A Hybrid Approach to Features Representation for Fine-grained Arabic Named Entity Recognition

Fahd Alotaibi
School of Computer Science
University of Birmingham, UK
fsa081@cs.bham.ac.uk
Faculty of Computing
King Abdulaziz University, KSA
fsalotaibi@kau.edu.sa

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

Abstract

Despite considerable research on the topic of Arabic Named Entity Recognition (NER), almost all efforts focus on a traditional set of semantic classes, features and token representations. In this work, we advance previous research in a systematic manner and devise a novel method to represent these features, relying on a dependency-based structure to capture further evidence within the sentence. Moreover, the work also describes an evaluation of the method involving the capture of global features and employing the clustering of unannotated textual data. To meet this set of goals, we conducted a series of evaluations to evaluate different aspects that demonstrate great improvement when compared with the baseline model.

1 Introduction

Traditionally, the focus of Arabic NER has been on a very limited number of semantic classes, i.e. PERSON, ORGANISATION and LOCATION, utilising the newswire domain such as those described by Benajiba and Rosso (2008), Benajiba et al. (2010) and Abdul-Hamid and Darwish (2010) . This limits higher-level applications (such as question answering) from extracting in-depth knowledge and working on a relatively open domains.

This paper addresses the issue of a fine-grained NER of 50 classes for Arabic and presents a comprehensive set of experiments that evaluate innovative means of representing the features set. Thus, the contribution of this paper falls into different categories with unique outcomes, as follows:

1. A novel approach to representing the features is used, relying on dependency representation. This representation overcomes the drawback of current window-based representations of features.

2. The representation of global evidence involves clustering unannotated textual data, employing hierarchical clusters (Brown et al., 1992).

3. Due to the fact that there is no comparable work to use as a comparison in the task of Arabic fine-grained NER, a baseline model was developed, based on Conditional Random Fields (CRF), using the best features, as established and used elsewhere in the literature.

4. Development of publically available gold-standard fine-grained NER corpora[1] from two different genres, i.e. Newswire and Wikipedia.

Each contribution is discussed in more detail during in the remainder of this paper.

2 Arabic Fine-grained Named Entity Corpora

The majority of Arabic NER approaches are supervised, ensuring that the machine learns from an annotated corpus and aims to predict unseen text. This approach requires a reasonable bank of labelled data. This section examines the availability of such an annotated dataset at the fine-grained level, and the creation of gold-standard corpora.

[1] Available at: http://sourceforge.net/projects/arabic-named-entity-corpora/ and
Mirror at: http://fsalotaibi.kau.edu.sa/Pages-Arabic-NE-Corpora.aspx

Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers, pages 984–995, Dublin, Ireland, August 23-29 2014.

2.1 Available Corpora

One of the earliest corpora publically released was ANERcorp, developed by Benajiba et al. (2007). This is a newswire based corpus and follows the CoNLL format. It annotates into four coarse-grained classes: PERSON, ORGANISATION, LOCATION, and MISCELLANEOUS. This dataset has been extensively used such as in (Benajiba et al., 2008b; Benajiba et al., 2010; Abdul-Hamid and Darwish, 2010).

Among corpora applying a fine-grained level of classes are those released by the Linguistic Data Consortium[2] (LDC). They released two multilingual NE corpora including Arabic (Mitchell et al., 2005; Walker et al., 2006). Both corpora were used in the Automatic Content Extraction (ACE) technology evaluation, at the coarse-grained level only. However, these corpora are governed by a costly annual license, which prevents the researcher from accessing and utilising them. At present, we are not aware of a study tackling fine-grained Arabic NER using this dataset.

Alotaibi and Lee (2013) released fine-grained Arabic NE corpora - WikiFANE$_{Selective}$ and WikiFANE$_{Whole}$. These were built automatically using the Arabic version of Wikipedia and released under the Creative Commons Attribution-ShareAlike 3.0 Unported Licence.[3]. These corpora apply a similar annotation taxonomy to that of the ACE corpus, but deliver increased coverage through the inclusion of a new class, i.e. PRODUCT, which includes Books, Movies, Sound, Hardware, Software, Food, Drugs and Other. Moreover, the corpora divide the PERSON class into 10 fine-classes, in order to provide wider coverage (i.e. Politician, Athlete, Businessperson, Artist, Scientist, Police, Religious, Engineer and Group). It is notable that this taxonomy can be easily mapped into CoNLL and ACE at either the coarse or fine-grained levels.

2.2 Creating Gold-standard Fine-grained Named Entity Corpora

Since the aim of this paper is to conduct an in-depth experiment for fine-grained Arabic NE, we manually created gold-standard fine-grained NE corpora for Arabic, drawing on two different genres. This gives a critical benchmark for evaluation and comparison with the automatically constructed corpus.

The first corpus is newswire-based, using the same textual data appearing in ANERcorp. The complete corpus was re-annotated to the fine-grained level. The second corpus is drawn from the Arabic version of Wikipedia. The selection of articles was made using a random heuristic, i.e. selecting articles discussing a named entity and maintaining a fair level of distribution among the classes. Moreover, the amount of textual data drawn from the Wikipedia article was restricted by avoiding such elements as lists, headings, and captions on images and tables.

2.3 Annotation Strategy and Quality Evaluation

For both corpora, the two-level taxonomy presented by Alotaibi and Lee (2013) was applied. This consists of 8 coarse-grained classes and 50 fine-grained classes. An in-house tool to facilitate the annotation process was developed. Two independent graduate-level Arabic native speakers were engaged to annotate the entire corpora. They were provided with extended instructions to guide them in the annotation process and a number of feedback sessions were conducted in the early stages of the process to ensure that any difficulties could be resolved.

