EDRAK: Entity-Centric Data Resource for Arabic Knowledge

Mohamed H. Gad-Elrab **Mohamed Amir Yosef** **Gerhard Weikum**
Max-Planck-Institut für Informatik
Saarbrücken, Germany
{gadelrab|mamir|weikum}@mpi-inf.mpg.de

Abstract

Online Arabic content is growing very rapidly, with unmatched growth in Arabic structured resources. Systems that perform standard Natural Language Processing (NLP) tasks such as Named Entity Disambiguation (NED) struggle to deliver decent quality due to the lack of rich Arabic entity repositories. In this paper, we introduce EDRAK, an automatically generated comprehensive Arabic entity-centric resource. EDRAK contains more than two million entities together with their Arabic names and contextual keyphrases. Manual evaluation confirmed the quality of the generated data. We are making EDRAK publicly available as a valuable resource to help advance research in Arabic NLP and IR tasks such as dictionary-based Named-Entity Recognition, entity classification, and entity summarization.

1 Introduction

1.1 Motivation

Rich structured resources are crucial for several Information Retrieval (IR) and NLP tasks; furthermore, resources quality significantly influence the performance of those tasks. For example, building a dictionary-based Named Entity Recognition (NER) system, requires a comprehensive and accurate dictionary of names (Darwish, 2013; Shaalan, 2014). Problems like Word Sense Disambiguation (WSD) and Named Entity Disambiguation (NED) require name and context dictionaries to resolve the correct word sense or entity respectively (Weikum et al., 2012).

Arabic digital content is growing very rapidly; it is among the top growing languages on the Internet [1]. However, the amount of structured or semi-structured Arabic content is lagging behind. For example, Wikipedia is one of the main resources from where many modern Knowledge Bases (KB) are extracted. It is heavily used in the literature for IR and NLP tasks. However, the size of the Arabic Wikipedia is an order of magnitude smaller than the English one. Furthermore, the structured data in the Arabic Wikipedia, such as info boxes, are on average of less quality in terms of coverage and accuracy.

On the other hand, the amount and quality of the English structured resources on the Internet are unrivaled. The English Wikipedia is frequently updated, and contains the most recent events for example. It is important to leverage English resources in order to augment the currently poor Arabic ones. For example, both the English and Arabic Wikipedia have articles about `Christian Dior` and `Eric Schmidt` and hence the Arabic Wikipedia knows, at least, one potential Arabic name for both (the Arabic page title). However, Arabic Wikipedia knows nothing about `Christian Schmidt`[2], although, at least, his name can be learned automatically from only the English and Arabic Wikipedia's interwiki links.

To this end, it is compelling to automatically generate Arabic resources using cross-language evidences. This would help overcome the scarcity problem of Arabic resources and improve the performance of many Arabic NLP and IR tasks.

1.2 Contributions

Our contributions can be summarized into:

- Introducing EDRAK: an automatically generated Arabic entity-centric resource built on top of the English and Arabic Wikipedia's.
- Manual assessment of EDRAK, conducted by Arabic native speakers.

[1] `www.internetworldstats.com/stats7.htm`

[2] German Federal Minister of Food and Agriculture, 2015

Proceedings of the Second Workshop on Arabic Natural Language Processing, pages 191–200,
Beijing, China, July 26-31, 2015.

- Making EDRAK publicly available to the research community to help advance the field of Arabic NLP.

1.3 EDRAK Use-cases

EDRAK is an entity-centric Arabic resource that is a valuable asset for many NLP and IR tasks. For example, EDRAK contains a comprehensive dictionary for different potential Arabic names for entities gathered from both the English and Arabic Wikipedia's. Such dictionary can be used for building an Arabic **Dictionary-based NER** (Darwish, 2013).

In addition to the name dictionary, the resource contains a large catalog of entity Arabic textual context in the form of keyphrases. They can be used to estimate **Entity-Entity Semantic Relatedness** scores such as in Hoffart et al. (2012).

Furthermore, both the name dictionary and the entity contextual keyphrases are the corner-stone of state-of-the-art **Named Entity Disambiguation** (NED) systems (Hoffart et al., 2011).

Entities in EDRAK are classified under the type hierarchy of YAGO (Hoffart et al., 2013). Together with the keyphrases, EDRAK can be used to build an **Entity Summarization** system as in (Tylenda et al., 2011), or to build a **Fine-grained Semantic Type Classifier** for named entities as in (Yosef et al., 2012; Yosef et al., 2013).

