Traitement automatique des langues

Traitement automatique de l'arabe et des langues apparentées / NLP for Arabic and Related Languages

sous la direction de
Mona Diab
Nizar Habash
Imed Zitouni

Vol. 58 - n°3 / 2017

ATALA

Traitement automatique de l'arabe et des langues apparentées / NLP for Arabic and Related Languages

ATALA

Revue de
l'Association
pour le Traitement
Automatique
des Langues

TAL
Vol.
58

n°3
2017

Traitement automatique de l'arabe et des langues apparentées / NLP for Arabic and Related Languages

ATALA

Traitement automatique des langues

Revue publiée depuis 1960 par l'Association pour le Traitement Automatique des LAngues (ATALA), avec le concours du CNRS, de l'Université Paris VII et de l'Université de Provence

ISSN 1965-0906

https ://www.atala.org/revuetal

Traitement automatique des langues

Secrétaire
Aurélie Névéol - LIMSI, CNRS

Traitement automatique des langues

Volume 58 – n°3 / 2017

TRAITEMENT AUTOMATIQUE DE L'ARABE ET DES LANGUES APPARENTÉES / NLP FOR ARABIC AND RELATED LANGUAGES

Table des matières

Préambule

Emmanuel Morin* — **Sophie Rosset**** — **Pascale Sébillot*****

* *Université de Nantes, LS2N, France*
emmanuel.morin@univ-nantes.fr
** *CNRS, LIMSI, France*
sophie.rosset@limsi.fr
*** *INSA Rennes, IRISA, France*
pascale.sebillot@irisa.fr

TAL. Volume 58 – n°3/2017, pages 7 à 8

Préambule

Ce numéro est marqué par la disparition de notre collègue et amie Isabelle Tellier, qui nous a quittés le premier juin 2018. Isabelle, en tant que rédactrice en chef de la revue *TAL*, était responsable de la gestion de ce numéro spécial. Elle a assumé cette responsabilité tant que ses forces le lui ont permis. Nous sommes admiratifs devant un tel courage et une telle opiniâtreté.

Isabelle était un membre très actif du comité de rédaction de la revue, comité qu'elle avait rejoint en septembre 2007, avant de devenir rédactrice en chef en janvier 2016. Isabelle était une collègue avec une grande culture scientifique qui avait la passion de son métier. Son enthousiasme, son intelligence et sa gaieté vont nous manquer. Ce numéro lui est dédié.

Emmanuel Morin, Sophie Rosset, Pascale Sébillot

Rédacteurs en chef de la revue *TAL*

NLP for Arabic and Related Languages

Mona Diab* — **Nizar Habash**** — **Imed Zitouni*****

* *The George Washington University, USA*
mtdiab@gwu.edu
** *New York University Abu Dhabi, UAE*
nizar.habash@nyu.edu
*** *Microsoft Research, USA*
izitouni@microsoft.com

1. On Arabic and Natural Language Processing

Arabic natural language processing (NLP) is a challenging field of research. This is due to many factors including Arabic's complex and rich morphology, its high degree of ambiguity as well as the presence of a number of dialects that vary quite widely. Furthermore, Arabic has many important geopolitical connections and is spoken by over 400 million people in countries with varying degrees of prosperity and stability. Arabic in its standard form, known as Modern Standard Arabic (MSA), is one of the 6 official languages of the United Nations. It is the primary language of the latest world refugee problem affecting the Middle East and Europe. The opportunities that are made possible by working on this language and its dialects cannot be underestimated in their consequence on the Arab World, the Mediterranean Region and the rest of the world.

