Adaptation and evaluation of systems for recognition and resolution of named entities in Arabic literacy

kawther A.Al-Dhlan - Associate professor – Department of Information and Computer Science - College of Computer Science and Engineering – University of Ha'il

Hail-KSA.

K_aldhlan@hotmail.com

Abstract

In thispaper, we propose chaining two existing tools, one for the recognition of named entities, and the other for the resolution of named entities in order to process texts from modern Arabic literature. The result provided by our processing chain projects the named entities from the Arabic text to its correct location on a map and finally showshow this idea can be interesting to the digital ecosystem.

Key Words

named-entity disambiguation, evaluation, linked data, digital humanities.

Introduction

Natural language processing(NLP) occupies an important place of the digital ecosystem spectrum and contributesparticularly to the analysis of digital literary works [Lecluze et al, 2014]. In this paper, we are particularly interested in named entities recognition (NER) and the resolution of named entities (RNE), we will use common NLP tasks in order to enrich texts from the Arabic literary. One part of the project consists of recognition of named entities which aims to identify and categorize linguistic expressions such as the names of person, place, and institution, on the other hand, the other part aims to determine the identity of the entities, mentioned in the text, from a databasesuch as DBpedia, Wikidata, Yago, Geonames.

In this work, we propose a processing chain based on two tools NER and RNE which will be adapted to the analysis texts from Arabic literature. We have set up a gold standard made up of two chapters from the novel " Cairo Trilogy: Palace Walk, Palace of Desire, Sugar Street " by Naguib Mahfouz and the first chapter of the novel “Al Nazarat” by MustafaLutfialManfaluti. We have also developed an online application in order to offer a dynamic rendering projecting the places marked in the texts on a map. The remainder of this paper is broken down into three parts. The first two parts are devoted to the presentation of the NER and RNE tools selected for our study, and the last part presents a conclusion evoking digital ecosystem in our projects.

1. About the processing chain

The proposed processing chain is illustrated In Figure 1, it integrates the " Segmenteur-Étiqueteur Markovien " system SEM [Dupont 2017] for the NERtask and "REDEN" [Frontini et al 2015; Brando et al 2016] for the RNEtask. The choice of the essential XML / TEL format was made for digital text, rendition was the choice of encoding made for the return of our chain. It is therefore essential that both tools support this format. For REDEN, the question does not arise. On the other hand, it was necessary to adapt SEM in order to give support to the XML / TEL format.

