Evaluating Lightweight text classification and Information Extraction for Arabic texts

Dhaou Ghoul, Lichao Zhu & Gael Lejeune

STIH (Sens Texte Informatique Histoire), Sorbonne Université

dhaou.ghoul@sorbonne-universite.fr, lichao.zhu@gmail.com, gael.lejeune@sorbonne-universite.fr

Abstract. Epidemic Surveillance aims at detecting disease outbursts in the world in order to provide useful information to health authorities. Many automatic systems have been conceived to help these authorities to mine the available data with a special focus on press articles. The main goal of these systems is to select relevant articles by means of text classification. The secondary goal is to extract valuable information from these relevant texts . One of the main challenge is to handle many languages with different properties, various availability of language resources (lexicons, POS taggers…) and annotated data. In this paper we present a state of the art on Text Classification as well as Information Extraction for the Arabic language and we test different options for designing a lightweight system to process texts in written Arabic. We show that Arabic language has particular properties, making it difficult to handle properly without improving existing approaches. We propose improvements of an existing lightweight approach that would be promising for Arabic as well as more poorly endowed languages.

Introduction

Available online press articles have become the main sources of information. The amount of published articles is increasing, making it difficult for users to get a comprehensive view of the data. This is why it is fundamental to have effective solutions in classifying the information to help web users find relevant documents in different areas. Among these areas is the epidemic surveillance domain. That's why we need expertise of web-based epidemic intelligence systems that allow us to easily and automatically detect epidemic disease outbursts [1].

Epidemic surveillance consists of detection and interpretation of unstructured, available information on the Internet. Epidemic surveillance is intended to provide duly selected and indexed documents [2]. Text classification and information extraction are powerful techniques that help to structure data in order to help experts of many fields including epidemiology. The main goal of classification is to select relevant documents for a particular task whereas Information

Extraction consists in extracting a structured representation from these documents in order to populate databases. Several approaches have been proposed for this task. Most of existing approaches are primarily designed for processing texts written in English, relying on sentence patterns [3], ontology-like lexical resources [4] or rather hybrid approaches [5]. Extending the multilingual coverage therefore implies to reproduce a pipeline with language dependent resources and processing tools. This approach is not suitable for all languages [6] even with the help of machine learning [7]. [8] pointed out the need to process poorly endowed languages or dialects without training data. These approaches did not seem suitable for building a disease surveillance system for a language like Arabic. In this paper we will test Daniel (Data Analysis for Information Extraction in any Language)[9], a lightweight approach which, according to the authors, allows to handle a large number of languages with a limited quantity of lexicon The system has been evaluated on 17 languages (English, French, Greek…) but has not, to the best of our knowledge, been appropriately evaluated on the Arabic language.

The article is organized as follows: in Section 2, we present some works on Information Extraction and Text classification for the Arabic language; in Section 3, we describe a lightweight approach for text classification and information extraction ; in Section 4 we analyze results and present some propositions to improve this approach.

Related Work

Text Classification

Text classification consists of assigning unknown documents into predefined classes. The process of text classification is usually summarized in the following steps [10]:

1. document pre-processing, i.e. tokenisation, stop-word removal, and stemming or lemmatisation
2. document modelling, i.e. representing a document in an appropriate form so that it can be processed by a machine learning algorithm
3. feature selection and projection
4. transforming features into classification rules
5. quality indicators and evaluation methods

Regarding step 4, there are three approaches to text classification: rule-based approach, machine learning approach and hybrid approach [11].

Rule-based approach organizes text into classes according to a set of handcrafted linguistic rules. If one wants to classify newspaper articles into two classes (economy and health), the easiest way would be to define two lists of tokens (usually words) that are most discriminant for each class. Then, when a new text needs to be classified, its class should be identified by comparing the tokens of the text to the list of tokens most representative of each class. Here, the rules are made to select relevant tokens and to assign appropriate weights to them. For instance, in emotion detection, lexicons can be used to compute a probabilty that the text belongs to a particular class [12]. These approaches can be quite simple to implement but, on the other hand, rule-based approaches have some disadvantages. These approaches require some knowledge of the domain and creating efficient language rules is expensive.