2 Related Work

Different approaches to enrich Arabic resources have used cross-lingual evidences. Among the generated resources, some are entity-aware and useful for semantic analysis tasks. Others are purely textual dictionaries without any notion of canonical entities.

2.1 Entity-Aware Resources

Wikipedia, as the largest comprehensive online encyclopedia, is the most used corpus for creating entity-aware resources such as YAGO (Hoffart et al., 2013), DBpedia (Auer et al., 2007) and Freebase (Bollacker et al., 2008). Due to the limited size of Arabic Wikipedia, building strong semantic resources becomes a challenge. Several research efforts have been exerted to go beyond Arabic Wikipedia to construct a rich entity-aware resource.

AIDArabic (Yosef et al., 2014) is an NED system for Arabic text that uses an entity-name dictionary and an entity-context catalog extracted from Wikipedia. They leveraged Wikipedia titles, disambiguation pages, redirects, and incoming anchor texts to populate the *entity-name dictionary*. In addition, Wikipedia categories, incoming Wikipedia links page titles, and outgoing anchor texts were used in building the *entity-context catalog*. In order to overcome the small size of Arabic Wikipedia, they proposed building an *entity catalog* including entities from both the English and Arabic Wikipedia's. While their *catalog* was comprehensive, their *name dictionary* as well as *context catalog* suffered from the limited coverage in the Arabic Wikipedia. Hence, the recall of the NED task was heavily harmed.

Google-Word-to-Concept(GW2C) (Spitkovsky and Chang, 2012) is a multilingual resource mapping strings (i.e. names) to English Wikipedia concepts (including NEs). For *entity-names*, they harvested strings from Wikipedia titles, inter-Wikipedia links anchors, as well as manually created anchor texts from non-Wikipedia pages (i.e. web dump) with links to Wikipedia pages. The resource did not offer any *entity-context information*. The full resource contained 297M string-to-concept mapping. Nevertheless, the share of the Arabic records did not exceed 800K mapping. Finally, using **GW2C** in the entity linking task achieved above median coverage for English. In contrast, the results for the multilingual entity linking were less than the median.

BabelNet (Navigli and Ponzetto, 2012) is a multilingual resource built using Wikipedia entities and WordNet senses. They used the sense labels, Wikipedia titles from incoming links, outgoing anchor texts, redirects and categories as sources for disambiguation context. In addition, machine translation services were used to translate Wikipedia concepts to other languages. Nevertheless, translation was not applied on Named-Entities. They achieved good results using BabelNet as resource for cross-lingual Word Sense disambiguation (WSD).

2.2 Entity-free Resources

There exist several multilingual name dictionaries without any notion of canonical entities. Steinberger et al. (2011) introduced **JRC-Names**, a multilingual resource that includes names of organizations and persons. They extracted these names from multilingual news articles and Wikipedia. JRC-

Names contained 617K multilingual name variants with only 17K Arabic records.

Attia et al. (2010), built an Arabic lexicon Named-Entity resource using Arabic WordNet (Black et al., 2006) and Arabic Wikipedia. They extracted instantiable nouns from WordNet as Named-Entity candidates. Then, they used Wikipedia categories and inter-lingual Wikipedia pages to identify name candidates exploiting cross-lingual evidences. The resource contained 45K Arabic names along their correspondent lexical information.

Azab et al. (2013) compiled **CMUQ-Arabic-NET** Lexicon corpus, an English-Arabic names dictionary from Wikipedia as well as parallel English-Arabic news corpora. They used off-the-shelf NER system on the English side of the data. NER results were projected onto the Arabic side according to the word-alignment information. Additionally, they included Wikipedia inter-lingual links titles in their dictionary as well as coarse-grained type information (`PERSON` or `ORGANIZATION`).

3 High-level Methodology

Our objective is to produce a comprehensive Arabic entity repository together with rich entity *Arabic names dictionary* and entity *Arabic keyphrases catalog*. We augment an Arabic Wikipedia-based entity repository by translating English names and keyphrases. Off-the-shelf translation systems are not suitable for translating named entities (Al-Onaizan and Knight, 2002; Hálek et al., 2011; Azab et al., 2013). Therefore, we incorporate three translation techniques:

1. **External Name Dictionaries:** We harness the existing English-Arabic name dictionaries via semantic and syntactic equivalence, for example, if two strings from one or more dictionaries are linking to the same canonical entity, we consider them a potential translation of each other.

