Automatically Developing a Fine-grained Arabic Named Entity Corpus and Gazetteer by utilizing Wikipedia

Fahd Alotaibi
School of Computer Science
University of Birmingham, UK
fsa081@cs.bham.ac.uk

Mark Lee
School of Computer Science
University of Birmingham, UK
m.g.lee@cs.bham.ac.uk

Abstract

This paper presents a methodology to exploit the potential of Arabic Wikipedia to assist in the automatic development of a large Fine-grained Named Entity (NE) corpus and gazetteer. The corner stone of this approach is efficient classification of Wikipedia articles to target NE classes. The resources developed were thoroughly evaluated to ensure reliability and a high quality. Results show the developed gazetteer boosts the performance of the NE classifier on a news-wire domain by at least 2 points F-measure. Moreover, by combining a learning NE classifier with the developed corpus the score achieved is a high F-measure of 85.18%. The developed resources overcome the limitations of traditional Arabic NE tasks by more fine-grained analysis and providing a beneficial route for further studies.

1 Introduction

Previous efforts that have been made to develop an Arabic NER either focused on traditional NE classes (Benajiba et al., 2010) or sought to expand only one class at a time (Shaalan and Raza, 2007). Applications such as Question Answering (QA) receive more benefits when a fine-grained NER is developed. This is true when we consider that, the majority of factoid questions are about named entities (Noguera et al., 2005). Having a finer NER, results in the possibility of extracting more semantic knowledge from the context. For example, if we consider the following sentence:

(شركة والت ديزني هي أكبر شركات وسائل الإعلام والترفيه في العالم /šrkħ wAlt dyzny hy Okbr šrkAt wsAŷl AlIʕlAm wAltrfyh fy AlʕAlm/ 'Walt Disney is the largest media company in the entertainment world')[1]

We would have more semantic information if we could tag (والت ديزني /wAlt dyzny/ 'Walt Disney') as [ORG-ENTERTAINMENT] rather than just [ORG]. This deeper semantics is very helpful when answering factoid question like "What is the largest entertainment company?"

Supervised machine learning technologies have been successfully adopted for several natural language tasks, including NER. These technologies require a reasonable portion of data to be accessible in the training phase, containing a number of positive and negative examples to learn from and to circumvent the problem of data sparseness. Traditional methods for compiling such data involve recruiting individuals to annotate a certain corpus manually. This is tedious work, as well as costly and time consuming. Moreover, manually annotating a large portion of a relatively open domain corpus beyond a news-wire and across various genres is not easy for an individual to achieve.

Therefore, when developing a reasonable fine-grained NE corpus two questions should be answered. First, *what proper fine-grained semantic classes should be established?* Second, *how to develop a reasonable sized fine-grained NE corpus at minimum cost?* This work answers those questions.

To these ends the methodology we devised was designed to utilise the availability and growth of Arabic Wikipedia to develop a large and extendable fine-grained named entity corpus and gazetteer with minimum human intervention. The contributions of this paper are:

1. It introduces a two-level tagset for Wikipedia NEs;
2. It develops a large fine-grained automatic NE corpus using minimum human intervention;
3. It develops a large fine-grained gazetteer; and
4. It thoroughly evaluates the resulting corpus and gazetteer.

2 Arabic Wikipedia and Named Entity

Wikipedia is an extensive collaborative project on the web in which articles are published and reviewed by volunteers from around the world. Wikipedia includes

[1] Throughout this paper, Arabic words are represented in three variants: (Arabic word /HSB transliteration scheme (Habash et al., 2007) / "English translation")

International Joint Conference on Natural Language Processing, pages 392–400,
Nagoya, Japan, 14-18 October 2013.

271 different languages, with the Arabic version ranked 27th with more than 210,000 articles. The annual increase in the number of articles is 30% (Wikipedia, 2013). The actual relationship between the Named Entity and Wikipedia is that a large percentage of Wikipedia articles are about named entities (Alotaibi and Lee, 2012). This provided the motivation to utilise Wikipedia's underlying structure to produce the target resources.

To this end, it is beneficial to provide an overview of the critical aspects of the Wikipedia structure:

- **Articles:** These can be one of the following:

1. **Normal article:** Each article has a unique title and contains authentic content; i.e. textual data, images, tables, items and links, related to the concept represented in the title. These are in the majority.

2. **Redirected article:** These contain a specific tag to redirect the enquirer to a normal article. For example: for the redirected article titled (بريطانيا العظمى /bryTAnyA AlςĎmý/ 'Great Britain'), there is a redirected tag to (المملكة المتحدة /Almmlkħ AlmtHdħ/ 'United Kingdom'). This tag is written thus #REDIRECTED[[المملكة المتحدة]].

3. **Disambiguation article:** These are used to list all the article titles that share ambiguities.