Unlike rule-based approaches, machine learning approaches take advantage of past observations, rather than explicit expert knowledge, using pre-labeled examples as training data. For this purpose, the amount of training data should be sufficiently high. The first step is to transform the text into a an appropriate representaion (usually vectors). One of the most frequently used approaches is bag of words, where a vector represents the frequency of a word in a predefined dictionary of words. Then, a learning algorithm (Naive Bayes, Support Vector Machine, KNN-neighbors, Deep Learning,...) is applied which takes a labeled learning corpus as input to create a classification model. With the appropriate amount of training data, text classification via machine learning exhibits more accurate results than rule-based approaches. The strength of these two approaches can be combined by a process called hybridization [13]. Expert knowledge and machine learning methods are used to find a symbiosis between the simplicity of the linguistic rules and the efficiency of machine learning.

Although much work has been devoted to the classification of available texts in English, Chinese and other common languages (Spanish, French ...), few works have studied the classification of texts in Arabic. We will present here some of the most interesting works on Arabic texts classification. These studies use different datasets with different algorithms but we unified the metrics to evaluate the performance of each approach presented in Table 1.A comparative description inspired by [14] and [15] is given in Table 1. We can see that most of existing works rely on machine learning approaches.

Reference	Corpus (# docs)	Classes	Algorithm	Accuracy	Year
Syiam et al. [16]	News (1132)	6	Rocchio	98%	2006
Duwairi [17]	Magazine/News (1000)	10	Naïve Bayes	95%	2007
Mesleh et al. [18]	News (1545)	9	SVM	98%	2008
Bawaneh et al. [13]	Unknown (242)	6	KNN	84%	2008
Ababneh et al .[19]	News (5121)	7	Cosine	95%	2014
Amina et al. [20]	News (6005)	9	SVM/Naïve Bayes	80%/ 70%	2017
Zinah et al .[21]	News (16757)	5	Master-slaves	88%	2018

Table 1: Some examples of studies performing Arabic texts classification

As mentioned earlier, the text classification process comprises three main steps: pre-processing, classification and evaluation. As Arabic is a morphologically rich language, the pre-processing phase is crucial but we lack efficient pre-processing tools(contrary to an isolated language like English).

For Arabic texts, the pre-processing, besides tokenization and lemmatization, involves normalization of some Arabic letters, for instance:

– we change ("ءْ" hamza),("اْ" aleph mad, ("اَ" aleph to hamza on top and ("وْ" hamza on w) to ("ا" alef)
– we change ("اِ" alef with hamza on the bottom) and ("يْ" hamza on ya) to ("ا" alef).

For Arabic, as well as for many other languages, there is an important gap compared to English regarding the quality of natural language processing applications. Regarding this, there is a stronger interest in the scientific community towards a better treatment of both morphologically rich languages [22] and poorly endowed languages [23]. [24] proposed Daniel, a lightweight approach for Classification and Information Extraction that shows convincing results for a bunch of morphologically rich languages (Greek, Polish and Russian) and remains competitive for rather isolate languages like Chinese or English. The rationale of their approach is to avoid classical pre-processing steps, leaving out the problems of tokenization and lemmatization, and to take advantage of text type properties. Their approach seems to be limited to news and has also been applied to Arabic and a set of languages like German, Spanish or Vietnamese but without proper evaluation

In the next section we will show our reproduction experiments of the Daniel approach and to properly evaluate this approach for Arabic texts.

Information Extraction

Information Extraction goes back up to the early days of automatic natural language processing in the 1970s. In general, Information Extraction (IE) aims to acquire knowledge from a text.Two basic information extraction tasks are named entity recognition and relationship extraction. The extraction task is carried out thanks to the filling of predefined forms. This model describes a set of entities, the relationships between them and the events involving these entities [25]. For example, a template(form) for a disease should specify fields such as: "name of disease", "place of disease", "number of victims". Information extraction methods can be classified into three categories: linguistic methods, statistical methods (Machine learning approaches) and hybrid methods. The linguistic methods are based on a syntactic study of text. On the other hand, statistical methods make it possible to extract information without prioritizing linguistic analysis. These methods are the most used in the processing of natural language. The hybrid methods consist of combining linguistic and statistical methods [26].

"*Arabic language is used by more than 330 million Arabic speakers who are spread over 22 countries.However, the IE in this language poses many problems because of the morphological and graphic changesin this language: polysemy, irregular and inflected derived forms, various spelling of certain words, various writing of certain combination character, short(diacritics) and long vowels, most* *of the Arabic words contain affixes*"[27]. The Named Entity Recognition(NER) is a sub-problem of Information Extraction (IE). In the Arabic language, several systems have been created on the recognition of named entities[28].