2. **Statistical Machine Translation:** We train an SMT system on English-Arabic parallel names corpora.

3. **Transliteration:** We build a transliteration system for persons names by training an SMT system on an English-Arabic parallel persons names corpora on the character level.

Data generated from all techniques are fused together to form a comprehensive Arabic resource obtained by translating an existing English one.

4 Creation of EDRAK

In this section, we start with describing EDRAK. Then, we explain the pre-processing steps applied on the data. The rest of the section explains in detail the creation process of EDRAK following the methodology explained in Section 3.

4.1 EDRAK in a Nutshell

EDRAK is an entity-centric resource that contains a catalog of entities together with their potential names. In addition, each entity has a contextual characteristic description in the form of keyphrases. Keyphrases and keywords are assigned scores based on their popularity and correlation with different entities.

EDRAK contains an entity catalog based on YAGO3 KB (Mahdisoltani et al., 2015), compiled from both English and Arabic Wikipedia's. We favored YAGO as our underlying KB over other available multilingual KBs because it is geared for precision instead of recall. Therefore, it is more salient for applying SMT techniques for example. We used the English Wikipedia dump of 12-January-2015 in conjunction with the Arabic dump of 18-December-2014 to build an Arabic YAGO3 KB.

EDRAK's *entity-name dictionary* is extracted from different pieces of Wikipedia that exist in YAGO3 KB. Namely, we harness Wikipedia page titles and redirects. In addition, we include YAGO3 *rdfs:labels* extracted from anchor texts and disambiguation pages in Wikipedia. *Entity context* is compiled from anchor texts, category names in the Wikipedia entity page. In addition, we include titles of Wikipedia pages linking to this entity.

The above data pieces extracted from the Arabic Wikipedia are included in EDRAK as it is, while those extracted from the English Wikipedia are translated/transliterated using one of the techniques introduced in the Section 3. We followed the same approach as in AIDA (Hoffart et al., 2011) to generate statistics about *entities importance* and *keyphrases weights*.

4.2 Data Pre-processing

Since Arabic is a morphologically-rich language, standard English text processing techniques are not directly suitable. Systems such as MADAMITA

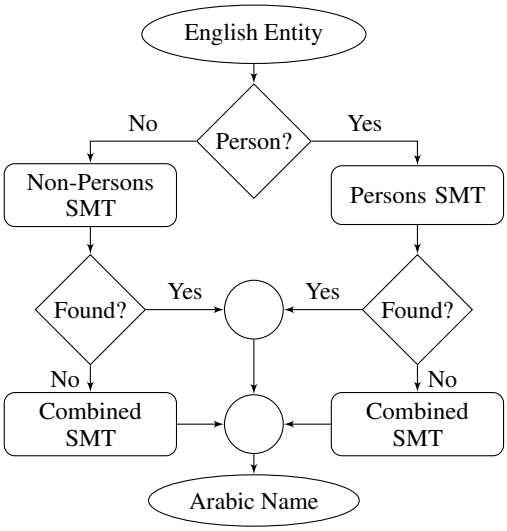

Figure 1: Architecture of Type-Aware Names Translation System

(Pasha et al., 2014) or Stanford Arabic Word Segmenter (Monroe et al., 2014) should be used to perform morphological-based pre-processing. Stanford Word Segmenter provides interpolatable handy Java API, hence has been used to pre-process the data. Text has been segmented by separating clitics, and normalized by *Removing Tatweel*, *Normalizing Digits*, *Normalizing Alif*, and *Removing Diacritics*. This helps achieving better coverage for our data, and computing more accurate statistics.

4.3 External Names Dictionaries

EDRAK harnesses *Google-Word-to-Concept (GW2C)* (Spitkovsky and Chang, 2012) multilingual resource in order to capture more names from the web. *GW2C* is created automatically without applying manual verification or post-processing. Therefore, it contains noise that should be filtered out. In order to include GW2C in EDRAK dictionary, we performed the following steps:

- **Language detection** We used off-the-shelf language detection tools developed by Shuyo (2010) to filter out non-Arabic records. Only 736K out of 297M were Arabic entries.