Apart from its geopolitical significance, Arabic poses interesting challenges to NLP in general. Given its complexity when considered with its dialects, the language pushes the boundaries of NLP as it forces researchers to think of creative solutions posed by the inherent nature of the language. The use of Arabic is diglossic, the standard language used in formal settings and in education is significantly different from the spoken vernaculars. The spoken vernaculars are quite diverse depending on whether they are spoken in cities or rural areas, settled communities or Bedouin environments. The variational dimensions are quite pronounced. We find social variations such as educated versus lay, male vs. female, urban vs. rural, re-

TAL. Volume 58 – n°3/2017, pages 9 à 13

ligious variants. The differences are not only in the accent but actually dialectal variants that are reflected in the lexical choice, morpho/phonological variations. To add to the complexity of the linguistic situation, Arabic spoken and written modalities typically code switch within utterance. In the written modality, we see pervasive code switching between the vernaculars and MSA; in the spoken variety, we observe rampant code switching with French, English, Italian and Spanish depending on the country. Moreover, in the written modality, even for MSA, Arabic is underspecified for short vowels leading to even more ambiguity over and above natural lexical/syntactic/phonological/semantic/pragmatic ambiguity present in natural language.

Before the onslaught of social media, Arabic in the digital world was perceived as strictly MSA with potential contrast to Classical Arabic (the language pertaining to old historical books dating back to the 6th-19th century). However with the ubiquity of current technology from SMS to chat rooms in the context of social media, we note the pervasive presence of written/spoken dialectal. The challenge with processing such resource is manyfold. Mainly, we have no written standard orthographies for these dialectal varieties that not only vary across Arab countries but actually within the same country along geographical and social continua. Accordingly, the nature of such linguistic expressions is quite low resource by definition. Therefore building resources and solutions that adopt domain adaptation techniques by default is necessary, especially if we need to scale solutions beyond one variety of Arabic to cater to all Arabics. Moreover robust solutions devised for Arabic processing could serve as solutions for other languages with multiple varieties living side by side such as Indonesian and Malay. Given the importance of Arabic, there has been a lot of progress in the last fifteen years in the area of Arabic NLP. This special edition of *TAL* intends to provide a forum for researchers to share and discuss their work.

2. Summaries of Articles

There are four contributions in this special issue. All of the articles have a common focus on computational semantics, although with different approaches and tasks. The first article presents a survey on Arabic sentiment analysis, a very popular area of research that has grown very fast in the last few years. The second article is a specific Arabic sentiment analysis study targeting the Algerian dialect of Arabic. The third article approaches computational semantics for Arabic from a Frame Semantics point of view, and describes the challenges of building an Arabic version of FrameNet. Last but not least, the fourth article presents a framework for information extraction in Arabic that supports Arabic's complex and rich morphology.

2.1. *Modern Trends in Arabic Sentiment Analysis: A Survey (Mulki, Haddad and Babaoğlu)*

The growth of Arabic textual content on social media platforms, together with the continuous crises in the Arab World, have evoked the need to analyze the opinions of

the public regarding ongoing events. As a result, Arabic Sentiment Analysis (ASA) has become the focus of many recent NLP studies. With several Arabic NLP resources being publicly available along with the emergence of deep learning techniques, researchers could handle the complex nature of Arabic language more efficiently. In the last decade, various ASA systems have been built. Yet, their achievements have not been investigated or compared against each other. This survey covers the ASA research carried out during the past five years. The survey compares and evaluates the performances and gives insight into the ability of the created resources to support future ASA research.

2.2. *Une approche fondée sur les lexiques d'analyse de sentiments du dialecte algérien (Guellil, Azouaou, Saâdane and Semmar)*

The paper presents a tool for sentiment analysis of the Algerian dialect, combining the use of lexicons of sentiments and the processing of agglutination. The proposed approach starts by building a lexicon of Algerian-dialect sentiments leveraging an English lexicon. Then, a morphological analysis for sentiment analysis is conducted, where the valence and intensity of the sentiment in the input text is defined. This is different from most of sentiment analysis tools that are currently available since they are capable of processing only MSA. Experiments are conducted using two sentiment lexicons and a test corpus of 700 messages. Results show improvement of performance at each processing step.