In general, the development of an information extraction system goes through three steps: (i) identify text fragments containing information, (ii) define the structure of information representation and (iii) develop the rules to identify the information and complete the proposed form. As mentioned previously, named entity recognition is a sub-task of information extraction. Named entities are textual elements allowing particularly relevant access to document content, that is why identifying and categorizing them is a key issue for the automatic understanding of texts. The following is a non-exhaustive list of NER tools that have been used in the Arabic NER literature.

System	Entity	Precision	Recall	F-measure	Method	Year
TAGARAB	Number	82.8	97.0	97.3	Rule based	1998
	Time	91.0	80.7	85.5		
	Location	94.5	85.3	89.7		
	Person	86.2	76.2	80.9		
Mesfar	Number	97.0	94.0	95.5	Rule based	2007
	Time	97.0	95.0	96.0		
	Location	82.0	71.0	76.0		
	Person	92.0	79.0	85.0		
PNAES	Person	93.0	86.0	89.0	Rule based	2009
ANERsys	Location	93.03	86.67	89.74	machine learning	2008
	Person	80.41	67.42	73.35		
	Misc	71.0	54.0	61.47		
	Organisation	84.23	53.94	65.76		
Abdul-hamid and Darwish	Location	93.0	83.0	88.0	machine learning	2010
	Person	90.0	75.0	81.0		
	Organisation	84.0	64.0	73.0		
Oudah and Shaalan	Location	x	x	90.0	Hybrid approach	2012
	Person	x	x	94.0		
	Organisation	x	x	88.0		

Table 2: Precision, Recall and F-Measure of Arabic Named Entity Recognition Systems.

To conclude, systems based on hybrid approaches have shown good performance in the recognition of Arabic named entities. It should be noted that the list of references provided here may not be complete. The Daniel system used in the next section does not use sentence level patterns but relies on document-level repetition (motifs) to perform NER and IE. The system extracts disease-location pairs for each relevant document.

Experiment the lightweight Daniel approach

For the experiments presented here, we used the code provided by the authors of Daniel [1]. In this section we will present the dataset we built for this experiment and the results obtained on Arabic texts. Then, we will discuss what we learned from these experiments and in the last part of this Section we will propose some improvements to this approach and confront it to datasets in other languages.

Getting Dataset and Resources

Since the Daniel code is designed to work with structured press articles, we managed to build a corpus of press articles in Arabic (which is presented in Table 3).

The method relies on repeated character strings to perform both classification and Information Extraction. For epidemic surveillance, the authors assume that a simple list of disease names obtained from Wikipedia pages is sufficient. The rationale is that in press articles, journalists use most common words in order to ensure that the information is conveyed properly. If scientific names are used, it is only additional to these common words since the target of the press articles is mainly composed by regular speakers rather than specialists. With this resource, the system provides a binary classification stating if a given press article is related to epidemics or not. An article is tagged as relevant if a substring S of a disease name D is found in salient positions (title, first paragraphand last paragraph) and if $\frac{length(S)}{length(D)} >= \theta$. θ is a threshold that can be manually tuned for each language but the authors report that $\theta = 0.8$ provides good results in various languages.

The definition of a relevant document would need to be well defined. It is not completely clear on how the frontier between a relevant and an irrelevant article is drawn since the guidelines used for human annotators[2] only indicate that in relevant articles ”the main theme of the article is epidemics”.

Once the relevant document is detected, the system tries to locate the event at country level. There again, common names from Wikipedia are used. If the name of a country is repeated, this country is supposed to be where the epidemics take place. Else, the location is the country where the article has been published. This rule is referred as ”implicit location”.

Therefore, when we built our dataset we created a resource indicating for each particular press source, the country where it is published (see Table 3).

Testing the Daniel approach

Creating a reference dataset for epidemic surveillance was not our first goal since it is rather costly to build a dataset of sufficient size so we only propose to evaluate the output of the system. This configuration does not allow us to evaluate

[1] https://github.com/rundimeco/daniel

[2] https://daniel.greyc.fr/guidelines.pdf

	ar Corpus	fr Corpus	multi Corpus
Documents	41,432	2,733	2,129
Paragraphs	$488 * 10^3$	$12 * 10^3$	$25 * 10^3$
Avg. paragraphs	$11.8(\pm23)$	$4.5(\pm2.3)$	$12(\pm8)$
Characters	$97 * 10^6$	$11 * 10^6$	$5 * 10^6$
Avg.characters	$2,347(\pm3010)$	$4,168(\pm4,604)$	$2,402(\pm23,99)$

Table 3: Statistics for the dataset in Arabic and the french and multilingual datasets

recall but at least we can evaluate precision. In our opinion, it is possible to verify the soundness of the approach for our purpose. Furthermore, previous research on this approach showed that the system usually produces worse results in precision than in recall. It is a somewhat counter-intuitive statement considering the lightness of the lexical resources.