- **Filtering ambiguous names** We utilized the provided conditional probability scores to filter out generic anchor texts such as "Read more", "Wikipedia page" or "المزيد على ويكيبيديا". We ignore strings with conditional probability less than a threshold of 0.01.

	PER	NON-PER	ALL
CMUQ-Ar.	28,493	34,116	62,609
Wikipedia	33,962	79,699	128,790
Both	62,455	113,815	191,399

Table 1: Entity Names SMT Training Data Size

- **Name-level post-processing** We post-processed the data by applying normalization and data cleaning. (e.g. removing punctuation and URLs).

- **Mapping to EDRAK Entities** We used Wikipedia pages URLs to map extracted names from GW2C to EDRAK's Entity repository.

In addition to *GW2C*, we used lexical named-entities resources as look-up dictionary to translate English entity names. English names were matched strictly against those dictionaries to get the accurate Arabic names. We used the multilingual resource *JRC-Names* (Steinberger et al., 2011) that includes several name-variants along with partial language tags. After automatically extracting the Arabic records, English-Arabic pairs were included in our lookup dictionary. Similarly, we included*CMUQ-Arabic-NET* lexicon corpus (Azab et al., 2013) the lookup dictionary.

4.4 Translation

We trained cdec (Dyer et al., 2010), a full fledged SMT system, to translate English Names into Arabic ones. As training data, we fused a parallel corpus of English-Arabic names from multiple resources. We used a dictionary compiled from Wikipedia interwiki links together with *CMUQ-Arabic-NET* dictionary (Azab et al., 2013). While the latter contains name-type information, for the interwiki links, we leveraged YAGO KB to restrict our training data to only named-entities and to obtain semantic types information for each. 5% of the data have been used for tuning the parameters of SMT. The properties of the training data are summarized in Table 1.

We implemented two different translation paradigms. The first is depicted in Figure 1. We train three different system, on `PERSONS`, `NON-PERSONS` and a fallback system trained on `ALL`. In the first approach, depending on the entity semantic type, we try to translate its English

Table Name	Major Columns	Description
entity_ids	- id - entity	Lists all entities together with their numerical IDs.
dictionary	- mention - entity - source	Contains information about the candidate entities for a name. It keeps track of the source of the entry to allow application-specific filtering.
entity_keyphrases	- entity - keyphrase - source - weight	Holds the characteristic description of entities in the form of keyphrases. The source of each keyphrase is kept for application-specific filtering.
entity_types	- entity - types []	Stores YAGO semantic types to which this entity belongs.
entity_rank	- entity - rank	Ranks all entities based on the number of incoming links in both the English and Arabic Wikipedia. This can be used as a measure for entity prominence.

Table 2: Main SQL Tables in EDRAK

name using the corresponding system. If it fails, we switch to the fallback system. In the second COMBINED approach, we use the system trained on ALL dataset to translate all names regardless of the entity type.

In addition, we are translating Wikipedia Categories to be included in entities contextual keyphrases. To this end, we train the SMT system on English-Arabic parallel data of categories names harvested from Wikipedia interwiki links. The size of the training data is 43K name pairs, of which 5% have been used for tuning SMT parameters as well.

4.5 Transliteration

Recent research has focused on building Arabization systems that are geared towards transliteration general and informal text, without any special handling for entity names (Al-Badrashiny et al., 2014).

To this end, we had to build a transliteration system optimized for names. Transliteration is applicable on many NON-PERSON entities. However, applying it for such entities will create a lot of inaccurate entries that should be either fully or partially translated, or those that can only be learned from manually crafted dictionaries such as movie names. It is also worth noting that ORGANIZATION names that contain a person name such as "*Bill Gates Foundation*" will be correctly translated using the COMBINED system explained above.

Transliteration has been applied on PERSONS names only. We used the PERSONS part of the training data (Table 1) used for translation, and trained an SMT system on the character-level. 5% of the data have been used for parameter tuning of the SMT system. Each PERSON entity has English *FirstName* and *LastName*. Transliteration has been applied for each, and on a *FullName* composed by concatenating both.