- **Links types**: There are two types of links in Wikipedia and they are described below:

1. **Non-piped links:** this type of links denotes that the display phrase of the link and the article's title are the same. For example: [[London]].

2. **Piped links:** this type of link allows for the text that appears in the contextual data to be different from the actual article it refers to. For example: [[UK|United Kingdom]], where "UK" appears in the display text, while "United Kingdom" refers to the titles of the article.

   Throughout this paper, the terms "link" and "link phrase" are used interchangeably to refer to the same thing.

- **Connectivity:** Used links, of any type, in the contextual data of any normal article, provide connectivity and thereby an underlying structure for Wikipedia; we are seeking to utilise to achieve our goal.

3 Transforming Arabic Wikipedia into a Fine-grained NE corpus and Gazetteer

In this section we present in detail the approach advised to automatically develop a tagged fine-grained named entity corpus and gazetteer based on Arabic Wikipedia.

3.1 The Conciseness of the Approach

Our assumption regarding this work is as follows:

If we are able to classify Wikipedia articles into NE classes, we will then be able to map the resultant labelling back into contextualised linked phrases. This involves the following steps:

1. Defining a fine-grained taxonomy suitable to Wikipedia;

2. Classifying Arabic Wikipedia articles into a predefined set of fine-grained NE classes;

3. Mapping the results of the classification back to the linked phrases in the text;

4. Detecting successive mentions of NE that have not been associated with links, while taking into account the Arabic morphological variation of the NE phrase; and

5. Selecting sentences to be included in the final corpus.

3.2 Defining Fine-grained Semantic NE Classes

Sekine et al. (2002) proposed a hierarchical named entity taxonomy that is very fine, with 150 subclasses. The methodology they used to construct semantic classes relies on analysing the named entities in a newswire corpus, in addition to analysing the answer type for a set of questions used in a Text Retrieval Conference TREC-QA task. WordNet noun hierarchy is also used to shape the classes further. Two years later, Sekine and Nobat (2004) added an extra 50 classes and decomposed some classes, such as "disease" and numeric expression respectively. Although the spectrum of classes is very wide, the specific descriptions and definitions for each class strives to avoid overlap and ambiguity, making it difficult to define. This taxonomy has been applied to both English and Japanese.

Some NLP applications, such as QA have designed their own named entity classes, based on the criteria they believe to be the most valuable. Harabagiu et al. (2003) developed a named entity recognition component in which one level consists of 20 defined fine grained classes. Knowing that factoid type questions require named entities, Li and Roth (2006) defined a fine grained taxonomy to answer certain types of questions. Although, their two layer taxonomy covered 50 fine grain classes of different types, some types were unrelated to named entities such as definition, description, manner and reason. Based on the same trend, Brunstein (2002) presented a two-level taxonomy in which 29 answer types are subdivided into 105 subtypes. Other researchers have adopted and used their taxonomy for named entity taxonomy (Nothman et al., 2008).

It is evident that there is no widely agreed fine grained taxonomy that can be directly adopted into Arabic; although ACE taxonomy is a reasonable choice

in the sense that it organises granularity into two layers, i.e. coarse and fine grained. In the evaluation of ACE (2008), the number of fine grain classes is 45. This taxonomy is designed in two levels of granularities and frequently used in the news-wire domain. Moreover, two-level taxonomy allows us to map a tagset into different traditional schemes easily, such as CoNLL or MUC.

Thus, ACE (2008) taxonomy was selected and because it is designed for a news-wire domain we applied some amendments to tailor it for use in a relatively open domain corpus, such as Wikipedia. For example, there are many articles in Wikipedia about people in different subclasses, such as scientists, athletes, artists, politicians, etc. These fine classes are not included in ACE, as it only involves three sub-classes: the individual, group and indeterminate. Another modification is performed; a new class called "Product" is added. This modified taxonomy is presented in Table 1.

Caorse-grained Classes	Fine-grained Classes
*PER: Person**	*Politician*, Athlete*, Businessperson*, Artist*, Scientist*, Police*, Religious*, Engineer**, Group, *Other**.
ORG: Organisation	Government, Non-Governmental, Commercial, Educational, Media, Religious, Sports, Medical-Science, Entertainment.
LOC: Location	Address, Boundary, Water-Body, Celestial, Land-Region-Natural, Region-General, Region-International.
GPE: Geo-Political	Continent, Nation, State-or-Province, County-or-District, Population-Center, GPE-Cluster, Special.
FAC: Facility	Building-Grounds, Subarea-Facility, Path, Airport, Plant.
VEH: Vehicle	Land, Air, Water, Subarea-Vehicle, Unspecified.
WEA: Weapon	Blunt, Exploding, Sharp, Chemical, Biological, Shooting, Projectile, Nuclear, Unspecified.
*PRO:Product**	*Book*, Movie*, Sound*, Hardware*, Software*, Food*, Drug*, Other**.