After its completion, the quality of the annotation was evaluated by calculating the inter-annotator agreement between both annotators. The entity F-measure was used to evaluate the inter-annotation agreement as in (Hripcsak and Rothschild, 2005; Zhang, 2013). The corpora were named NewsFANE$_{Gold}$ and WikiFANE$_{Gold}$, referring to News-based, and Wikipedia-based, Fine-grained Arabic Named Entity Gold corpus, respectively. Micro-averaging was used while matching exact phrases, in order to calculate the agreement. The size and the inter-annotator agreement of NewsFANE$_{Gold}$ is 170K of tokens and 91% while WikiFANE$_{Gold}$ is 500K of tokens and 89% .

[2] https://www.ldc.upenn.edu/

[3] Available at: http://www.cs.bham.ac.uk/~fsa081/resources.html
Mirror at: http://sourceforge.net/projects/arabic-named-entity-corpora/

Corpus	Token level	Phrase level
NewsFANE$_{Gold}$	10.7	6.7
WikiFANE$_{Gold}$	13.1	7.4
WikiFANE$_{Selective}$	10.8	6.4
WikiFANE$_{Whole}$	7.08	4.9

Table 1: The density of NEs on token and phrase levels

Corpus	Length							
	1	2	3	4	5	6	7	8
NewsFANE$_{Gold}$	58.19	30.77	8	1.73	0.82	0.21	0.2	0.04
WikiFANE$_{Gold}$	51.75	31.55	10.88	3.48	1.34	0.46	0.21	0.12
WikiFANE$_{Selective}$	48.27	37.95	10.22	2.98	0.41	0.11	0.05	0.01
WikiFANE$_{Whole}$	66.22	25.85	6.02	1.58	0.05	0.02	0.01	0.01

Table 2: The distribution of NE phrases relative to length.

3 Corpus-based Evaluation and Comparison

It is important to closely evaluate and compare different corpora. The nature of the distribution of NE phrases is expected to differ to some extent, affecting the performance of learning the probabilistic model. Therefore, the coverage of NE phrases related to different aspects was studied, including the distribution of density, length and semantic classes.

3.1 The Density of NE

The density represents the coverage of NEs at the level of tokens and phrases. As can be seen in Table 1, WikiFANE$_{Gold}$ has the greater density at both levels. This demonstrates that the Wikipedia-based gold corpus tends to represent more NE phrases in context than that of the newswire-based. Although WikiFANE$_{Gold}$ is 0.7% denser than NewsFANE$_{Gold}$ in the phrase level, it reveals a notable difference (2.4%) in the token level. This indicates that WikiFANE$_{Gold}$ possess a greater variety in the length of NE phrases than the newswire-based corpus. However, the automatically developed corpus, WikiFANE$_{Selective}$, has a similar density of coverage as NewsFANE$_{Gold}$ whereas the WikiFANE$_{Whole}$ demonstrates a low level of density, due to its method of compilation.

3.2 The Distribution of the Length of Named Entity Phrases

It can be seen in Table 2, NewsFANE$_{Gold}$ and WikiFANE$_{Whole}$ tend to have more single-word NE phrases than other corpora. When it comes to the newswire corpus, this is due to differences in the way the NE phrases are written in a newswire domain. On the other hand, the boundaries of multi-word NE phrases are difficult to detect, in Arabic, due to the fact that the language has a complex morphology. This is demonstrated in the Wikipedia corpora, i.e. the gold and the selective - less than half the NE phrases in WikiFANE$_{Selective}$ are single-word, with a slightly higher rate found in WikiFANE$_{Gold}$.

3.3 The Distribution of the Fine-grained Classes

This demonstrates the distribution of NE phrases into fine-grained classes according to their annotation. As shown in Figure 1, the majority of classes tend (to some extent) to have a relatively harmonic distribution. In general, the newswire-based corpus tends to include more NE phrases related to politics, government, commerce, nations and cities, whereas the automatically-built corpora score a very high frequency on NE types such as 'Nation' and 'Population-centre'. Moreover, WikiFANE$_{Gold}$ shows wide distribution on most of the fine-grained classes of 'PERSON', 'LOCATION', 'FACILITY', 'VEHICLE' and 'PRODUCT', compared to other corpora.