5 Statistics and Technical Details

5.1 Technical Description

We are publicly releasing EDRAK for the research community. EDRAK is available in the form of an SQL dump, and can be downloaded from the *Downloads* section in AIDA project page http://www.mpi-inf.mpg.de/yago-naga/aida/. We followed the same schema used in the original AIDA framework (Hoffart et al., 2011) for data storage. Highlights of the SQL dump are shown in Table 2. EDRAK's comprehensive entity catalog is stored in SQL table entity_ids. Each entity has many potential Arabic names together stored in SQL table dictionary. In addition, each entity is assigned a set of Arabic contextual keyphrases stored in SQL table entity_keyphrases.

It is worth noting that sources of dictionary entries as well as entities keyphrases are kept in the schema (YAGO3_LABEL, REDIRECT, GIVEN_NAME, or FAMILY_NAME). Furthermore, generated data (by translation or transliteration) are differentiated from the original Arabic data extracted directly from the Arabic Wikipedia. Different generation techniques and data sources entail different data quality. Therefore, keeping data sources enables downstream applications to filter data for precision-recall trade-off.

Semantic Type	# entities
PERSON	1,220,032
EVENT	199,846
LOCATION	360,108
ORGANIZATION	196,305
ARTIFACT	359,071

Table 3: Number of Entities per Type in EDRAK

Technique	# of entries
Google W2C	241,104
CMUQ-Arabic-NET	23,338
JRC	4148
Translation	11,222,876
Transliteration	9,578,658

Table 4: Number of Entity-Name pairs per Generation Technique

5.2 Statistics

EDRAK is the largest publicly available Arabic entity-centric resource we are aware of. It contains around 2.4M entities classified under YAGO type hierarchy. The numbers of entities per high level semantic type are summarized in Table 3. The contributions of each generation technique are summarized in Table 4. Numbers show that automatic generation contributes way more entries than name dictionaries. In addition, translation delivers more entries than transliteration since it is applied on all types of entities (in contrast to only persons for transliteration).

The most similar resource to EDRAK is the one used in AIDArabic system to perform NED on Arabic text. However, AIDArabic resource is compiled solely from *manual* entries in both English and Arabic Wikipedia's such as Wikipedia categories, without incorporating any *automatic* data generation techniques. Therefore, the size of AIDArabic resource is constrained by the amount of Arabic names and contextual keyphrases available in the Arabic Wikipedia. In order to show the impact of our automatic data enrichment techniques, we compare the size of EDRAK to that of AIDArabic resource. Detailed statistics are shown in Table 5. Clearly, EDRAK is an order of magnitude larger than the resource used in AIDArabic.

	AIDArabic	EDRAK
Unique Names	333,017	9,354,875
Entities with Names	143,394	2,400,340
Entity-Name Pairs	495,245	21,669,568
Unique Keyphrases	885,970	7,918,219
Entity-Keyphrase Pairs	5,574,375	211,681,910

Table 5: AIDArabic vs EDRAK

5.3 Data Example

Many prominent entities do not exist in the Arabic Wikipedia, and hence do not appear in any Wikipedia-based resource. For example, Christian_Schmidt, the current German Federal Minister of Food and Agriculture, and Edward_W._Morley, a famous American scientist, are both missing in the Arabic Wikipedia[3]. EDRAK's data enrichment techniques managed to automatically generate reasonable potential names as well as contextual keyphrases for both. Table 6 lists a snippet of what EDRAK knows about those two entities.

6 Manual Assessment

6.1 Setup

We evaluated all aspects of data generation in EDRAK. Entity names belong to four different sources: *First Name*, *Last Name*, Wikipedia *redirects*, and *rdfs:label* relation which carries names extracted from Wikipedia page titles, disambiguation pages and anchor texts.

As explained in Section 4, we implemented two different name translation approaches, the first considers entity semantic type (which we refer to as **Type-Aware** system), and the second uses a universal system for translating all names (which is referred to as **Combined**).

Data assessment experiment covered all types of data against both translation approaches. Additionally, we conducted experiments to assess the quality of translating Wikipedia categories. Finally, we evaluated the performance of transliteration when applied on English person names. We randomly sampled the generated data and conducted an online experiment to manually assess the quality of the data.