Table 1: ACE (2008) modified taxonomy. The modified or added classes are represented with italics and asterisks

3.3 Wikipedia Document Classification

The aim of classifying Wikipedia articles is to produce a list of two tuples, like <article's title, fine-grained NE tag>. The following sub sections describe the steps taken to achieve this goal.

3.3.1 Fine-grained Document Annotation and Quality Evaluation

In order to classify Arabic Wikipedia articles into named entity classes, we manually annotated 4000 articles into two levels of granularity, i.e. coarse and fine grained, using the modified taxonomy shown in Table 1. Two Arabic natives were involved in the annotation process and the inter-annotator agreement between the annotators was calculated using Kappa Statistic (Carletta, 1996). Table 2 shows that the inter-annotator agreement was calculated for different sizes of documents, i.e. 500, 2000 and 4000. This revealed difficulties that might be encountered during the annotation process.

Level	$Kappa$: n=500	$Kappa$: n=2000	$Kappa$: n=4000
Coarse-grained	92	98	99
Fine-grained	80	95	97

Table 2: Inter-annotator agreement in coarse and fine grained levels

3.3.2 Features Engineering and Representation

We developed our classification model relying on the set of features proposed by Alotaibi and Lee (2012) as these score 90% on the F-measure for coarse grained level. The features were:

1. **Simple Features (SF):** which represent the raw dataset as a simple bag of words without further processing.
2. **Filtered Features (FF):** involving removing the punctuation and symbols, filtering stop words and normalising digits.
3. **Language-dependent Features (LF):** represent the tokens in their stem form.
4. **Enhanced Language-dependent Features (ELF):** involving tokenising the sentence and assigning parts of speech for each token. This allows filtering of the dataset by involving only nouns (for instance) in the classifier.

In addition, we extended this set of features by extracting two more features:

1. **First paragraph:** Instead of just relying on the first sentence as in (Alotaibi and Lee, 2012), we identified useful features spread across the first paragraph.
2. **Bigram:** By using this feature, we aim to examine the effects of the collocation of tokens. Here we added the representation of a bigram while still preserving the unigram.

We represent the feature space using the term frequency-inverse document frequency (tf-idf).

3.3.3 Fine-grained Document Classification Results

The annotated dataset was divided into training and test at 80% and 20% respectively. We chose the Support Vector Machine (SVM) and Stochastic Gradient Decent (SGD) as a probabilistic model for the classifier. In each round of the classification, we tested one set of features and selected the one that performed best.

Table 3 shows the overall results for the fine-grained classification. There are three main findings. First, both classifiers tend to perform in a very similar way; therefore, in practice, use of either classifier to perform the final classification for the whole Wikipedia dataset will be expected to deliver very similar results. The second finding is that, the bigram features have little effect when different features are set. Finally, the best result for both classifiers was achieved using the ELF_{Uni} feature.

Features set	SVM			SGD		
	P	R	F	P	R	F
SF_{Uni}	0.78	0.79	0.78	0.78	0.79	0.78
$\text{SF}_{Uni+Bigram}$	0.80	0.81	0.80	0.80	0.81	0.79
FF_{Uni}	0.80	0.81	0.80	0.81	0.82	0.80
$\text{FF}_{Uni+Bigram}$	0.81	0.82	0.81	0.81	0.82	0.81
LF_{Uni}	0.77	0.78	0.77	0.78	0.79	0.78
$\text{LF}_{Uni+Bigram}$	0.79	0.80	0.79	0.79	0.80	0.79
ELF_{Uni}	0.82	0.83	0.82	0.82	0.83	0.82
$\text{ELF}_{Uni+Bigram}$	0.81	0.82	0.81	0.82	0.82	0.81

Table 3: The average fine-grained classification results when using SGD and SVM over different features sets where (tf-idf) is applied

3.4 Compiling the Corpus

Compilation of the final corpus was achieved according to the pipeline steps as follows:

1. Prepare and extract the features for all Arabic Wikipedia datasets, according to the method presented in Section 3.3.2;

2. Train an SVM classifier using the training dataset (4000 articles);

3. For each Wikipedia article, classify the article into the target fine-grained NE class;

4. Prepare final list of all articles' titles and their tags; and

5. Detect successive mentions of the named entity that have not been associated with the link:

   As a convention, a linking phrase in the text of any Wikipedia article should only be assigned the first time it appears in context; successive mentions of the phrase appear with no link. Therefore, not all NE phrases are linked every time. Detecting successive mentions works by finding and matching possible NE phrases in the text that share similarity, to a certain extent, with each phrase in the list of linked NE phrases. The main goal of this step is to augment the plain text with NE tags and to address some of the lexical and morphological variations that arise when a named entity is contextualised. For example, a named entity of (سعود الفيصل /sʕwd AlfySl/ 'Saud Alfaisal') is expected to be repeated in context with either the first name (سعود /sʕwd/ 'Saud') or the last name (الفيصل /AlfySl/ 'Alfaisal') or both together. This can also be difficult when prefixes are used. For example (و لسعود /wlsʕwd/ 'and for Saud'). Therefore, we prepare for and match all the variations of prefixes that can be attached to the NE.