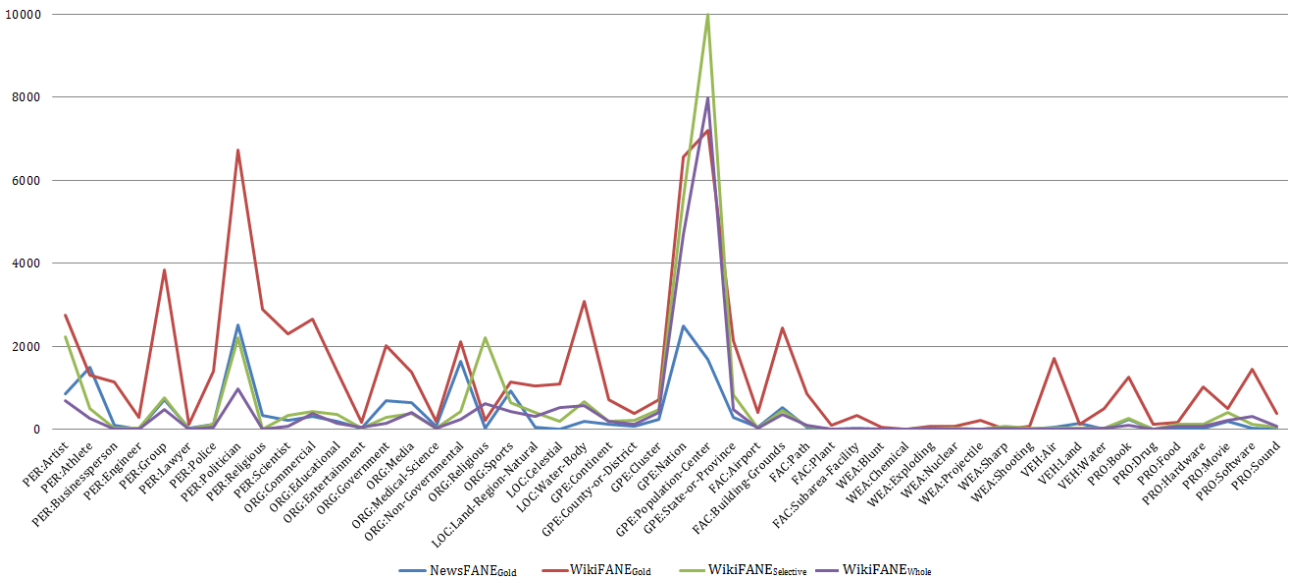


Figure 1: Distribution of Fine-grained Classes

4 The Baseline Model for Fine-grained Arabic NER

In order to prepare the baseline model and conduct successive experiments, the dataset for each corpus was divided into training, development and test. It is important to emphasise that, due to the limitations of computation power and the space allocated for the machine used, only a portion of WikiFANE$_{Selective}$ and WikiFANE$_{Whole}$ were selected with a size of ˜500K tokens each. The following table shows each corpus and its size.

Corpus	Type	Training	Dev	Test
NewsFANE$_{Gold}$	gold-standard	120K	25K	25K
WikiFANE$_{Gold}$	gold-standard	350K	75K	75K
WikiFANE$_{Selective}$	automatically-developed	354K	73K	73K
WikiFANE$_{Whole}$	automatically-developed	356K	72K	72K

Table 3: The size of the training, development and test for each corpus

Since there is no comparative work in the form of a fine-grained Arabic NER to use as a comparison, a baseline model based on Conditional Random Fields (CRF) was developed. It was decided to use the most successful features of the coarse-grained NER. For this purpose, the following features were extracted: ***Lexical and contextual features*** (current token, two tokens before and after the current token, first and last three characters of the token, and length of the token); ***Morphological features*** (gender, number and person); ***Syntactical features*** (part of speech and base phrase chunk); and ***External knowledge*** (the presence of the token in the gazetteer developed by Alotaibi and Lee (2013)). It was decided to use the BILOU scheme representation for the baseline model and successive experiments, as suggested by Ratinov and Roth (2009). The performance of the baseline model is presented in Table 4.

Corpus	Development			Test		
	P	R	F	P	R	F
NewsFANE$_{Gold}$	79.58	57.87	67.01	73.07	53.34	61.67
WikiFANE$_{Gold}$	62.19	43.67	51.31	68.13	44.78	54.04
WikiFANE$_{Selective}$	89.01	68.92	77.69	88.69	60.37	71.84
WikiFANE$_{Whole}$	82.35	49.83	62.09	84.27	58.63	69.15

Table 4: The results of the baseline model by learning CRF classifier with traditional features

5 Dependency based Features Representation

The current representation of the sequence tagging classifier involves using a predefined window of tokens (e.g. with size 5, including the current token) in order to capture local evidence. This representation has the following three drawbacks:

1. It is restricted to only capturing local evidence.
2. It fails to capture the relationship between dependent tokens, particularly for long sentences and multiword NE phrases.
3. Since Arabic has a relatively free word order, the window-based feature representation cannot capture the order variation for different examples.

In this paper, a new approach has been devised to utilise further evidence within a sentence in the classification process. The key idea informing this approach was to rely on the dependency-based representation of sentences in order to extract valuable features.

The dependency structure is one of syntactical representations, where a sentence is analysed by connecting its words in a word-to-word relationship. These relationships specify the head and dependent tokens in context, and assign a grammatical role for each token, e.g. subject, object and modifier.

To elaborate on the amount of knowledge that can be utilised based on the dependency structure, consider the following sentences:

