, Doha, Qatar, October. Association for Computational Linguistics.