6. Produce the NE annotated corpus by selecting sentences to be included in the final corpus.

3.4.1 To Which Extent to Select Sentences to be Involved in the Final Corpus?

We decided to compile two versions of the developed corpus. The first version is called "WikiFANE_{Whole}", which means that we retrieved all the sentences from the articles. On the other hand, the second version, i.e. $\text{WikiFANE}_{Selective}$, is compiled by selecting only the sentences, which have at least one named entity phrase. This creates a Wikipedia corpus that has as high a density of tags as possible.

In this paper and for evaluation purposes, we compiled the corpus for more than 2 million tokens as shown in Table 4. Meanwhile, this methodology allows all of Arabic Wikipedia to become a tagged fine-grained NE corpus. Moreover, both versions of this dataset were freely distributed to the research community[2].

Corpus	# of sentences	# of tokens
WikiFANE_{Whole}	76821	2,023,496
$\text{WikiFANE}_{Selective}$	57126	2,021,177

Table 4: The total number of sentences and tokens for the compiled corpora

4 Introducing a Fine-grained Arabic NE Gazetteer

The process of classifying Wikipedia articles into NE classes provides the benefit of compiling a large Arabic NE Gazetteer at two levels of granularity. Based on our best knowledge, the only Arabic NE gazetteer currently available is that produced by Benajiba et al. (2007) covering only three traditional NE classes, i.e. PER, ORG and LOC. The size of this gazetteer

[2]The fine-grained Arabic NE corpora, i.e. WikiFANE_{Whole} and $\text{WikiFANE}_{Selective}$ are freely available at http://www.cs.bham.ac.uk/~fsa081/resources.html

is 4132 entities. Table 5 compares the distribution between ANERgazet and WikiFANE$_{Gazet}$. Due to the space limitation, we only present the coarse level distribution of WikiFANE$_{Gazet}$. It is clearly shown that, WikiFANE$_{Gazet}$ has superiority in the sense of type and coverage. The gazetteer produced is freely available to the research community to use and extend[3].

Class	ANERgazet	WikiFANE$_{Gazet}$
PER	1920	30821
ORG	262	6664
LOC	1950	1424
GPE	NA	20785
FAC	NA	2182
VEH	NA	518
WEA	NA	274
PRO	NA	5624
Total	4132	68355

Table 5: The distribution of named entities for different gazetteers across coarse-grained NE classes

5 Evaluation and Results

To evaluate the fine-grained NE corpus and gazetteer produced, we conducted a set of thorough experiments. The aims of the evaluation were to answer the following questions:

- What is the quality of the corpus produced and the gazetteer in terms of annotation?
- How efficient is the NE classifier when used with WikiFANE$_{Whole}$ and WikiFANE$_{Selective}$ and tested over cross-domain and within-domain datasets?

5.1 Evaluating the Annotation Quality

The performance of document classification across all Wikipedia articles is crucial to avoid error propagation from the document classification stage when compiling the final version of the annotated corpus. Therefore, the first evaluation focused on this aspect. After classifying all articles to the target NE classes, we drew another 4000 articles, to be represented as a sample for all Wikipedia articles, and manually annotated them. The selection of the articles was made by selecting the first 4000 articles with identical glyphs to those used most frequently in other Wikipedia articles. This criteria ensured that the most frequent NE were classified properly with a minimum error rate. After this, we calculated the inter-annotation agreement between the manually annotated, gold-standard documents, and that classified based on step 3 in Section 3.4. Table 6 shows the result for both levels of granularity. The overall Kappa for the fine-grained level is 82.6% and this is consistent with the results shown in Section 3.3.3. This gives the impression that, the error rate is at a minimum, even when performing the classification across all Wikipedia articles with small amounts of training data.

[3]The fine-grained Arabic NE gazeetter WikiFANE$_{Gazet}$ is freely available at http://www.cs.bham.ac.uk/~fsa081/resources.html

Level	Accuracy	Overall Kappa
Coarse-grained	85.8	84.02
Fine-grained	82.9	82.6

Table 6: Inter-annotation agreement between the classified articles and the gold-standard

Features
Lexical features
Current token
Two tokens before and after the current token
First and last three characters of the token
Length of the token
The tag of the previous token
Morphological features
Gender
Number
Person
Syntactical features
Part of speech
Base phare chunck
External knowledge features
The token appears in gazeetter

Table 7: The set of language dependent and independent features extracted to be used by the classifier