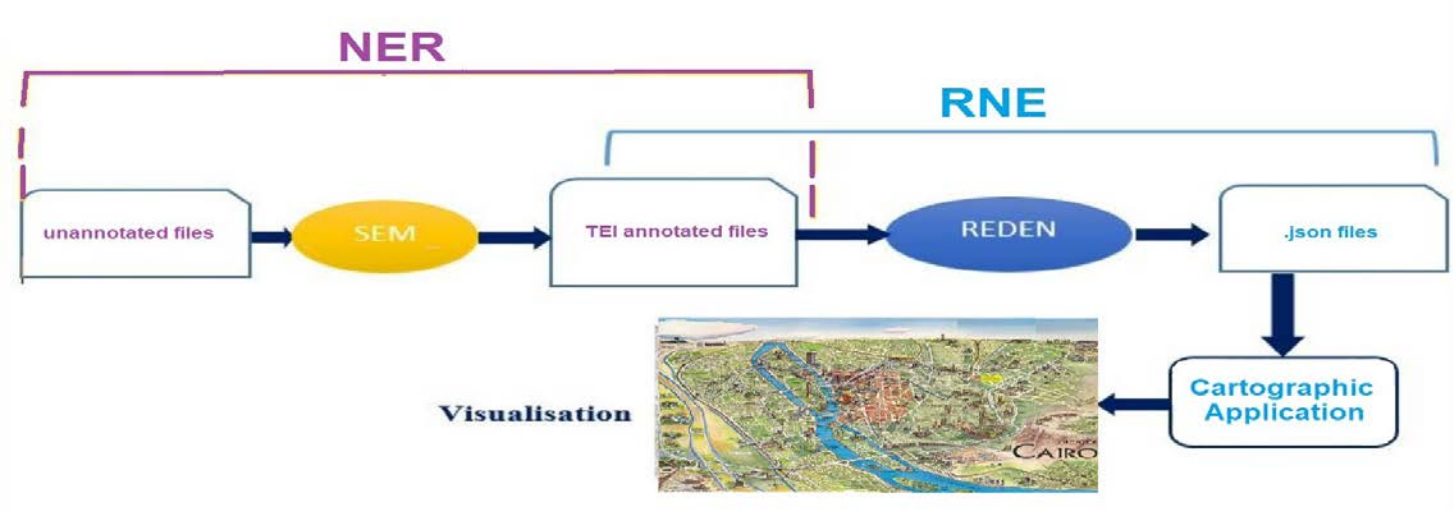

Figure1. Proposed processing chain

1.1. Adaptation and evaluation of the SEM system

SEM is a NER system that relies on a supervised learning approach [Raymond et al., 2010]. The automatics learning of the system is based on training a model from examples in order to reproduce a prediction sheet. A model for SEM is available, based on journalistic texts from the Arabic TreeBank [Mohamed Maamouri et al., 2005]. For this work, we have trained a NER model for SEM from Arabic literary texts that we have manually annotated (the gold standard) by two distinct annotators. The inter-annotator agreed to give the following values: 0.88 for recall, 0.96 for precision and 0.91 for F-measure. The proximity of the inter-annotator fitting values, all close to 1, deduces the similarity of the annotation produced by the two experts.

We have evaluated the performance of this model in terms of recall, precision and F-measure, as well as analyzed recurring annotation errors. The experiments on SEM concern two types of named entities, namely people and places, and are carried out on the gold standard. They are divided into three main parts, described below.

1.2Evaluation of SEM with the Arabic Tree Bank model:

This experiment shows that the model used on contemporary journalistic texts is not sufficiently portable on literary texts. In particular, Named Entities (NE) type “Person” pose more problems with variable F-measure results which can reach a minimum of 10%. This is because the names of people correspond to fictitious names. However, for the recognition of places, the results are better with F-measure values varying between 24% and 33%. This is because the names of places represent real places that exist and are therefore known and learned by the model and present in the dictionary. The following table 1 shows the results of the experiments.

	ArabicTreeBank SEM Model Palace of Desire		ArabicTreeBank SEM Model Sugar Street		Adaptation1	Adaptation2
	Location	Person	Location	Person		
Global Precision	0.43	0.175	0.23	0.12	0.95	0.5
Global Recall	0.24	0.06	0.25	0.09	0.68	0.27
Global F-Measure	0.31	0.11	0.21	0.09	0.81	0.32

Table 1. SEM experiment results

Three types of errors were noted: (1) NE type error, (2) partial annotation, (3) absence annotation. In particular, the non-recognition of site triggers was responsible for some of the Errors. For this reason, adaptation to the domain required the development of a dictionary of triggers (boulevard, street ...) to improve the retraining phase.

1.3. Adaptation Domain

Adaptation A: The first adaptation consists of training on an excerpt from the novel " Palace Walk” and an assessment of another part of the same novel. The results of this experiment show us that a model trained and tested on the same author, for our case Naguib Mahfouz, gives quite good results with a overall accuracy of 1, an overall recall that reaches 0.67 and an F-measure between 0.81 (see tab 1). These results can be explained by the fact that a chapter is a very narrow area of reference, in which same names of person and place tend to repeat themselves. Which means that training on part of a chapter produces good chapter’s annotation on the other.

Adaptation B: It's a workout on an excerpt from the novel " Palace Walk " and an evaluation onthe other novel by a " The Yacoubian Building".The same learning model from (Adaptation 1) and tested on “Sugar Street” novel, givesquite weak measures. An accuracy of 0.5, a recall of 0.27 and an F-measure of 0.32. These results are justifiedbecause the model is less portable from one author to another.

1.4. Progress:

In this experiment, the training corpus of Mahfouz extracts is gradually increased in order to find the optimal size of the learning sample. The learning sample is made up of 1/3, 2/3 and 3/3 of a chapter respectively and the models are tested on two different extracts from Mahfouz andAl Manfaluti.

Figure 2. Respective progressions for Mahfouz and Al Manfaluti

For the test on Mahfouz's text, by increasing the size of the training corpus the performance increases progressively. However, by conducting a training with a text from Mahfouz and testing it on the text of Al Manfaluti, the results improve first but decrease when the training corpus is composed of 3 extracts. This can be explained by the fact that the training corpus has become too suitable for Mahfouz which could lead to a problem of over-learning and lack of generalization [Douglas M., 2004].

2. Adaptation and evaluation of the REDEN system

REDEN [Frontini et al, 2015, Brando et al, 2016], is anRNE tool based on the graph theorywhich permits the resolutionof NE based on web-based data sources. This tool takes an XML/TEL file as an entrytagged into named entities and produces the same file enriched with an identifier, an IRI (acronym for Internationalized Resource Identifier), for each entity and a NIL for entities without reference. We are onlyinterested on places and have adapted three databases (DBpedia, BNF, WikiData) from queriesSPARQL. Each query is related to a single knowledge base and produces a dictionarywhich is used by REDEN to search for candidates. We finally evaluated REDEN from textsannotated by SEM for each database. The evaluation metrics depend on the RNE phase in question, i.e.(1) Search for candidates, (2) Selection of the right candidate. These metrics extend those of the NER (see [Brandoet al, 2016] for details). The results obtained during the experiments are presented in Table 2.