2.3. *L'analyse et l'annotation à base de FrameNet: contribution à l'étude contrastive des événements de mouvement en arabe et en anglais (Lakhfif and Laskri)*

The paper describes a computational approach based on Frame Semantics for Arabic language processing. This approach is based on the adaptability of Berkeley FrameNet database and the transferability of FrameNet tools for Arabic, a language that differ typologically from English. The paper describes an attempt to build an equivalent Arabic FrameNet where it shows the use of such a semantic resource for Arabic text semantic analysis, representation and annotation. A frame based contrastive study of motion-events expressions in bilingual text (English-Arabic) is presented, using FrameNet based tool for semantic annotation. Results conducted on a corpus of motion events expressions confirm the cross-linguistic nature of Frame Semantics approach and the suitability of the theory for Arabic processing.

2.4. *Morphology-based Entity and Relational Entity Extraction Framework for Arabic (Jaber and Zaraket)*

Rule-based techniques and tools to extract entities and relational entities from documents allow users to specify desired entities using natural language questions, finite state automata, regular expressions, structured query language statements, or proprietary scripts. These techniques and tools require expertise in linguistics and programming. They lack support of Arabic morphological analysis which is key to process Arabic text. This work presents MERF, a morphology-based entity and relational entity extraction framework for Arabic text. MERF provides a friendly interface where the user, with basic knowledge of linguistic features and regular expressions, defines tag types and interactively associates them with regular expressions defined over Boolean formulae. Boolean formulae range over matches of Arabic morphological features, and synonymity features. Users define relations with tuples of subexpression matches and can associate code actions with subexpressions. MERF computes feature matches, regular expression matches, and constructs entities and relational entities from the user-defined relations. MERF is evaluated with several case studies and compared with existing application-specific techniques. The results show that MERF requires shorter development time and effort compared to existing techniques and produces reasonably accurate results within a reasonable overhead in run time.

Acknowledgments

We want to thank the *TAL journal* editors and committee as well as the specific scientific committee. We are particularly grateful to the reviewers for their time and effort to improve this special issue.

We dedicate this issue to the memory of Isabelle Tellier.

Specific scientific committee (by alphabetical order): Tiba Zaki Abdulhameed (Western Michingan University, USA); Muhammad Abdulmageed (University of British Columbia, Canada); Motasem Alrahabi (University of Sorbonne Abu Dhabi, UAE); Mohammad Attia (Google Inc., USA); Eric Atwell (Leeds University, UK); Yassine Benajiba (Symanto Group, USA); Houda Bouamor (Carnegie Mellon University in Qatar, Qatar); Tim Buckwalter (University of Maryland, USA); Violetta Cavalli-Sforza (Al Akhawain University, Morocco); Khalid Choukri (ELRA , France); Mona Diab (The George Washington University, USA); Joseph Dichy (University Lyon 2, France); Heba Elfardy (Columbia University, USA); Yannick Estève (University of Le Mans, France); Mahmoud Ghoneim (The George Washington University, USA); Nizar Habash (New York University Abu Dhabi, UAE); Bassam Haddad (University of Petra, Jordan); Kais Haddar (University of Sfax, Tunisia); Lamia Hadrich-Belguith (University of Sfax, Tunisia); Hazem Hajj (American University of Beirut, Lebanon); Denis Jouvet (INRIA, France); Omar Larouk (ENSSIB Lyon, France); Mohsen Rashwan (Research and Development International (RDI), Egypt); Khaled Shaalan (British University of Dubai, UAE); Khalil Sima'an (University of

Amsterdam, Netherlands); Kamel Smaïli (University of Lorraine, France); Abdelhadi Soudi (École Nationale Supérieure des Mines, Morocco); Nadi Tomeh (University of Paris 13, France); Stephan Vogel (Qatar Computing Research Institute, Qatar); Wajdi Zaghouani (Carnegie Mellon University in Qatar, Qatar); Aya Zirikly (The George Washington University, USA); and Imed Zitouni (Microsoft, USA).