Together with the Daniel code, we have been provided a reference dataset of more than 2,000 annotated articles in French. We will also perform experiments with the reference dataset ”corpus_daniel”[3] used in [24] which contains around 2,100 annotated documents in five languages. These two datasets have been annotated with the same guidelines. These datasets will be exploited to test our improvement proposals for the Daniel approach.

Results and discussion

In this section, we present the first results obtained on the Arabic corpus. Among the relevant documents we annotated 50 documents in order to assess precision. 54% of them were devoted to epidemic events. This was not a good result and was far from what was reported in the literature for other languages. Therefore, we wanted to find out more about this rather disappointing and to get insight about the classification errors.

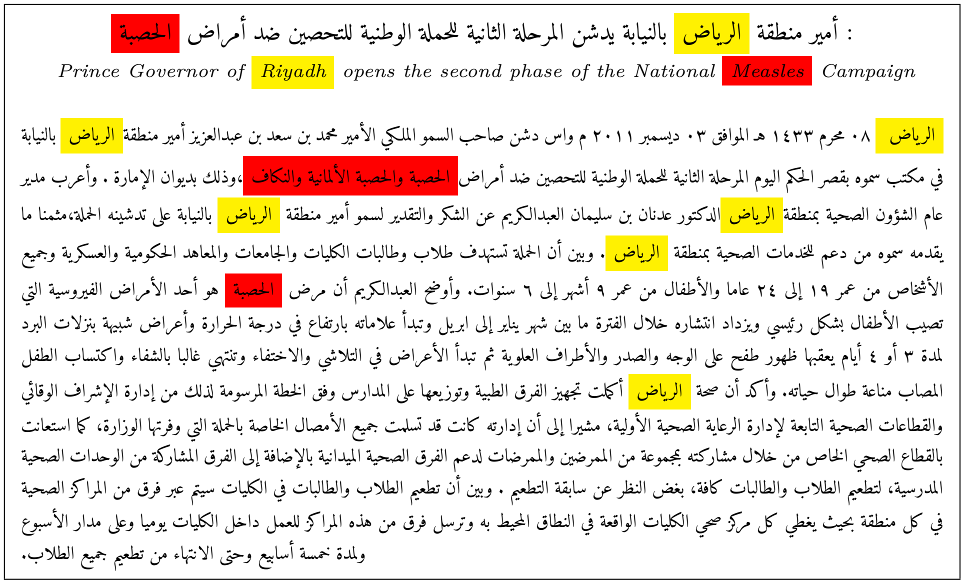
We present here two examples of documents deemed relevant by Daniel. In these two illustrations, the red color represents the disease and the yellow color the place of the event The article presented in Figure 1 describes cases of measles in Riyadh (Saudi Arabia) and is, according to the guidelines mentioned earlier, a true positive. On the contrary, Figure 2 shows a False Positive: because this article speaks about a city which is called ”measles”.

We can see with the True Positive example that the repetition patterns, the so-called ”relevant content” algorithm, succeeds in detecting the main subject of the article. this is not surprising as the 5W rule used in this article is also known in Arabic rhetorics[4].

In the other example, the mis-classification is not due to the algorithm itself but rather to the lexical resources it exploits. The character string extracted is a rather small one and therefore tends to be more frequent and is prone to

[3] Found through Research Gate : https://tinyurl.com/ResearchGate-DanielCorpus

[4] https://en.wikipedia.org/wiki/Five_Ws



: أمير منطقة الرياض بالنيابة يدشن المرحلة الثانية للحملة الوطنية للتحصين ضد أمراض الحصبة

Prince Governor of Riyadh opens the second phase of the National Measles Campaign