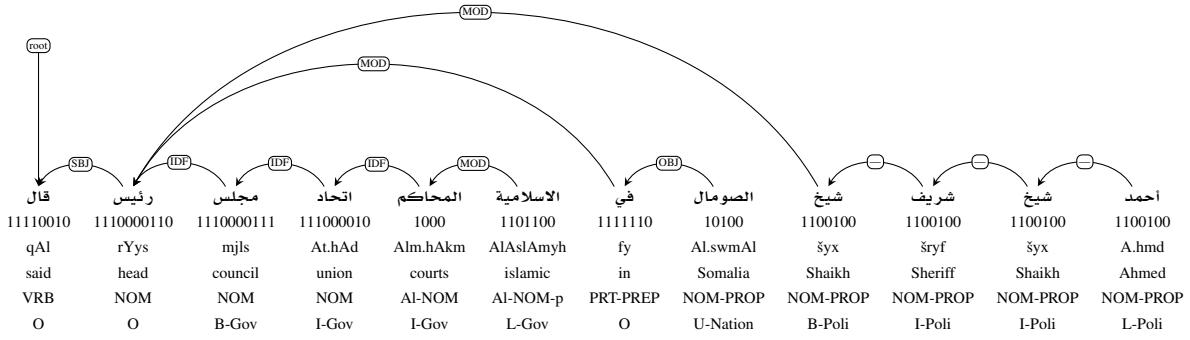
- (قال رئيس مجلس اتحاد المحاكم الاسلامية في الصومال شيخ شريف شيخ أحمد ...الخ /qAl rŷys mjls AtHAd AlmHAkm AlAslAmyħ fy AlSwmAl šyx šryf šyx OHmd fy ...Alx/ 'The head of the Council of the Islamic Courts Union, Sheikh Sharif Sheikh Ahmed, said in Somalia ...etc.')
- (يقول شارلز مورفي السياسي الانجليزي بعد الزيارة الأخيرة التي قام بها جون ميجور رئيس وزراء بريطانيا ...الخ /yqwl šArlz mwrfy AlsyAsy AlAnjlyzy bς d AlzyArħ AlOxyrħ Alty qAm bhA jwn myjwr rŷys wzrA' bryTAnyA ...Alx/ 'Charles Murphy, the English politician, said after the recent visit by John Major, Britain's prime minister ... etc.')
- (يذكر أن صلاد حسن انتخب رئيساً للصومال في اغسطس آب ٢٠٠٠ /yðkr On SlAd Hsn Antxb rŷysAã llSwmAl fy AγsTs Āb 2000/ 'It was mentioned that, Salad Hassan was elected as president of Somalia in August 2000')

The dependency representation and an English gloss of each example are shown in Figure 2. The parsed output includes a new set of information, which can be utilised as features, as follows:

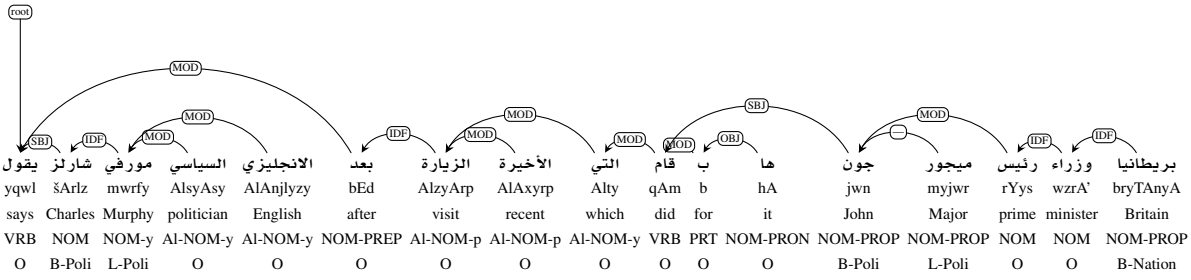
1. Head and Dependent Relation: The relationship between the head and the dependent is one of the most important features to capture. Consider the token (شيخ /šyx/ 'Shaikh'), in Figure 2a; the head (رئيس /rŷys/ 'the head of') is located far away and cannot be captured in the local window-based representation. Moreover, the vice versa relationship between the dependent and head is also useful. Consider the example in Figure 2b: the token (جون /jwn/ 'John') has two dependents (ميجور /myjwr/ 'Major') and (رئيس /rŷys/ 'Prime')[4] where the latter dependent (i.e. 'رئيس') gives a useful clue of the way in which it has been used in political contexts. The sequence of heads or dependents can also be utilised in the same way.

2. Sibling Relation: The sibling tokens are those dependent on the same head. Siblings can be located near each other in context, or appear at a distance. For example: the sibling of the token (شيخ /šyx/ 'Shaikh') is (مجلس /mjls/ 'council'), in Figure 2a, is expected to appear in a political context, which gives a clue towards the target NE class. Meanwhile, the token (في /fy/ 'in') is also a sibling, and can be avoided as it is a stop word. This is also the case in the example presented in Figure 2c, where the token (صلاد /SlAd/ 'Salad') is a sibling to the token (انتخب /Antxb/ 'elected'), which relates to the political context.

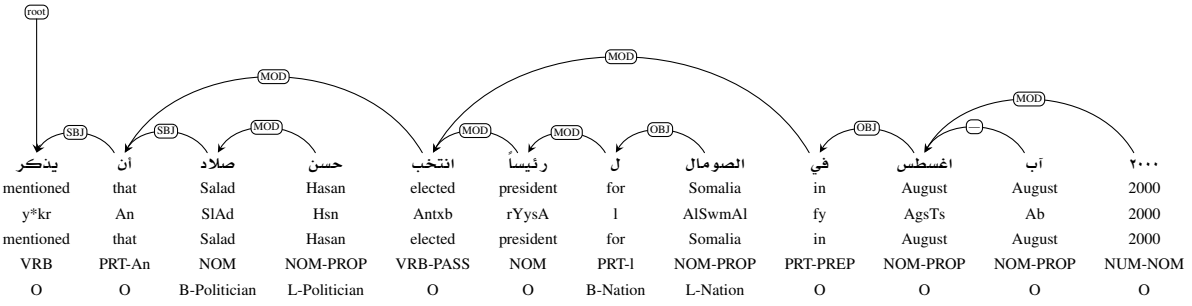
[4]Different contexts yield different English translation of the token "رئيس" as "the head of" and "Prime"



(a) The first example. (The second row represents the clusters according to the Brown algorithm)



(b) The second example



(c) The third example

Figure 2: The examples of a dependency structure. The rows show the Arabic token, Buckwalter transliteration, English gloss, POS and NE tag, respectively (the sentence is displayed left to right).