[3] as of June 2015

Entity	Generated Arabic Names	Generated Keyphrases
Christian_Schmidt	جيسون	وزارة الدفاع الاتحادية الالمانية
	سشميد	سياسيون الاتحاد في بافاريا اجتماعية مسيحية
	شميت	مجتمع الاطلسي
	شميدت	وزارة الدفاع الالمانية الفيدرالية
	كرستيان	كريستيان شميدت مستقل
	كريستيان	وزراء المان الزراعة
	كريستيان تشميدت	هانز بيتر فريدرش
	كريستيان شميت	وزارة الدفاع الفيدرالية الالمانية
	كريستيان شميت مستقل	هانز بيتر فريدريك
	كريستيان شميدت	وزراء المان زراعة
	كريستيان شميدت مستقل	مجموعة الاطلسي
	كريستين	برلمانيون ألمان
		المجموعة الاطلسي
		سرطان الحكومة الثالثة
		هانز بيتر فريدريش
		سياسيون الاتحاد اجتماعية مسيحية في بافاريا
		كريستيان شميت مستقل
		سرطان الحكومة الثالث
		وزراء الزراعة المان
Edward_W._Morley	ادوار	كيميائيون فيزيائية امريكيون
	إدوارد	جائزة غيبس
	ادوارد	تجربة فيزو
	ادوارد دبليو مورلى	اكاديمية كيس و+ سترن
	ادوارد مورلي	فيزيائيون التجريبي
	ادوارد مورلى	جمعية فلكية الامريكية
	ادوارد وليامز مورلى	خريجو جامعة المحقق الفيدرالي الغربية
	ادوارد و+ مورلي	كاس جامعة كيس و+ سترن
	ادوارد و+ مورلى	فوهة مورلى
	ادوارد و+ يليامز مورلي	ف+ يزيائيون طيف
	ادوارد و+ يليامز مورلى	كيميائيون امريكيون فيزيائية الاميركى
	ادوار مورلى	الاكاديمية كيس و+ سترن
	ادورد	تاريخ الكيمياء البدنية
	اي و+ مورلى	الجمعية الامريكية فلكية
	دوارد	فائزون بوسام إليوت كريسون
	مرلي	التسلسل الزمني ل+ لكيمياء البدنية
	مورلي	مجلة شكوكية
	مورلى	غرب هارتفورد
	ميرلي	تجربة ميكلسون ومورلي
		اختبار فيزو

Table 6: Examples for Entities in EDRAK with their Generated Arabic Names and Keyphrases

	Approach	Source	Translations @ Top-K			Precision @ Top-K		
			1	2	3	1	2	3
Persons	Type-Aware	First Name	8	10	12	87.50	80.00	66.67
		Last Name	14	17	19	92.86	88.24	78.95
		rdfs:label	156	288	383	79.49	63.19	57.44
		redirects	113	210	285	69.91	57.62	50.18
	Combined	First Name	7	10	12	100.00	90.00	75.00
		Last Name	16	22	25	87.50	81.82	76.00
		rdfs:label	160	307	421	81.25	64.82	57.24
		redirects	108	210	288	67.59	60.00	54.51
	Transliteration	First Name	26	52	76	80.77	61.54	56.58
		Last Name	94	188	279	70.21	63.83	55.91
Non-Persons	Type-Aware	rdfs:label	269	519	742	53.16	43.16	36.66
		redirects	191	370	526	45.55	34.86	30.99
	Combined	rdfs:label	273	533	770	49.82	41.84	36.75
		redirects	195	378	539	46.67	39.42	34.69
Categories	Categories	Categories	118	234	340	67.80	52.99	46.18

Table 7: Assessment Results of Applying SMT for Translating Entities and Wikipedia Categories Names

6.2 Task Description

We asked a group of native Arabic speakers to manually judge the correctness of the generated data using a web-based tool. Each participant was presented around 150 English Names together with the top three potential Arabic translations or transliteration proposed by cdec (or less if cdec proposed less than three translations). Participants were asked to pick all possible correct Arabic names. Evaluators had the option to skip the name if they needed to. Each English Name was evaluated by three different persons.

6.3 Assessment Results

In total, we had 55 participants who evaluated 1646 English surface forms, that were assigned 4463 potential Arabic translations. Participants were native Arabic speakers that are based in USA, Canada, Europe, KSA, and Egypt. Their homelands span Egypt, Jordan, and Palestine. Translation assessment results are shown in Table 7. Evaluation results are given per entity type, translation approach and name origin. Since cdec did not return three potential translations for each name, we computed the total number of translations added when considering up to top one or two or three results. For each case, we computed the corresponding precision based on participants annotations.