Modern Trends in Arabic Sentiment Analysis: A Survey

Hala Mulki* — **Hatem Haddad**** — **Ismail Babaoğlu***

* *Department of Computer Engineering, Selcuk University, Turkey.*
** *Department of Computer & Decision Engineering, Université Libre de Bruxelles, Belgium.*

ABSTRACT. *The growth of the Arabic textual content on social media platforms has been caused by the continuous crises in the Arab World evoking the need to analyze the opinions of the public against the ongoing events. Arabic Sentiment Analysis (ASA) is, therefore, becoming the focus of many recent NLP studies. With several Arabic NLP resources being publicly available along with the emergence of deep learning techniques, researchers could handle the complex nature of Arabic language more efficiently. In the last decade, various ASA systems have been built. Yet, their achievements have not been investigated or compared against each other. This survey covers the ASA research carried out during the past five years. We compare and evaluate the performances and give insight into the ability of the created Arabic resources to support the future ASA research.*

RÉSUMÉ. *La croissance du contenu arabe dans les médias sociaux a été causée par les crises dans le monde arabe, évoquant la nécessité d'analyser les réactions du public à l'égard des événements en cours. L'analyse des sentiments de la langue arabe est au centre d'intérêt de plusieurs études en TAL. Avec l'émergence de plusieurs TAL ressources en arabe accessibles au public ainsi que l'émergence de techniques d'apprentissage approfondi, les chercheurs pourraient gérer la nature complexe de la langue arabe plus efficacement. Au cours de la dernière décennie, plusieurs systèmes d'analyse du sentiment dans le contenu arabe (ASA) ont été développés. Cependant, leurs performances n'ont pas été étudiées ou comparées entre elles. Cette enquête couvrirait les travaux de ASA réalisés au cours des cinq dernières années. Nous comparons les résultats, évaluons les performances et donnons un aperçu de la capacité des ressources arabes créées à soutenir la recherche future dans le domaine ASA.*

KEYWORDS: *Arabic Sentiment Analysis, supervised learning, lexicon-based, word embeddings.*

MOTS-CLÉS : *analyse du sentiment dans le contenu arabe, apprentissage supervisé, approche basé sur le lexique,* embedding *de mots.*

1. Introduction

Online shared opinions towards events or products are becoming a rich source of information required for analytical studies. Sentiment Analysis (SA) is a Natural Language Processing (NLP) task that facilitates performing such studies by mining the opinionated content in a piece of text (Piryani *et al.*, 2017). The uprisings in the Arab world that started in 2010 have led to a significant growth in the online Arabic content shared across social media platforms. The ability to analyze such vast opinionated content has attracted the attention of NLP researchers. Consequently, multiple Arabic sentiment analysis (ASA) systems could be developed to capture the sentiment at the different analysis levels; some of them used traditional machine learning approaches, others exploited semantic resources through lexicon-based models, while more recent studies employed the newly-emerged deep learning techniques. Through the presented ASA research, several semantic resources, annotated datasets and trained word embedding vectors were created. Therefore, it is crucial to conduct a comprehensive evaluation of what has been achieved in order to give insight into the potential improvements that could be done in terms of SA tools and resources.

In this paper, we present a survey of ASA research introduced during the last five years. The reviewed research works have been classified according to the used method and the analysis level. In addition, we compare the obtained results and evaluate the performances to recognize by which method, using which semantic resource and for which dataset a better performance can be achieved. Moreover, we shed light on the potential future exploitation of the produced tools and resources to develop more efficient ASA systems.

2. Sentiment Analysis

Sentiment refers to the people's opinions or emotions towards entities, events or ideas. It represents the opinionated content that implies a positive, negative or neutral polarity expressed within a written text (Turney, 2002). According to Liu (2012), sentiment analysis (SA) or opinion mining aims to develop automated techniques to analyze the opinions embedded in a piece of text.