3. Syntactic Roles: The syntactical roles also benefit by being utilised to capture NE phrases in context. Among those with concern for NER are:

a. SBJ and OBJ: defines which subject and object roles are assigned to the head token of the NE phrase. For example, the tokens (**صلاد** /SlAd/ 'Salad') and (**شارلز** /šArlz/ 'Charles') are tagged as subjects.

b. IDF[5]*:* the Idafa chain is another important syntactical role, which helps to identify multiword NE phrases. For example: the token (**مورفي** /mwrfy/ 'Murphy') is tagged as an IDF of its previous token (**شارلز** /šArlz/ 'Charles'), where this indicates a multiword NE phrase. This is also the case for the example (**مجلس اتحاد المحاكم الإسلامية** /mjls AtHAd AlmHAkm AlIslAmyħ/ 'Council of the Islamic Courts Union') where all tokens are assigned an IDF role except the last token.

c. Flat relation (—): is a special role used by a CATiB pipeline parser for the sequence of proper nouns. For example: NE phrases such as (**شيخ شريف شيخ أحمد** /šyx šryf šyx OHmd/ 'Sheikh Sharif Sheikh Ahmed'), in which all tokens are assigned a flat relation.

[5]The naming of this abbreviation is used in CATiB to represent the syntactical role of idafa.

Corpus	Development			Test			+|-
	P	R	F	P	R	F	
NewsFANE$_{Gold}$	79.84	56.75	66.34	76.14	57.70	65.65	+3.98
WikiFANE$_{Gold}$	71.17	46.95	56.58	75.18	45.10	56.38	+2.34
WikiFANE$_{Selective}$	87.00	73.55	79.71	85.78	69.18	76.59	+4.75
WikiFANE$_{Whole}$	88.58	66.97	72.22	85.15	59.01	69.71	+0.56

Table 5: The results of the dependency-based features representation. ("+|-" represents the variation compared with the previous experiment)

5.1 Dependency-based Features set

The representation of the dependency structure presents each token as a node. A particular token (T) should have one node and only one head (H), except for the root, and zero or more dependents (D). A token (T) can have zero or more siblings (S), where they are connected, (i.e. are dependent), to the same head. Therefore, the following set of features has been extracted:

1. The current token T
2. POS of T
3. The presence of T in the Gazetteer
4. Syntactical role of T
5. Token of 1st, 2nd and 3rd Head H
6. Syntactical role of 1st, 2nd and 3rd H
7. POS of 1st, 2nd and 3rd H
8. Token of 1st, 2nd and 3rd Dependent D or 'NA' otherwise
9. Syntactical role of 1st, 2nd and 3rd D or 'NA' otherwise
10. POS of 1st, 2nd and 3rd D or "NA" otherwise
11. Token of 1st, 2nd and 3rd Sibling S or 'NA' otherwise
12. Syntactical role of 1st, 2nd and 3rd S or 'NA' otherwise
13. POS of 1st, 2nd and 3rd S or 'NA' otherwise

The 1st, 2nd and 3rd 'H' represent the parent, grandparent and great grandparent heads; while the 1st, 2nd and 3rd 'S' represent the first three siblings (if any).

5.2 Evaluation

It was decided to use the CATiB pipeline tool[6] (produced by Marton et al. (2013)), to parse all corpora and produce the set of features presented in the previous section. Since the POS tag set produced using the CATiB pipeline tool is very limited, it was decided instead to rely on the output of the AMIRA tokeniser and POS tagger produced by Diab (2009). The same classifier (CRF) was used, with a similar encoding scheme. Two experiments were conducted: the first was intended to evaluate the dependency-based representations. This was important in examining the effectiveness of the approach, compared with the window-based representation of local evidence. This is shown in Table 5, where in all corpora the performance of dependency-based representation alone outperforms that with window-based representation. The recall metrics reveal improvement across corpora, suggesting that the dependency-base representation has the ability to capture an increased number of NE phrases comparing to the traditional window-based representation.

The second experiment is intended to evaluate the integration in the classification process of dependency-based and window-based representations. This evaluation is expected to attain maximum benefit from both approaches in one model. The results in Table 6 demonstrate that the classifier tends to efficiently utilise both dependency-based and window-based representations in all corpora, apart from WikiFANE$_{Whole}$. The reason behind the degradation of the performance over the WikiFANE$_{Whole}$ dataset is due to the nature of the compiling of the corpus. Alotaibi and Lee (2013) state that this version includes entire sentences from Wikipedia articles, with no further filtering, ensuring that it is

[6]Not yet released to the public. We would like to thank the author for permission for its use.

Corpus	Development			Test			+|-
	P	R	F	P	R	F	
NewsFANE$_{Gold}$	82.08	57.77	67.81	80.21	61.58	69.68	+4.03
WikiFANE$_{Gold}$	89.31	49.11	63.37	83.34	50.48	62.88	+4.63
WikiFANE$_{Selective}$	87.03	73.29	79.57	87.31	76.17	77.81	+1.22
WikiFANE$_{Whole}$	82.44	57.91	68.03	75.88	52.45	62.03	-7.68

Table 6: The results of the hybrid approach using dependency-based and window-based features representation

possible to have sentences including NE phrases that are mistakenly assigned to 'O' class when using an automatic approach, as these NE phrases have no known destination in a Wikipedia article. This variety of mis-annotation is expected to propagate at this stage. It is worth noting that NewsFANE$_{Gold}$ and WikiFANE$_{Gold}$, as gold-standard corpora of different genres, reveal notable improvements of 4.03 and 4.63 F-measure respectively by using hybrid representation.