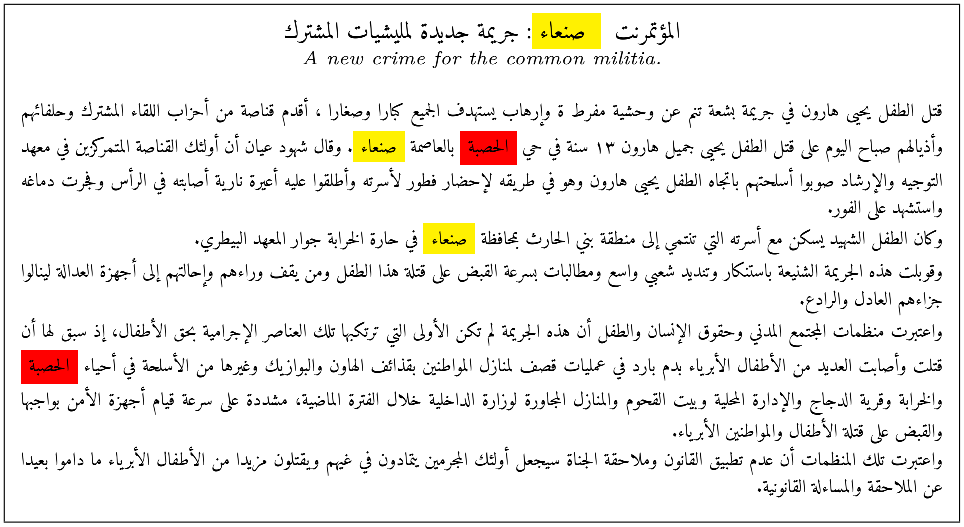
الرياض ٠٨ محرم ١٤٣٣ هـ الموافق ٠٣ ديسمبر ٢٠١١ م واس دشن صاحب السمو الملكي الأمير محمد بن سعد بن عبدالعزيز أمير منطقة الرياض بالنيابة في مكتب سموه بقصر الحكم اليوم المرحلة الثانية للحملة الوطنية للتحصين ضد أمراض الحصبة والحصبة الألمانية والنكاف ،وذلك بديوان الإمارة . وأعرب مدير عام الشؤون الصحية بمنطقة الرياض الدكتور عدنان بن سليمان العبدالكريم عن الشكر والتقدير لسمو أمير منطقة الرياض بالنيابة على تدشينه الحملة،مثمنا ما يقدمه سموه من دعم للخدمات الصحية بمنطقة الرياض . وبين أن الحملة تستهدف طلاب وطالبات الكليات والجامعات والمعاهد الحكومية والعسكرية وجميع الأشخاص من عمر ١٩ إلى ٢٤ عاما والأطفال من عمر ٩ أشهر إلى ٦ سنوات. وأوضح العبدالكريم أن مرض الحصبة هو أحد الأمراض الفيروسية التي تصيب الأطفال بشكل رئيسي ويزداد انتشاره خلال الفترة ما بين شهر يناير إلى ابريل وتبدأ علاماته بارتفاع في درجة الحرارة وأعراض شبيهة بنزلات البرد لمدة ٣ أو ٤ أيام يعقبها ظهور طفح على الوجه والصدر والأطراف العلوية ثم تبدأ الأعراض في التلاشي والاختفاء وتنتهي غالبا بالشفاء واكتساب الطفل المصاب مناعة طوال حياته. وأكد أن صحة الرياض أكملت تجهيز الفرق الطبية وتوزيعها على المدارس وفق الخطة المرسومة لذلك من إدارة الإشراف الوقائي والقطاعات الصحية التابعة لإدارة الرعاية الصحية الأولية، مشيرا إلى أن إدارته كانت قد تسلمت جميع الأمصال الخاصة بالحملة التي وفرتها الوزارة، كما استعانت بالقطاع الصحي الخاص من خلال مشاركته بمجموعة من الممرضين والممرضات لدعم الفرق الصحية الميدانية بالإضافة إلى الفرق المشاركة من الوحدات الصحية المدرسية، لتطعيم الطلاب والطالبات كافة، بغض النظر عن سابقة التطعيم . وبين أن تطعيم الطلاب والطالبات في الكليات سيتم عبر فرق من المراكز الصحية في كل منطقة بحيث يغطي كل مركز صحي الكليات الواقعة في النطاق المحيط به وترسل فرق من هذه المراكز للعمل داخل الكليات يوميا وعلى مدار الأسبوع ولمدة خمسة أسابيع وحتى الانتهاء من تطعيم جميع الطلاب.