6.4 Discussion

Data was randomly sampled from all generated data, and the size of each test set reflects the distribution of the sources included in the original data. For example, names originating from *rdfs:label* relation are an order of magnitude more than those coming from *FirstName*, and *LastName* relations.

The quality of the generated data varies according to the entity type, name source and generation technique. For example, the quality of translated Wikipedia *redirects* is consistently less than that of other sources. This is due to the nature of *redirects*. They are not necessarily another variation of the entity name. In addition, *redirects* tend to be longer strings, and hence are more error-prone than *rdfs:labels*. For example, "European Union common passport design" which redirects to the entity Passports_of_the_European_Union could not be correctly translated. Each token was translated correctly, but the final tokens order was wrong. Evaluators were asked to annotate such examples as wrong. However, such ordering problems are less critical for applications that incorporate partial matching techniques. Categories tend to be relatively longer than entity names, hence they exhibit the same problems as redirects.

Although the size of the evaluated *FirstName* and *LastName* data points is small, the assessment

results are as expected. Translating one token name is relatively an easy task. In addition, cdec returned only one or two translations for the majority of the names as shown in Table 7.

Results also show that the type-aware translation system does not necessarily improve results, and using one universal system can deliver comparable results for most of the cases.

Person names transliteration unexpectedly achieved less quality than translation. Names are pronounced differently across countries. For example, a USA-based annotator is expecting "Friedrich" to be written "فريدريك", while a Germany-based one is expecting it to be written as "فريدريش".

Inter-annotator agreement was measured using Fleiss' kappa to be 0.484 indicating moderate agreement.

7 Conclusion

In this paper we introduced EDRAK: and entity-centric Arabic resource. EDRAK is an entity repository that contains around 2.4M entities, with their potential Arabic names. In addition, EDRAK associates each entity with a set of keyphrases. Data in EDRAK has been extracted from the Arabic Wikipedia and other available resources. In addition, we automatically translated parts of the English Wikipedia and used them to enrich EDRAK. Data have been manually assessed. Results showed that the quality is adequate for consumption by other NLP and IR systems. We are making the resource publicly available to help advance the research for the Arabic language.

Acknowledgments

We would like to thank the anonymous reviewers for their valuable feedback. We would also like to show our gratitude to the volunteers who participated in our manual assessment task. Finally, we are also immensely grateful to Waleed Ammar, Chris Dyer, and Hassan Sajjad for their valuable advice about translation and transliteration using cdec framework.

References

Mohamed Al-Badrashiny, Ramy Eskander, Nizar Habash, and Owen Rambow. 2014. Automatic Transliteration of Romanized Dialectal Arabic. In *Proceedings of the 18th Conference on Computational Natural Language Learning*, CoNLL 2014, Baltimore, Maryland, USA.

Yaser Al-Onaizan and Kevin Knight. 2002. Translating Named Entities Using Monolingual and Bilingual Resources. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL 2002, Stroudsburg, PA, USA.

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 7th International Conference on Language Resources and Evaluation*, LREC 2010, Valletta, Malta.

Sren Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, and Zachary Ives. 2007. DBpedia: A Nucleus for a Web of Open Data. In *Proceedings of the 6th International Semantic Web Conference*, ISWC 2007, Busan, Korea.

Mahmoud Azab, Houda Bouamor, Behrang Mohit, and Kemal Oflazer. 2013. Dudley North visits North London: Learning When to Transliterate to Arabic. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAAC-HLT 2013, Atlanta, Georgia.

William Black, Sabri Elkateb, and Piek Vossen. 2006. Introducing the Arabic Wordnet Project. In *In Proceedings of the 3rd International WordNet Conference*, GWC 2006, Jeju Island. Korea.

Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. 2008. Freebase: A Collaboratively created Graph Database for Structuring Human Knowledge. In *Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data*, SIGMOD 2008, New York, NY, USA.

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*, ACL 2013, Sofia, Bulgaria.

Chris Dyer, Adam Lopez, Juri Ganitkevitch, Johnathan Weese, Ferhan Ture, Phil Blunsom, Hendra Setiawan, Vladimir Eidelman, and Philip Resnik. 2010. cdec: A Decoder, Alignment, and Learning Framework for Finite-state and Context-free Translation Models. In *Proceedings of the Association for Computational Linguistics*, ACL 2010, Uppsala, Sweden.