6 Further Exploiting of Global Evidences

Thus far, this study has examined the window-based and dependency-based representation of evidence, in order to increase the performance of the classification process. However, there is still room for improvement. Both approaches focus only at the sentence level. This section will investigate the approach to capturing global evidence. One means of achieving this is by utilising unannotated textual data, by clustering tokens into semantic groups based on context similarity. The reasoning behind this approach is that a NE token such as (الطائف /AlTAŷf/ 'Taif') (which is not seen in the training process) cannot be correctly classified, as it contains neither window-based nor dependency-based evidence in the training phase. Meanwhile, the token 'الطائف' is assigned to the same cluster of (لندن /lndn/ 'London') where the classifier knows that 'لندن' is a location. In this way, the knowledge capacity of the classifier has been broadened to a global level. A number of efforts have been undertaken for languages other than Arabic that demonstrate the usefulness of injecting clustering into NLP tasks, e.g. PCFG parsing (Candito and Crabbé, 2009) and dependency parsing (Koo et al., 2008). Utilising unannotated textual data in the supervised NER has already been variously studied with reference to English. The studies in (Turian et al., 2009; Turian et al., 2010; Tkachenko et al., 2012; Ratinov and Roth, 2009; Miller et al., 2004) reveal improvements when using the Brown clustering algorithm (Brown et al., 1992) to extract useful features.

This paper focuses on extracting a useful set of features from unannotated Arabic textual data, by relying on the Brown algorithm. We are not aware of any other study employing the Brown algorithm to Arabic textual data and in an Arabic NER task.

6.1 Brown Clustering and NER

The Brown clustering algorithm works by maximising the mutual information of bigrams. It uses hierarchical representation for the clusters. The hierarchal representation of the Brown clusters algorithm allows inclusion of different semantic levels of granularity. The output from the clustering delivers valuable information, which can be utilised by NER. This information can be divided into three categories:

1. The cluster of tokens belongs to the named entity category. For example, (شيكاغو /šykAγw/ 'Chicago') and (طوكيو /Twkyw/ 'Tokyo') belong to the same cluster, where both are NE type 'LOCATION'. In addition, (هديل /hdyl/ 'Hadeel') and (ممدوح /mmdwH/ 'Mamdooh') fall into similar clusters, and are both Personal NE.

2. The cluster of keyword tokens that have an informal insight to the target NE classes. For example, (كتائب /ktAŷb/ 'Brigades') and (منظمة /mnĎmħ/ 'Organisation') are keywords which infer the context of organisational NE. The context is expressed, for instance, as (كتائب شهداء الأقصى /ktAŷb šhdA' AlOqSý/ 'Al Aqsa Martyrs Brigades') or (منظمة العفو الدولية /mnĎmħ Alςfw

Aldwlyħ/ ‘Amnesty International’). Both head tokens in the NE phrases refer to the same cluster, which indicates the ‘ORGANISATION’.

3. The cluster of the head and dependent tokens the current token is pointing to. For example, the token (شيخ /šyx/ ‘Shaikh’), as shown in Figure 2a, is pointed to the head token (رئيس /rŷys/ ‘President’) where the ‘رئيس’ belongs to the cluster ‘1110000111’. This clustering knowledge permits the building of an abstract semantic representation for tokens. This implies that the token ‘رئيس’ can be replaced as (مدير /mdyr/ ‘Manager’) in other sentences where both tokens belong to the same cluster.

Further examples are presented in the Figure 3, where the group’s heading shows both name and cluster.

Locations: 0101101100	**First names: 000011111111101**
(بكين /bkyn/ ‘Beijing’)	(هديل /hdyl/ ‘Hadeel’)
(تكساس /tksAs/ ‘Texas’)	(حميدان /HmydAn/ ‘Homaidan’)
(طوكيو /Twkyw/ ‘Tokyo’)	(ممدوح /mmdwH/ ‘Mumdooh’)
Last names: 0000110001(01|10)	**Organisational keywords: 0111111111111011000**
(الساهر /AlsAhr/ ‘Alsaher’)	(كتائب /ktAŷb/ ‘battalions’)
(البخاري /AlbxAry/ ‘Albokhari’)	(جبهة /jbhħ/ ‘front’)
(الحازمي /AlHAzmy/ ‘Alhazmi’)	(منظمة /mnĎmħ/ ‘organization’)
Locational keywords: 011110110000	**Facility-related keywords: 101101100111011**
(مستوطنة /ktAŷb/ ‘settlement’)	(استاد /AstAd/ ‘stadium’)
(ضاحية /DAHyħ/ ‘suburb’)	(جسر /jsr/ ‘bridge’)
(محمية /mHmyħ/ ‘protectress’)	(مطار /mTAr/ ‘airport’)

Figure 3: Examples of the output of the Brown algorithm when applied to Arabic textual data.

6.2 Evaluation

The goal of this experiment was to evaluate the usefulness of injecting the clustering information from Brown algorithm into the supervised model. However, the actual size of the corpora mentioned in section 2.3 is too small to apply the Brown algorithm. Instead, a different set of different unannotated corpora, of a reasonably large size from different sources, was prepared for use in this experiment, as shown in Table 7.