Fig. 1: An example of True Positive : Measles in Saudi Arabia

ambiguity. In [24] the only parameter used to avoid such False Positive cases is the θ ratio between the found substring and the lexical entry of the database. In the code published online, some corrections are made to take into account different positions in the document. But neither of these two methods is suitable to resolve the problem identified here. Tuning the ratio would not help here since the full string is found in the document. Modifying the relevant positions would not help either since repetitions in first paragraph and body of the article are usually quite efficient to assess the theme of the document. Another solution might be to remove this disease name from the database but this would surely lead to an increasing number of False Negatives. The French corpus and the multilingual corpus provided by the authors show similar False Positives cases. In the multilingual corpus we had a False Positive case regarding ”Odra” (”Measles” in Polish) because Odra is also the name of a river and the name of a small city. Another example involved the French name for ”scabies” which is ”gale”. The substring ”gale” is not uncommon in French so that some of the False Positives identified in the data were due to that particular disease names.

An approach may be to try linguistic pre-processing and Named Entity recognition but, it would imply a paradigm shift. So we want to find a solution that keeps the originality, and the multilinguality, of the original approach.

We believe that the length of the disease name is the key. The longer the disease name is, the less ambiguous it is and the more confident the system should be. Setting a minimum length threshold would not fulfill the purpose. Some disease names are short and the length would need to be tuned according to the language. The solution we propose here is to take into account not only



المؤتمرنت صنعاء : جريمة جديدة لمليشيات المشترك

A new crime for the common militia.

قتل الطفل يحيى هارون في جريمة بشعة تنم عن وحشية مفرط ة وإرهاب يستهدف الجميع كبارا وصغارا ، أقدم قناصة من أحزاب اللقاء المشترك وحلفائهم
وأذيالهم صباح اليوم على قتل الطفل يحيى جميل هارون ١٣ سنة في حي الحصبة بالعاصمة صنعاء . وقال شهود عيان أن أولئك القناصة المتمركزين في معهد
التوجيه والإرشاد صوبوا أسلحتهم باتجاه الطفل يحيى هارون وهو في طريقه لإحضار فطور لأسرته وأطلقوا عليه أعيرة نارية أصابته في الرأس وفجرت دماغه
واستشهد على الفور.

وكان الطفل الشهيد يسكن مع أسرته التي تنتمي إلى منطقة بني الحارث بمحافظة صنعاء في حارة الخرابة جوار المعهد البيطري.

وقوبلت هذه الجريمة الشنيعة باستنكار وتنديد شعبي واسع ومطالبات بسرعة القبض على قتلة هذا الطفل ومن يقف وراءهم وإحالتهم إلى أجهزة العدالة لينالوا
جزاءهم العادل والرادع.

واعتبرت منظمات المجتمع المدني وحقوق الإنسان والطفل أن هذه الجريمة لم تكن الأولى التي ترتكبها تلك العناصر الإجرامية بحق الأطفال، إذ سبق لها أن
قتلت وأصابت العديد من الأطفال الأبرياء بدم بارد في عمليات قصف لمنازل المواطنين بقذائف الهاون والبوازيك وغيرها من الأسلحة في أحياء الحصبة
والخرابة وقرية الدجاج والإدارة المحلية وبيت القحوم والمنازل المجاورة لوزارة الداخلية خلال الفترة الماضية، مشددة على سرعة قيام أجهزة الأمن بواجبها
والقبض على قتلة الأطفال والمواطنين الأبرياء.

واعتبرت تلك المنظمات أن عدم تطبيق القانون وملاحقة الجناة سيجعل أولئك المجرمين يتمادون في غيهم ويقتلون مزيدا من الأطفال الأبرياء ما داموا بعيدا
عن الملاحقة والمساءلة القانونية.

Fig. 2: A False Positive example : there is a confusion because a city is named ”Measles”

the θ ratio but also the length of the disease name itself to compute a confidence score.

Taking into account the length of the disease names

In this configuration we sort the documents selected by the system with respect to the length of the disease names. We take advantage of the two annotated corpora at our disposal. We observe that False Positives come mostly from short disease names. We try different configurations, first with the longest disease names ($length >= 10$) and then with names of length 9, 8 ...and so on till all the disease names are introduced. The rationale of this experiment is to see from a ROC curve how recall and precision evolve.

Figure 3 shows the results obtained on the french corpus and multilingual corpus. One can see a property of the length of the disease names. Longer disease names leads to a better precision with a recall quite low. Shorter disease names are introduced step by step, increasing the recall but at some point with an important cost in precision. In order to maintain precision some improvements of the algorithm need to be performed.