5.2 Evaluating the Corpus Developed by Learning NE Classifier

This evaluation was designed to evaluate the corpus developed by using it as training data to test it over cross-domain and within-domain datasets. Moreover, this assists evaluation of the efficiency of using gazetteer as external knowledge resource. We parsed the different datasets and tokenised the sentences using AMIRA (Diab, 2009) relying on the scheme (Conjunction + Preposition + Prefix). The concept behind using this tokenisation scheme is that, the notable sparseness issues regarding Arabic NE are caused by agglutination of the prefixes. In this scheme, we guaranteed that the named entities like (**خالد** /xAld/ 'Khalid') in the training data also refer to (**و لخالد** /wlxAld/ 'and for Khalid') in the test data. This happens by tokenising the words and splitting the prefixes, so the result will be three different tokens (**و** /w/ 'and'), (**ل** /l/ 'for') and (**خالد** /xAld/ 'Khalid').

We extracted traditional sets of features at different levels; including lexical, morphological, syntactical and external knowledge. Table 7 summarises the features used where a window of five features are encoded in the classifier including the current position.

The following set of experiments was conducted relying on the Conditional Random Field (CRF) probabilistic model to perform the sequence labelling. In all the experiments, we divided the datasets into training and test at 80% and 20% respectively. We used the three metrics, precision, recall and F-measure, to evaluate the results.

5.2.1 Tags Distribution for the Gold-standard Newswire-based NE Corpora and WikiFANE

Different corpora have been used by researchers to develop NER. The first one is ANERcorp, which developed by Benajiba et al. (2007) and is freely accessible. It is a 150K news-wire based corpus tagged with CoNLL traditional coarse classes, i.e. PER, ORG, LOC and MISC. ACE produced two datasets named ACE 2004[4] and ACE 2005[5] which are subject to a costly licence. This prevents us using those corpora in the evaluation. However, ACE also produced a multilingual small corpus called REFLEX Entity Translation Training/DevTest (REFLEX for short), which consists of about 60K of tokens with two levels of classes. This is divided according to its origin into news-wire (NW), treebank (TB) and web blogs (WL). We used both the ANERcorp and the Arabic portion of REFLEX as gold-standard corpora to conduct the evaluation.

Table 8 shows the tag distribution for each corpus per class and the total per token and phrase. We use (NA) as an indication of no availability in the dataset. It is clearly shown that WikiFANE$_{Selective}$ has wider distributed tags compared with WikiFANE$_{Whole}$.

5.2.2 Gazetteer Evaluation

Using gazetteer as an external knowledge source in NER helps to boost the performance of NER (Carreras et al., 2002). To evaluate the gazetteer produced, we learned the classifier by news-wire dataset one at a time. Each time, we evaluated the presence and absence of WikiFANE$_{Gazet}$. Due to ANERcorp dataset being coarse-grain level, we decided to map the REFLEX dataset to the same scheme used by ANERcorp. In addition, we eliminated the MISC class used by ANERcorp because there is no direct equivalent in REFLEX. Three main points arose from this experiment. First, the F-measure increased by at least 2 points for all datasets, showing the overall positive effect of the developed gazetteer. Second, the recall metric clearly boosted the classifier enabling retrieval of more NE phrases than would be possible without WikiFANE$_{Gazet}$. Third, the TB sub-dataset of REFLEX showed dramatic improvement in comparison with other datasets, because that TB dataset had comparatively less noise.

[4]http://www.ldc.upenn.edu/Catalog/catalogEntry.jsp?catalogId=LDC2005T09

[5]http://www.ldc.upenn.edu/Catalog/catalogEntry.jsp?catalogId=LDC2006T06

Coarse-grained classes	ANERcorp	REFLEX	WikiFANE $whole$	WikiFANE $selective$
PER	6505	2701	37009	84757
ORG	3454	2457	14133	35479
LOC	5069	220	7533	9886
MISC	1707	NA	NA	NA
GPE	NA	3472	70523	94099
FAC	NA	175	3651	6790
VEH	NA	47	269	482
WEA	NA	25	612	748
PRO	NA	NA	7717	14063
Total (token level)	16735	9097	141447	246304
Total (phrase level)	11175	7566	96562	142932

Table 8: The distribution of the coarse-grained NE tags across different corpora

		No Gazetteer			WikiFANE$_{Gazet}$		
Corpus		P	R	F	P	R	F
ANERcorp		87.13	69.27	77.18	87.86	72.34	79.35
REFLEX	NW	88.51	69.37	77.78	88.21	72.79	79.76
	TB	79.09	70.16	74.36	89.20	76.61	82.43
	WL	83.78	62.23	71.41	84.69	66.61	74.57

Table 9: The comparison for using WikiFANE$_{Gazet}$ as external knowledge over news-wire dataset