Source of unannotated dataset	Size	Used for
NewsFANE$_{Gold}$ + Gigaword	1.17M	NewsFANE$_{Gold}$
WikiFANE$_{Gold}$ + 1/2(WikiFANE$_{Selective}$ & WikiFANE$_{Whole}$)	2.1M	WikiFANE$_{Gold}$
WikiFANE$_{Selective}$	2M	WikiFANE$_{Selective}$
WikiFANE$_{Whole}$	2M	WikiFANE$_{Whole}$

Table 7: Different textual data used in Brown algorithm

The first and second columns in Table 7 show the source of the unlabelled textual data used in the Brown algorithm and the respective size. The final column shows the target corpus using the knowledge in the CRF classifier.

Random stories were selected from Arabic Gigaword (Parker et al., 2011) as well as textual data from NewsFANE$_{Gold}$, to form unannotated data sized as 1.17M tokens. The Gigaword subset was selected due to the similarity of its genre to NewsFANE$_{Gold}$. The textual data for WikiFANE$_{Gold}$, and half of both WikiFANE$_{Selective}$ and WikiFANE$_{Whole}$ were compiled into one in order to induce clustering knowledge for WikiFANE$_{Gold}$.

The Brown algorithm was run in order to cluster the tokens into 1000 clusters, as suggested in (Miller et al., 2004; Liang, 2005; Ratinov and Roth, 2009; Tkachenko et al., 2012). The output of the Brown algorithm (which involves 1000 clusters) was injected as a set of features by extracting the clustering

Corpus	Development			Test			+|-
	P	R	F	P	R	F	
NewsFANE$_{Gold}$	86.13	70.38	77.46	81.66	68.36	74.42	+4.74
WikiFANE$_{Gold}$	77.80	62.36	69.23	79.87	60.19	68.64	+5.76
WikiFANE$_{Selective}$	89.17	74.04	80.90	88.64	73.18	80.17	+2.36
WikiFANE$_{Whole}$	90.39	69.97	78.88	84.98	65.00	73.66	+11.63

Table 8: The results of the injecting the output of Brown clustering into the CRF model

bits of (4, 6, 8, 10, 12) in a way that is similar to that presented by (Turian et al., 2010; Tkachenko et al., 2012). The reason behind this representation of the output is to allow a flexible level of grouping tokens into semantic clusters. For example, the tokens '**البخاري**' and '**الحازمي**' are clustered into '000011000101' and '000011000110', respectively, where both are personal NE. They share the first 10 bits of the cluster. This information allows for the extraction of useful knowledge to classify both tokens into the same class.

Table 8 shows notable improvement across all corpora. WikiFANE$_{Whole}$ and WikiFANE$_{Gold}$ score the highest, while other corpora gain improvements. It can be seen that the recall has sharply improved for approximately 7 to 13 points for NewsFANE$_{Gold}$, WikiFANE$_{Gold}$ and WikiFANE$_{Whole}$. This implies that the injecting of Brown clusters has improved the recall metric as a means of delimiting an increased number of NE phrases.

7 Related Work

This paper has addressed a series of issues, along with a discussion of the literature relevant to the context discussed in each section. Additional works of particular relevance are noted here. A large number of studies undertaking traditional Arabic NER have been developed, using a variety of methodologies to attain different goals. Using machine learning for the traditional task of NER has been addressed in different dimensions. Sequence labelling has also emerged, i.e. Maximum Entropy (Benajiba et al., 2007; Benajiba and Rosso, 2007); Support Vector Machine (Benajiba et al., 2008a); Conditional Random Fields (Benajiba and Rosso, 2008) and Structured Perceptron (Farber et al., 2008). Other hybrid approaches reliant on rule-based and ML are presented by (Shaalan and Oudah, 2013), a semi-supervised pattern is described in (AbdelRahman et al., 2010; Althobaiti et al., 2013) and the involvement of machine translation system to boost the performance of NER presented by (Zitouni and Florian, 2008). The researcher is not aware of any study tackling the fine-grained level of Arabic NER. Even that which has been developed for other languages (such as English) remains limited (Ling and Weld, 2012).

In terms of the representation of features, almost all studies in the Arabic NER apply the predefined window-based representation as examples when using this approach (Shaalan and Oudah, 2013; Benajiba et al., 2009). In English, Ratinov and Roth (2009) implemented two ways of capturing non-local features. The first approach is 'context aggregation'. This works by searching the entire document for a given token and returning the context of size two around each matched token. Ratinov and Roth (2009) limited the search to within 200 tokens. The second approach is 'extended prediction history', which looks up the 1000 previous tokens and counts the frequency of the label of the target class.

8 Conclusion

The majority of attempts to date to address NER focus on a limited number of semantic classes. This limitation has implications for other applications, such as question answering. This paper has presented an extended series of experiments and ideas, with the aim of constructing a fine-grained NER detailing resource creation to evaluation. Two approaches have been presented that rely on the output of the dependency parser and the clustering algorithm, instead of on a local window-based representation.

References

Samir AbdelRahman, Mohamed Elarnaoty, Marwa Magdy, and Aly Fahmy. 2010. Integrated machine learning techniques for arabic named entity recognition. *IJCSI International Journal of Computer Science*, 7(4):27–36.

Ahmed Abdul-Hamid and Kareem Darwish. 2010. Simplified feature set for arabic named entity recognition. In *Proceedings of the 2010 Named Entities Workshop*, pages 110–115, Uppsala, Sweden. Association for Computational Linguistics.