Discussion

We advocate that there are two ways to do this. The first one would be to use a measure to assess the potential ambiguity of the substrings found in the document. This can be done using additional lexical resources, which would impair the ability to scale multingual corpus. Or, using measures like the adaptation

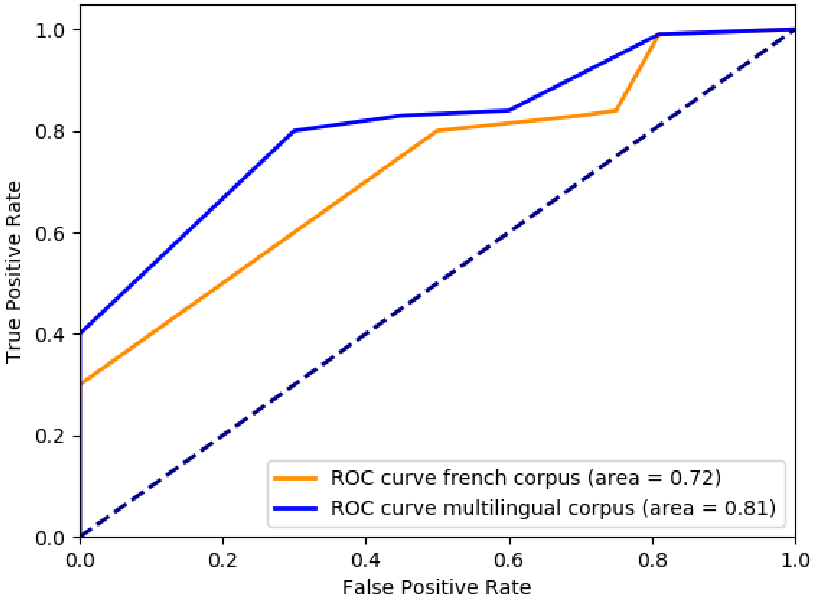


Fig. 3: ROC curve of the Daniel system with different compositions of the lexical resource

measure [29] would surely help to improve results without setting language dependent length thresholds.

The other way, and more promising, would be to use long disease names to bootstrap the system by getting for each language a bunch of annotated documents with great confidence scores. These annotated documents will then be used to learn words of the domain that are not disease names so that it would be possible to resolve ambiguities for shorter words independently of expert data. In the future, we plan to build an annotated dataset of sufficient size for Arabic to experiment with this solution, because it would help to assess how the lightweight approach for text classification and Information Extraction can be useful for Arabic texts.

References

[1] Barboza, P., Vaillant, L., Mawudeku, A., Nelson, N.P., Hartley, D.M., Madoff, L.C., Linge, J.P., Collier, N., Brownstein, J.S., Yangarber, R., Astagneau, P., on behalf of the Early Alerting, R.P.o.t.G.H.S.I.: Evaluation of epidemic intelligence systems integrated in the early alerting and reporting project for the detection of a/h5n1 influenza events. PLOS ONE **8** (2013) 1–9

[2] Lejeune, G., Lucas, N., Doucet, A.: Tentative d'approche multilingue en extraction d'information. In: JADT Journées internationales d'Analyse statistique des Données Textuelles, rome, Italy (2010) 1259–1267

[3] Du, M., Von Etter, P., Kopotev, M., Novikov, M., Tarbeeva, N., Yangarber, R.: Building support tools for Russian-language information extraction. In Habernal, I. and Matoušek, V., ed.: Proceedings of the 14th international conference on Text, Speech and Dialogue, Pilsen, Czech Republic, Springer (2011) 380–387

[4] Collier, N.: Towards cross-lingual alerting for bursty epidemic events. Journal of Biomedical Semantics **2** (2011) 1–11

[5] Freifeld, C.C., Mandl, K.D., Reis, B.Y., Brownstein, J.S.: HealthMap: Global Infectious Disease Monitoring through Automated Classification and Visualization of Internet Media Reports. Journal of the American Medical Informatics Association **15** (2008) 150–157

[6] Steinberger, R.: A survey of methods to ease the development of highly multilingual text mining applications. Language Resources and Evaluation (2011) 1–22

[7] Etzioni, O., Fader, A., Christensen, J., Soderland, S.: Open Information Extraction: The Second Generation. Proceedings of the 22nd International Joint Conference on Artificial Intelligence (2011) 3–10

[8] Munro, R.: Processing short message communications in low-resource languages. PhD thesis, Stanford University (2012)

[9] Lejeune, G., Brixtel, R., Lecluze, C., Doucet, A., Lucas, N.: Daniel : Veille épidémiologique multilingue parcimonieuse. In: Proceedings of TALN 2013. (2013) 787–788