5.2.3 Cross-domain Evaluation

The purpose of cross-domain evaluation is to train the classifier on a certain domain and then test this over different datasets with different domains or genres. The aim behind this experiment is to evaluate the effect when using WikiFANE$_{Whole}$ and WikiFANE$_{Selective}$ as training data versus news-wire domain datasets. This experiment helps to clarify the suitability of using WikiFANE as a relatively open domain corpus. It is evident from Table 10 that, self-training of ANERcorp and REFLEX produces the best performance. Meanwhile, there are some interesting findings. Even though REFLEX is a news-wire based corpus, its performance is dramatically lower when it is used as training dataset and tested over ANERcorp. This is also the case when training ANERcorp and testing it over REFLEX. This implies that, even within the same domain, news-wire, there is less generalisability for the current news-wire dataset across different datasets. Another interesting finding is that, the version of WikiFANE$_{Selective}$ performs better than WikiFANE$_{Whole}$ on different test sets, except for with ANERcorp. This might be because WikiFANE$_{Selective}$ has a greater tag density than WikiFANE$_{Whole}$, which leads to more positive examples in the dataset.

Training	Testing											
	ANERcorp			REFLEX								
				NW			TB			WL		
	P	R	F	P	R	F	P	R	F	P	R	F
ANERcorp	87.86	72.34	79.35	80.60	58.38	67.71	79.31	64.92	71.40	74.23	52.55	61.54
REFLEX	73.57	50.07	59.59	88.21	72.79	79.76	89.20	76.61	82.43	84.69	66.61	74.57
WikiFANE$_{Whole}$	81.53	43.10	56.39	71.43	37.84	49.47	84.11	51.21	63.66	71.43	36.50	48.31
WikiFANE$_{Selective}$	88.10	37.52	52.62	86.99	42.16	56.80	86.49	51.61	64.65	84.43	37.59	52.02

Table 10: The result of cross-domain evaluation

Corpus		P	R	F
ANERcorp + WikiFANE$_{Selective}$		90.40	58.21	70.81
REFLEX + WikiFANE$_{Selective}$	NW	90.55	62.16	73.72
	TB	86.52	62.10	72.30
	WL	86.01	52.74	65.38

Table 11: The result of combining WikiFANE$_{Selective}$ with news-wire corpora

To elaborate more on cross-domain evaluation we evaluated the merging of WikiFANE$_{Selective}$, since it performed best in the previous experiment, with both ANERcorp and REFLEX. The idea behind this experiment was to understand how the classifier performs when different domains and genera are combined together. The most notable findings, as shown in Table 11 are that, the recall metric shows a sharp drop in all datasets. However, the precision shows high scores, suggesting the Wikipedia corpus is strong in difference when compared with the news-wire domain.

5.2.4 Within-domain Evaluation

The traditional practice of learning NE classifier is to draw the training and test datasets from single domain. Therefore, we divided WikiFANE$_{Whole}$ and WikiFANE$_{Selective}$ into training and test for 80% and 20% respectively and then training the CRF classifier on WikiFANE$_{Whole}$ and WikiFANE$_{Selective}$ separately with and without the injection of the WikiFANE$_{Gazet}$ as an external knowledge source. Table 12 shows that, the use of WikiFANE$_{Gazet}$ creates a notable improvement across datasets by at least 3 points on the F-measure. In addition, WikiFANE$_{Selective}$ has a slightly superiority over WikiFANE$_{Whole}$ advising that both datasets are performing at a promising level of accuracy.

6 Related Work

A promising trend in the research is towards automatically developing an annotated NE corpus that extends beyond both traditional classes and the domain of newswire, in order to create novel resources. One of the earliest of these approaches was presented by An et al. (2003) in which the web was used to build a target corpus, using bootstrapping to build an annotated NE corpus. A further approach utilises parallel corpora to build an NE corpus automatically. This relies on the suggestion that once one corpus is annotated then other parallel corpora can be easily annotated using projection. Ehrmann et al. (2011) developed multilingual NE corpora for English, French, Spanish, German and Czech. Similarly, Fu et al. (2011) developed a Chinese annotated NE corpus exploiting an English aligned corpus. The difference here is that the alignment is conducted between both corpora at the word-level.

Beyond the newswire-based corpora, Wikipedia becomes more attractive for different NLP tasks. Some researchers have exploited the unrestricted accessibility of Wikipedia to establish an automatic fully annotated NE corpus with different granularity; meanwhile others are merely focusing on partially utilising Wikipedia to achieve specific goals, such as developing a NE gazetteer (Attia et al., 2010) or classifying Wikipedia articles into NE semantic classes (Saleh et al., 2010).

Tkatchenko et al. (2011) expanded the classification into an 18 fine-grain taxonomy extracted from (BNN). To prepare training data for use in the classification stage, a small set of seeds is constructed, as undertaken by Nadeau et al. (2006), in which a semi-supervised bootstrapping approach was used to construct long lists of entities in different fine-grain NE classes from the web. After the list is constructed, the entities are then intersected with Wikipedia articles so as to classify each article according to its target class. Therefore, a set of 40 articles per fine-grain class was produced for use in training with the Naïve Bayes and Support Vector Machine (SVM). Several similar features have been selected (e.g. (Saleh et al., 2010; Dakka and Cucerzan, 2008)).