Fahd Alotaibi and Mark Lee. 2013. Automatically developing a fine-grained arabic named entity corpus and gazetteer by utilizing wikipedia. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 392–400, Nagoya, Japan, October. Asian Federation of Natural Language Processing.

Maha Althobaiti, Udo Kruschwitz, and Massimo Poesio. 2013. A semi-supervised learning approach to arabic named entity recognition. In *Proceedings of the International Conference Recent Advances in Natural Language Processing RANLP 2013*, pages 32–40, Hissar, Bulgaria, September. INCOMA Ltd. Shoumen, BULGARIA.

Yassine Benajiba and Paolo Rosso. 2007. Anersys 2.0: Conquering the ner task for the arabic language by combining the maximum entropy with pos-tag information. In *Proceedings of the Workshop on Natural Language-Independent Engineering, 3rd Indian Int. Conf. on Artificial Intelligence, IICAI-2007*, Pune, India.

Yassine Benajiba and Paolo Rosso. 2008. Arabic named entity recognition using conditional random fields. In *Proceedings of the Workshop on HLT & NLP Within the Arabic World. Arabic Language and Local Languages Processing: Status Updates and Prospects, 6th International Conference on Language Resources and Evaluation*, volume 8, pages 26–31, Marrakech, Morocco. LREC-2008.

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.

Yassine Benajiba, Mona Diab, and Paolo Rosso. 2008a. Arabic named entity recognition: An svm-based approach. In *Proceedings of 2008 Arab International Conference on Information Technology (ACIT)*, pages 16–18, Amman, Jordan. Association of Arab Universities.

Yassine Benajiba, Mona Diab, and Paolo Rosso. 2008b. Arabic named entity recognition using optimized feature sets. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 284–293, Honolulu, Hawaii. Association for Computational Linguistics.

Yassine Benajiba, Mona Diab, and Paolo Rosso. 2009. Using language independent and language specific features to enhance arabic named entity recognition. *IEEE Transactions on Audio, Speech and Language Processing. Special Issue on Processing Morphologically Rich Languages*, 17(5).

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.

Peter F Brown, Peter V Desouza, Robert L Mercer, Vincent J Della Pietra, and Jenifer C Lai. 1992. Class-based n-gram models of natural language. *Computational linguistics*, 18(4):467–479.

Marie Candito and Benoît Crabbé. 2009. Improving generative statistical parsing with semi-supervised word clustering. In *Proceedings of the 11th International Conference on Parsing Technologies*, pages 138–141. Association for Computational Linguistics.

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*.

Benjamin Farber, Dayne Freitag, Nizar Habash, and Owen Rambow. 2008. Improving ner in arabic using a morphological tagger. In *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, pages 2509–2514, Marrakech, Morocco. European Language Resources Association (ELRA).

George Hripcsak and Adam S Rothschild. 2005. Agreement, the f-measure, and reliability in information retrieval. *Journal of the American Medical Informatics Association*, 12(3):296–298.

Terry Koo, Xavier Carreras, and Michael Collins. 2008. Simple semi-supervised dependency parsing.

Percy Liang. 2005. *Semi-supervised learning for natural language*. Ph.D. thesis, Massachusetts Institute of Technology.

Xiao Ling and Daniel S Weld. 2012. Fine-grained entity recognition. In *Proceedings of the 26th Conference on Artificial Intelligence (AAAI)*.

Yuval Marton, Nizar Habash, and Owen Rambow. 2013. Dependency parsing of modern standard arabic with lexical and inflectional features. *Computational Linguistics*, 39(1):161–194.

Scott Miller, Jethran Guinness, and Alex Zamanian. 2004. Name tagging with word clusters and discriminative training. In *HLT-NAACL*, volume 4, pages 337–342. Citeseer.

Alexis Mitchell, Stephanie Strassel, Shudong Huang, and Ramez Zakhary. 2005. Ace 2004 multilingual training corpus [ldc2005t09], March 15. [accessed 20 December 2013].

Robert Parker, David Graff, Ke Chen, Junbo Kong, and Kazuaki Maeda. 2011. Arabic gigaword fifth edition [ldc2011t11], October 21. [accessed 20 December 2013].

Lev Ratinov and Dan Roth. 2009. Design challenges and misconceptions in named entity recognition. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning*, pages 147–155, Boulder, Colorado. Association for Computational Linguistics.

Khaled Shaalan and Mai Oudah. 2013. A hybrid approach to arabic named entity recognition. *Journal of Information Science*.

Maksim Tkachenko, Andrey Simanovsky, and St Petersburg. 2012. Named entity recognition: Exploring features. In *Proceedings of KONVENS*, pages 118–127.

Joseph Turian, Lev Ratinov, Yoshua Bengio, and Dan Roth. 2009. A preliminary evaluation of word representations for named-entity recognition. In *NIPS Workshop on Grammar Induction, Representation of Language and Language Learning*.

Joseph Turian, Lev Ratinov, and Yoshua Bengio. 2010. Word representations: a simple and general method for semi-supervised learning. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 384–394. Association for Computational Linguistics.

Christopher Walker, Stephanie Strassel, Julie Medero, and Kazuaki Maeda. 2006. Ace 2005 multilingual training corpus [ldc2006t06], February 15. [accessed 20 December 2013].

Ziqi Zhang. 2013. *Named entity recognition: challenges in document annotation, gazetteer construction and disambiguation*. Ph.D. thesis.

Imed Zitouni and Radu Florian. 2008. Mention detection crossing the language barrier. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 600–609. Association for Computational Linguistics.