Database	Candidate Precision	Candidate recall	NIL Precision	NIL recall	Disambiguation accuracy	Ambiguity rate	Overall linking accuracy
DBpedia	1	0.816	0.367	1	None	0	0.834
BNF	0.760	0.630	0.580	0.972	1	0.005	0.7
Wikidata	0.912	0.830	0.440	1	1	0.29	0.85

Table 2. Results of REDEN experiments

Phase metrics (1) allow us to determine its effectiveness. The " Candidate Precision " is 1 for database DBpedia, 0.912 for Wikidata and 0.76 for BNF. So generally, REDEN is able to find the appropriate IRI in the database among a set of non-empty candidates. In addition, for the "Candidate recall" we notice that the results with DBpedia (0.816) and in Wikidata is (0.83) which outperforms BNF (0.63). This translates into that REDEN finds more correct references with DBpedia and Wikidata compared to BNF. In addition, since we are interested in toponyms ofthe 1900 eraof the Arabic literature, some places have changed their name.

To assess the ability of the algorithm to produce correct NIL annotations for mentions that do not have of refer in the gold. We observe that the result of the "NIL Precision" is slightly better with the BNF base 0.580 compared to Wikidata 0.44 and DBpedia 0.367. However, the results remain low, this is explained by the fact that REDEN processes the candidate search phase with the algorithm of perfect match between character strings (exact string match). On the other hand, the "NIL Recall" is high for the three databases, 1 for DBpedia and Wikidata, 0.972 for BNF. This means thatcomparing results to the NIL in gold, causes REDEN to return correct NILs. As for the measurements of phase (2), the result obtained for the “Disambiguation accuracy” for DBpedia is empty, compared to the BNF and Wikidata 1. It is important to note that this measurement is interesting when we have sets of candidates of size greater than 1. The ambiguity rate is zero for DBpedia, 0.05 for the BNF and 0.29 for Wikidata. Which means that the average of the sets of candidates with more than 2 candidates is low for DBpedia as well as BNF and slightly more important with Wikidata. Finally, the “Overall Linking” measurement obtained is 0.834 for DBpedia, 0.85 for Wikidata and 0.7 for BNF. REDEN was therefore effective. This metric tries to assess the overall efficiency of the system, not by phase, and expresses the reliability of REDEN for the named entity resolution task. We can thus conclude that all three bases are effective for the RNE task, and give correct results. However, Wikidata is slightly better than DBpedia and BNF.

3. Conclusion

my work will prove useful for the creation of a literary section of a Cairo Time Machine. This ambitious project aims to reconstruct the past of major Egyptian cities through information extraction from historical documents, including literary ones. Our contribution is particularly to have produced a model for the recognition of NE in Arabic literary texts as well as Web databases of relevant data for the RNE task in this same context. In order to illustrate our remarks, figures 3 and 4, propose two possible cartographic renderings from the proposed processing chain. Since the databases used list the location coordinates for each location type entity, so it is possible, after the RNE phase, to repatriate this information (and others) automatically through IRIs. The first view shows the places that were mentioned in the novels and marks them with an indicator on the map. The second view allows us to visualize on a map for each place mentioned in the text the number of occurrences of this NE in the novels.

Figure 3. #1 cartographic view

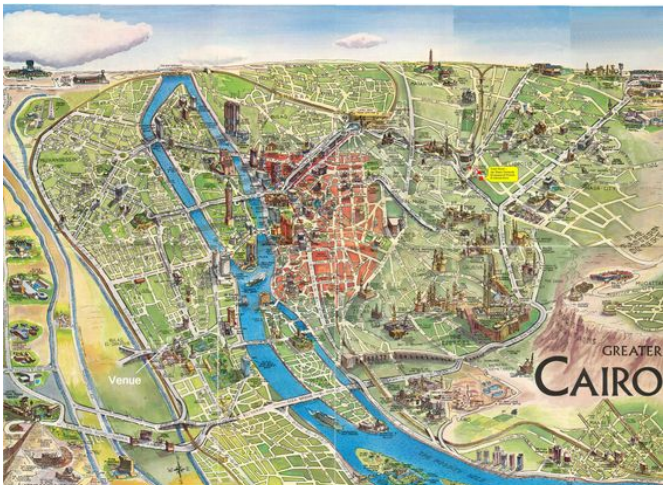

Figure 4. #2 cartographic view

Bigraphy of Auther :

kawther A.Al-Dhlan , Associate professor in Computer science , she has 10 years experince in data science and more than 25 years in academic fields. moreover she interested in many branches of Cs such as security, cloud computing and deep learning , currently she teaches in University of Hail .