Instead of relying on machine learning, Richman and Schon Richman and Schon (2008) defined a set of heuristics involving using assigned category links to classify articles. Phrasal patterns for each semantic NE class were specified when a matching article was classified; alternatively the procedure searched the upper level of categories to find candidates. These articles are still classified according to traditional coarse grain classes.

Closely related to our work are attempts to build a completely annotated NE corpus free from human in-

Corpus	PER			ORG			LOC			Overall		
	P	R	F	P	R	F	P	R	F	P	R	F
WikiFANE$_{Whole}$ *(no gaz)*	93.15	85.41	89.11	93.69	89.34	91.46	83.39	66.81	74.19	88.51	76.18	81.88
WikiFANE$_{Selective}$ *(no gaz)*	92.82	85.80	89.17	93.41	88.83	91.06	81.76	72.24	76.70	86.92	78.62	82.56
WikiFANE$_{Whole}$	97.35	88.61	92.78	97.74	93.10	95.36	84.58	70.37	76.83	91.10	79.62	84.98
WikiFANE$_{Selective}$	96.37	88.75	92.40	96.12	91.73	93.87	82.55	75.73	78.99	88.77	81.86	85.18

Table 12: The result for within-domain evaluation

tervention. The first attempt to transform Wikipedia into an annotated NE corpus was made by Nothman et al. (2008); they assumed that many NEs are associated with Wikipedia inter-links, i.e. the hyperlinks associated with a phrase in contexts pointing to another article. Therefore, the procedure first identified NEs using heuristics to exploit capitalisation, and then the target articles were classified into NE semantic classes. A bootstrapping approach is then used to extract seeds from a set of 1300 articles. Two distinguishing features were extracted per article; i.e. the head noun for the category links and the head noun for the definitional sentence. The corpus produced covered 60 fine-grained classes in two layers. An alternative approach to the same data set is presented by Tardif et al. (2009), in which the classification relies on supervised machine learning. Like Dakka and Cucerzan (2008), both Naïve Bayes and the Support Vector Machine (SVM) have been used as statistical interfaces for the purpose of classification. A total of 2311 articles have been manually annotated and a combination of structured and unstructured features extracted.

The corpus produced by Nothman et al. (2008) has been thoroughly experimented with to evaluate the impact of its performance. Three different gold-standard corpora, i.e. MUC, CoNLL and BNN, were used for comparative purposes and separate models built for each corpus. The experiment showed that, when in conjunction with other gold-standard corpora the Wikipedia-based corpus could raise their performance; it also performs well for non-Wikipedia texts (Nothman et al., 2009).

7 Conclusion

We presented a methodology to develop a large size fine-grained named entity corpus and gazetteer using an automatic approach. This involved recruiting document classifications. Using this methodology, we produced constantly evolving NE resources that will exploit the yearly growth rate of Arabic Wikipedia.

The freely fine-grained NE corpus and gazetteer produced when used on their own are of a very promising quality and extend the scope of research beyond traditional NE tasks.

References

ACE. 2008. Ace (automatic content extraction) english annotation guidelines for entities, 06. [accessed 10 April 2013].

Fahd Alotaibi and Mark Lee. 2012. Mapping Arabic Wikipedia into the named entities taxonomy. In *Proceedings of COLING 2012: Posters*, pages 43–52, Mumbai, India, December. The COLING 2012 Organizing Committee.

Joohui An, Seungwoo Lee, and Gary Geunbae Lee. 2003. Automatic acquisition of named entity tagged corpus from world wide web. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics*, pages 165–168. Association for Computational Linguistics.

Mohammed Attia, Antonio Toral, Lamia Tounsi, Monica Monachini, and Josef van Genabith. 2010. An automatically built named entity lexicon for arabic. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC10)*, Valletta, Malta. European Language Resources Association (ELRA).

Yassine Benajiba, Paolo Rosso, and José Miguel Benedíruiz. 2007. Anersys: An arabic named entity recognition system based on maximum entropy. In Alexander Gelbukh, editor, *Computational Linguistics and Intelligent Text Processing*, volume 4394 of *Lecture Notes in Computer Science*, pages 143–153. Springer Berlin / Heidelberg.

Y. Benajiba, I. Zitouni, M. Diab, and P. Rosso. 2010. Arabic named entity recognition: Using features extracted from noisy data. In *Proceedings of the ACL 2010 Conference Short Papers*, pages 281–285, Uppsala, Sweden. Association for Computational Linguistics.

Ada Brunstein. 2002. Annotation guidelines for answer types. LDC2005T33 [accessed 02 January 2012].