[10] Mirończuk, M.M., Protasiewicz, J.: A recent overview of the state-of-the-art elements of text classification. Expert Systems with Applications **106** (2018) 36 – 54

[11] Mesleh, A.M.: Support vector machines based arabic language text classification system: Feature selection comparative study. In: Advances in Computer and Information Sciences and Engineering, Proceedings of the 2007 International Conference on Systems, Computing Sciences and Software Engineering (SCSS), part of the International Joint Conferences on Computer, Information, and Systems Sciences, and Engineering (CISSE 2007), Bridgeport, CT, USA, December 3-12, 2007. (2007) 11–16

[12] Recupero, D.R., Dragoni, M., Buscaldi, D., Alam, M., Cambria, E., eds.: Proceedings of 4th Workshop on Sentic Computing, Sentiment Analysis, Opinion Mining, and Emotion Detection (EMSASW 2018) Co-located with the 15th Extended Semantic Web Conference 2018 (ESWC 2018), Heraklion, Greece, June 4, 2018. In Recupero, D.R., Dragoni, M., Buscaldi, D., Alam, M., Cambria, E., eds.: CEUR Workshop. Volume 2111 of CEUR Workshop Proceedings., CEUR-WS.org (2018)

[13] Bawaneh, M., Alkoffash, M., Alrabea, A.: Arabic text classification using k-nn and naive bayes. Journal of Computer Science **4** (2008)

[14] Al-Tahrawi, M.M., Al-Khatib, S.N.: Arabic text classification using polynomial networks. Journal of King Saud University - Computer and Information Sciences **27** (2015) 437 – 449

[15] Elhassan, R., Ahmed, M.: Arabic text classification review. International Journal of Computer Science and Software Engineering (IJCSSE) **4** (2015) 2409–4285

[16] Syiam, M., Tolba, M., Fayed, Z., Abdel-Wahab, M., Ghoniemy, S., Habib, M.: An intelligent system for arabic text categorization. International journal of cooperative information systems **6** (2006) 1–19

[17] Duwairi, R.: Arabic text categorization. Int. Arab J. Inf. Technol. **4** (2007) 125–132

[18] Mesleh, A.M., Kanaan, G.: Support vector machine text classification system: Using ant colony optimization based feature subset selection. In: 2008 International Conference on Computer Engineering Systems. (2008) 143–148

[19] Ababneh, J., Almanmomani, O., Hadi, W., El-Omari, N., Alibrahim, A.: Vector space models to classify arabic text. International Journal of Computer Trends and Technology (IJCTT **7** (2014) 219–223

[20] Chouigui, A., Khiroun, O.B., Elayeb, B.: Ant corpus: An arabic news text collection for textual classification. In: 2017 IEEE/ACS 14th International Conference on Computer Systems and Applications (AICCSA). (2017) 135–142

[21] Abutiheen, Z.A., Aliwy, A.H., Aljanabi, K.B.S.: Arabic text classification using master-slaves technique. Journal of Physics: Conference Series **1032** (2018) 012052

[22] Marton, Y., Habash, N., Rambow, O., Alkhulani, S.: Spmrl'13 shared task system: The cadim arabic dependency parser. In: SPMRL@EMNLP. (2013)

[23] Gales, M.J.F., Knill, K.M., Ragni, A., Rath, S.P.: Speech recognition and keyword spotting for low-resource languages: Babel project research at CUED. In: SLTU, ISCA (2014) 16–23

[24] GaÃ«l, L., Romain, B., Antoine, D., Nadine, L.: Multilingual event extraction for epidemic detection. Artificial Intelligence in Medicine (2015) doi: 10.1016/j.artmed.2015.06.005.

[25] Grishman, R., Sundheim, B.: Message understanding conference 6: A brief history. In: Proc COLING. Volume 96. (1996) 466–471

[26] Elsadig, M., Ahmed, A., Himmat, M.: Information extraction methods and extraction techniques in the chemical document's contents: Survey. ARPN Journal of Engineering and Applied Sciences **10** (2015) 1068–1073

[27] Hajjar, M., Zreik, K.: Classification of arabic information extraction methods. In: 2nd International Conference on Arabic Language. (2009)

[28] Alruily, M., Ayesh, A., Zedan, H.: Arabic language in the context of information extraction task. In: Journal of Computational Linguistics Research. (2012)

[29] Church, K.: Empirical Estimates of Adaptation: The chance of Two Noriega's is closer to p/2 than p 2. In: Proceedings of the 18th International Conference on Computational Linguistics. (2000) 173–179