Most of the proposed SA models adopt a general pipeline as follows:

1) Preprocessing: it reduces the noisy nature of the input text, especially for data derived from social media. NLP techniques such as tokenization, normalization, stemming, lemmatization, Part Of Speech (POS) tagging, denoising and stopwords removal are usually used. Consequently, the none-sentimental content represented by special characters, punctuation, duplicate-characters, typos etc. could be eliminated. Moreover, stemming and lemmatization reduce the size of features to be used in the subsequent phase as inflected words are returned to their roots or lemmas (Assiri *et al.*, 2015);

2) Feature Extraction and Selection: the preprocessed data forwarded to this phase facilitates syntactic features extraction since some preprocessing tasks like POS tagging, stemming and lemmatization, negation and emotion tagging can be considered key indicatives of the sentiment (Assiri *et al.*, 2015). In addition, through tokenization, the common bag-of-words and n-grams feature schemes are produced. Feature vectors can then be formulated via binary weighting due to the presence/absence of a word or n-gram in a specific input text. Furthermore, the relative importance of a term or an n-gram which is usually decided by its frequency of occurrence in the dataset, reduces the features' dimensionality by keeping terms of specific frequency values. On the other hand, sentiment lexicons provide another set of features where a term's sentiment score or intensity value define the text features. All these features are called "hand-crafted"; they have been used in most of the presented SA works (Mohammad *et al.*, 2013; Al-Osaimi and Badruddin, 2014; Abdulla *et al.*, 2013; Salameh *et al.*, 2015). More recently, a novel type of features has emerged, the so-called text embeddings where words, phrases and sentences are mapped into real-valued, low-dimensional feature vectors to be used within deep learning systems (Collobert *et al.*, 2011; Al Sallab *et al.*, 2015; Mdhaffar *et al.*, 2017);

3) Sentiment Classification: in natural languages, the sentiment is normally included in the subjectivity concept as the latter represents the language's aspects of opinions and impressions (Liu, 2012). Therefore, SA involves performing a subjectivity classification task first so that a unit of text (term, phrase, sentence or document) is classified as either objective or subjective. Then, the subjective text is classified into the polarity it implies which might be positive, negative, neutral or even mixed. Sentiments can be annotated at various levels of granularity: word or phrase, aspect, sentence and document. Regardless of the level at which sentiment is captured, the sentiment classification process is conducted using one of two main approaches: Machine Learning (ML) or rule-based. Both approaches exploited syntactic, lexicon-derived or embeddings features and were applied successfully in many SA research (Turney, 2002; Salameh *et al.*, 2015; Altowayan and Tao, 2016).

3. Arabic Sentiment Analysis Challenges

Sentiment analysis has become a very active area of NLP research since the advent of Web 2.0 technologies. Nevertheless, most of the presented SA research has been dedicated towards Indo-European languages while under-represented languages such as Arabic were remarkably less tackled. Despite the recent growth of the public Arabic content across social networks and with the continuous development of Arabic NLP tools, ASA research still faces challenges, most of which are related to the Arabic language itself. Arabic has three main variants: Classical Arabic used in Quran, Modern Standard Arabic (MSA), which is the formal type of Arabic, and the informal Arabic known as colloquial or Dialectal Arabic (DA) which combines several different dialects (Al-Kabi *et al.*, 2013). With such variety, where each form of Arabic has its own complexities which are represented by special linguistic and morphological features, SA has to handle further issues beyond those already existing for textual data.

Here, we highlight the major challenging issues encountered while conducting ASA:

– *Complex morphology*: being a Semitic language, Arabic adopts the root-and-pattern representation where a single set of consonants called the "root" is used to derive a variety of words by adding vowels (a,o,i) (ا، و، ي) or short vowels (diacritics) in addition to other consonants (Habash, 2010). The inflectional morphology, however, is observed through the ability of Arabic language to express a word in several grammatical categories while keeping the same meaning. The word's inflected forms can be obtained for several categories such as person, tense, voice (active/passive), number, gender, etc. Consequently, with such high derivational and inflectional morphology, handling Arabic texts through customizing current English SA systems and tools might be limited (Habash, 2010). Thus, special preprocessing tasks supported by Arabic-oriented morphological analyzers should be combined in ASA systems;