Ondrej Hálek, Rudolf Rosa, Ales Tamchyna, and Ondrej Bojar. 2011. Named Entities from Wikipedia for Machine Translation. In *Proceedings of the Conference on Theory and Practice of Information Technologies*, ITAT 2010, Velká Fatra, Slovak Republic.

Johannes Hoffart, Mohamed Amir Yosef, Ilaria Bordino, Hagen Fürstenau, Manfred Pinkal, Marc Spaniol, Bilyana Taneva, Stefan Thater, and Gerhard

Weikum. 2011. Robust Disambiguation of Named Entities in Text. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP 2011, Edinburgh, UK.

Johannes Hoffart, Stephan Seufert, Dat Ba Nguyen, Martin Theobald, and Gerhard Weikum. 2012. KORE: Keyphrase Overlap Relatedness for Entity Disambiguation. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, CIKM 2012, Hawaii, USA.

Johannes Hoffart, Fabian M. Suchanek, Klaus Berberich, and Gerhard Weikum. 2013. YAGO2: A Spatially and Temporally Enhanced Knowledge Base from Wikipedia. *Journal of Artificial Intelligence*.

Farzaneh Mahdisoltani, Joanna Biega, and Fabian M Suchanek. 2015. Yago3: A Knowledge Base from Multilingual Wikipedias.

Will Monroe, Spence Green, and Christopher D. Manning. 2014. Word Segmentation of Informal Arabic with Domain Adaptation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, ACL 2014, Baltimore, MD, USA.

Roberto Navigli and Simone Paolo Ponzetto. 2012. BabelNet: The Automatic Construction, Evaluation and Application of a Wide-coverage Multilingual Semantic Network. *Journal of Artificial Intelligence*.

Arfath Pasha, Mohamed Al-Badrashiny, Mona Diab, Ahmed El Kholy, Ramy Eskander, Nizar Habash, Manoj Pooleery, Owen Rambow, and Ryan M Roth. 2014. MADAMIRA: A Fast, Comprehensive Tool for Morphological Analysis and Disambiguation of Arabic. *Proceedings of the Language Resources and Evaluation Conference*, LREC 2014, Reykjavik, Iceland.

Khaled Shaalan. 2014. A Survey of Arabic Named Entity Recognition and Classification. *Computational Linguistics*.

Nakatani Shuyo. 2010. Language Detection Library for Java.

Valentin I. Spitkovsky and Angel X. Chang. 2012. A Cross-lingual Dictionary for English Wikipedia Concepts. In *Proceedings of the 8th International Conference on Language Resources and Evaluation*, LREC 2012, Istanbul, Turkey.

Ralf Steinberger, Bruno Pouliquen, Mijail Kabadjov, Jenya Belyaeva, and Erik van der Goot. 2011. JRC-NAMES: A Freely Available, Highly Multilingual Named Entity Resource. In *Proceedings of the International Conference Recent Advances in Natural Language Processing*, RANLP 2011, Hissar, Bulgaria.

Tomasz Tylenda, Mauro Sozio, and Gerhard Weikum. 2011. Einstein: Physicist or Vegetarian? Summarizing Semantic Type Graphs for Knowledge Discovery. In *Proceedings of the 20th International Conference on World Wide Web*, WWW 2011, Hyderabad, India.

Gerhard Weikum, Johannes Hoffart, Ndapandula Nakashole, Marc Spaniol, Fabian M Suchanek, and Mohamed Amir Yosef. 2012. Big Data Methods for Computational Linguistics. *IEEE Data Engineering Bulletin*.

Mohamed Amir Yosef, Sandro Bauer, Johannes Hoffart Marc Spaniol, and Gerhard Weikum. 2012. HYENA: Hierarchical Type Classification for Entity Names. In *Proc. of the 24th International Conference on Computational Linguistics*, COLING 2012, Mumbai, India.

Mohamed Amir Yosef, Sandro Bauer, Johannes Hoffart Marc Spaniol, and Gerhard Weikum. 2013. HYENA-live: Fine-Grained Online Entity Type Classification from Natural-language Text. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, ACL 2013, Sofia, Bulgaria.

Mohamed Amir Yosef, Marc Spaniol, and Gerhard Weikum. 2014. AIDArabic: A Named-Entity Disambiguation Framework for Arabic Text. In *The EMNLP 2014 Workshop on Arabic Natural Language Processing*, ANLP 2014, Dohar, Qatar.