Jean Carletta. 1996. Assessing agreement on classification tasks: the kappa statistic. *Comput. Linguist.*, 22(2):249–254, June.

Xavier Carreras, Lluis Marquez, and Lluís Padró. 2002. Named entity extraction using adaboost. In *proceedings of the 6th conference on Natural language learning-Volume 20*, pages 1–4. Association for Computational Linguistics.

Wisam Dakka and Silviu Cucerzan. 2008. Augmenting wikipedia with named entity tags. In *Proceedings of the 3rd International Joint Conference on Natural Language Processing*, pages 545–552, Hyderabad, India. Asian Federation of Natural Language Processing.

Mona Diab. 2009. Second generation amira tools for arabic processing: Fast and robust tokenization, pos tagging, and base phrase chunking. In *2nd International Conference on Arabic Language Resources and Tools*.

Maud Ehrmann, Marco Turchi, and Ralf Steinberger. 2011. Building a multilingual named entity-annotated corpus using annotation projection. In *Proceedings of Recent Advances in Natural Language Processing (RANLP)*, pages 118–124, Hissar, Bulgaria.

Ruiji Fu, Bing Qin, and Ting Liu. 2011. Generating chinese named entity data from a parallel corpus. In *Proceedings of the 5th International Joint Conference on Natural Language Processing*, pages 264–272, Chiang Mai, Thailand. Asian Federation of Natural Language Processing (AFNLP).

Nizar Habash, Abdelhadi Soudi, and Tim Buckwalter. 2007. On arabic transliteration. In *Arabic Computational Morphology*, volume 38 of *Text, Speech and Language Technology*, pages 15–22. Springer Netherlands.

Sanda Harabagiu, Dan Moldovan, Christine Clark, Mitchell Bowden, John Williams, and Jeremy Bensley. 2003. Answer mining by combining extraction techniques with abductive reasoning. In *Proceedings of 12th Text Retrieval Conference*, volume 2003, pages 375–382. NIST.

Xin Li and Dan Roth. 2006. Learning question classifiers: the role of semantic information. *Natural Language Engineering*, 12(03):229–249.

David Nadeau, Peter D. Turney, and Stan Matwin. 2006. Unsupervised named-entity recognition: Generating gazetteers and resolving ambiguity. *Advances in Artificial Intelligence*, pages 266–277.

Elisa Noguera, Antonio Toral, Fernando Llopis, and Rafael Muńoz. 2005. Reducing question answering input data using named entity recognition. In *Text, Speech and Dialogue*, pages 428–434. Springer.

Joel Nothman, James Curran, and Tara Murphy. 2008. Transforming wikipedia into named entity training data. In *Proceedings of the Australian Language Technology Association Workshop*, pages 124–132, Hobart, Australia. ALTA.

Joel Nothman, Tara Murphy, and James Curran. 2009. Analysing wikipedia and gold-standard corpora for ner training. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, pages 612–620, Athens, Greece. Association for Computational Linguistics.

Alexander Richman and Patrick Schon. 2008. Mining wiki resources for multilingual named entity recognition. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1–9, Columbus, Ohio, USA. Association for Computational Linguistics.

Iman Saleh, Kareem Darwish, and Aly Fahmy. 2010. Classifying wikipedia articles into ne's using svm's with threshold adjustment. In *Proceedings of the 2010 Named Entities Workshop*, pages 85–92, Uppsala, Sweden. Association for Computational Linguistics.

Satoshi Sekine and Chikashi Nobat. 2004. Definition, dictionaries and tagger for extended named entity hierarchy. In *Proceedings of the 4th International Conference on Language Resources And Evaluation*, pages 1977–1980, Lisbon, Portugal. ELRA.

Satoshi Sekine, Kiyoshi Sudo, and Chikashi Nobata. 2002. Extended named entity hierarchy. In *Proceedings of the third International Conference on Language Resources and Evaluation*, volume 2, Las Palmas, Spain. ELRA.

Khaled Shaalan and Hafsa Raza. 2007. Person name entity recognition for arabic. In *Proceedings of the 2007 Workshop on Computational Approaches to Semitic Languages: Common Issues and Resources*, pages 17–24, Prague, Czech Republic. Association for Computational Linguistics.

Sam Tardif, James Curran, and Tara Murphy. 2009. Improved text categorisation for wikipedia named entities. In *Australasian Language Technology Association Workshop 2009*, pages 104–108, Sydney, Australia.

Maksim Tkatchenko, Alexander Ulanov, and Andrey Simanovsky. 2011. Classifying wikipedia entities into fine-grained classes. In *Data Engineering Workshops (ICDEW), 2011 IEEE 27th International Conference on*, pages 212–217. IEEE.

Wikipedia. 2013. The statistic of arabic wikipedia, 05. [accessed 10 May 2013].