References

[1] Brando, C., Frontini, F., Ganascia, J.G.: Disambiguation of named entities in cultural heritage texts using linked data sets (accepted). In: Proceedings of the First International Workshop on Semantic Web for Cultural Heritage in Conjunction with 19th East-European Conference on Advances in Databases and Information Systems (2015)

[2] Cucerzan, S.: Large-scale named entity disambiguation based on Wikipedia data. In: Proceedings of the 2007 Joint Conference on mpirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL), pp. 708–716. Association for Computational Linguistics, Prague, Czech Republic, Jun 2007. https://www.aclweb.org/anthology/D07-1074

[3] Daiber, J., Jakob, M., Hokamp, C., Mendes, P.N.: Improving efficiency and accuracy in multilingual entity extraction. In: Proceedings of the 9th International Conference on Semantic Systems (I-Semantics) (2013)

[4] Dredze, M., McNamee, P., Rao, D., Gerber, A., Finin, T.: Entity disambiguation for knowledge base population. In: Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010), pp. 277–285. Coling 2010 Organizing Committee, Beijing, China, August 2010. https://www.aclweb.org/anthology/C10-1032

[5] Ehrmann, M., Romanello, M., Bircher, S., Clematide, S.: Introducing the CLEF 2020 HIPE shared task: named entity recognition and linking on historical newspapers. In: Jose, J.M., et al. (eds.) ECIR 2020, Part II. LNCS, vol. 12036, pp. 524–532. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-45442-5_68

[6] Ehrmann, R., Clematide, F.: HIPE - Shared Task Participation Guidelines, January 2020. https://doi.org/10.5281/zenodo.3677171

[7] Freeman, L.C.: A set of measures of centrality based on betweenness. Sociometry pp. 35–41 (1977)

[8] Frontini, F., Brando, C., Ganascia, J.G.: Semantic web based named entity linking for digital humanities and heritage texts. In: Proceedings of the First International Workshop Semantic Web for Scientific Heritage at the 12th ESWC 2015 Conference. pp. 77–88 (2015), http://ceur-ws.org/Vol-1364/

[9] Ganea, O.E., Hofmann, T.: Deep joint entity disambiguation with local neural attention. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. pp. 2619–2629. Association for Computational Linguistics (2017). https://doi.org/10.18653/v1/D17-1277

[10] Hachey, B., Radford, W., Curran, J.R.: Graph-based named entity linking with wikipedia. In: Web Information System Engineering–WISE 2011, pp. 213–226. Springer (2011)

[11] Heino, E., et al.: Named entity linking in a complex domain: case second world war history. In: Gracia, J., Bond, F., McCrae, J.P., Buitelaar, P., Chiarcos, C., Hellmann, S. (eds.) LDK 2017. LNCS (LNAI), vol. 10318, pp. 120–133. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-59888-8_10

[12] Hoffart, J., et al.: Robust disambiguation of named entities in text. In: Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, pp. 782–792. Association for Computational Linguistics, Edinburgh, Scotland, UK, July 2011. https://www.aclweb.org/anthology/D11-1072

[13] Kolitsas, N., Ganea, O.E., Hofmann, T.: End-to-end neural entity linking. In: Proceedings of the 22nd Conference on Computational Natural Language Learning, pp. 519–529. Association for Computational Linguistics (2018). https://doi.org/10.18653/v1/K18-1050

[14] Lehmann, J., et al.: DBpedia - a large-scale, multilingual knowledge base extracted from Wikipedia. Semant. Web J. 6(2), 167–195 (2015). https://doi.org/10.3233/SW-140134

[15] Rao, D., McNamee, P., Dredze, M.: Entity linking: Finding extracted entities in a knowledge base. In: Multi-source, Multilingual Information Extraction and Summarization, pp. 93–115. Springer (2013)

[16] Rochat, Y.: Character Networks and Centrality. Ph.D. thesis, University of Lausanne (2014)

[17] Sinha, R.S., Mihalcea, R.: Unsupervised graph-basedword sense disambiguation using measures of word semantic similarity. In: ICSC. vol. 7, pp. 363–369 (2007) 9. Usbeck, R., Ngonga Ngomo, A.C., Auer, S., Gerber, D., Both, A.: Agdistis - graphbased disambiguation of named entities using linked data. In: 13th International Semantic Web Conference (2014)

[18] van Hooland, S., De Wilde, M., Verborgh, R., Steiner, T., Van de Walle, R.: Exploring entity recognition and disambiguation for cultural heritage collections. Digit. Sch. Humanit. 30(2), 262–279 (2013). https://doi.org/10.1093/llc/fqt067

[19] Wilde, M.: Improving retrieval of historical content with entity linking. In: Morzy, T., Valduriez, P., Bellatreche, L. (eds.) ADBIS 2015. CCIS, vol. 539, pp. 498–504. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-23201-0_50