– *Lack of resources*: despite the abundant online Arabic content, there is a lack of Arabic sentiment datasets and sentiment lexicons. During the last decade, some datasets have been constructed either for MSA or DA, nevertheless, the number of sentiment datasets which are publicly available remains little (Assiri *et al.*, 2015). Besides, most of these datasets do not have enough amount of data which affects the evaluation of ASA systems when compared to English SA models since the sentiment analysis accuracy depends on the size of the manipulated data. On the other hand, the difficulties that accompany the construction and annotation process of sentiment lexicons have hindered the provision of large-scale and highly-coverage Arabic lexicons, especially with the existence of different Arabic dialects and domains;

– *Negation and sarcasm*: negation in Arabic is expressed using specific negation words which indicate the meaning "not"; some of them are: "ما", "لم" and "لا". Negation should be accurately detected and handled as it can convert the meaning of a sentence yielding a quite opposite polarity. This task becomes more difficult and challenging when dealing with DA where negation words are so different from formal MSA ones and have several meanings such as "مو" meaning "not" in the Levantine dialect that can be used for negation (e.g. السلطة مو تازة[1]) or interrogative (e.g. مو، بوكرا تجي[2]) which might mislead the sentiment classifier. Another ambiguity faced by ASA models is the sarcasm issue in which the explicit polarity totally opposites the meant sentiment as in e.g. بعد الانتظار لساعتين، نفدت كل التذاكر، كم انا محظوظ[3], where the word "محظوظ", which means "lucky", indicates a positive sentiment while in the example it actually refers to the opposite;

– *Arabizi usage*: Arabizi is considered a newly-emerged Arabic variant written using the Arabic numeral system and Roman script characters (Assiri *et al.*, 2015). It is commonly used while expressing DA across social media and poses a challenge

1. "The salad is not fresh."
2. "You're coming tomorrow, aren't you?"
3. "After waiting for two hours, all tickets were sold; Lucky me."

to sentiment analysis when it is mentioned along with Arabic (e.g. 3an jad كتير الفلم 7elou. [4]). This requires proper tools to interpret Arabizi into either MSA or DA before conducting the sentiment classification task;

– *Dialects variances*: DA forms the majority of the online opinionated Arabic content as it is commonly used across social media platforms. DA combines various dialects which differ according to the geographical location. Each dialect has its own vocabulary, syntactic and grammatical rules in addition to special idioms. On the other hand, despite that all dialects are derived from MSA and hence do share some vocabulary, common words or expressions among two dialects might have drastically different sentiments. For example, "يعطيك العافية" is a compliment of a positive sentiment that means "May God grant you health" in the Levantine dialect, while this very same phrase has an aggressive meaning of "Burn in hell" in the Tunisian dialect. Considering these variances, an ASA system that targets one dialect might not be efficient for another as it is developed with a dialect-dependent tools such as the morphological analyzer, stopwords/negation words and sentiment lexicons.

4. Arabic Sentiment Analysis Research

Earlier ASA studies had to handle the complex nature of Arabic through limited feature types and resources. However, with more MSA and DA morphological analysis and disambiguation tools becoming available, the ASA task was facilitated as these tools could provide a wide variety of syntactic and stylistic features such as 1-best tokenization, POS tags, stems, lemmas and diacritization in one fell swoop. On the other hand, exploitation of web forums and social media enabled the provision of sentiment datasets and lexicons needed for developing and evaluation of MSA and multi-dialectal SA systems (Rushdi-Saleh *et al.*, 2011; Mourad and Darwish, 2013; Badaro *et al.*, 2014; Nabil *et al.*, 2015).

ASA research has been conducted at different linguistic levels: word or phrase, aspect, sentence and document. The following subsections review the recent major studies achieved at each level.

4.1. *Words-level Sentiment Analysis*

Determining the semantic orientation of sentiment-bearing words or phrases in a corpus is essential for sentiment lexicon construction. Sentiment lexicons are fundamental for computing the sentence or document sentiment through lexicon-based methods or as features for machine learning methods. Sentiment lexicons can be compiled by means of three strategies: manually with the assistance of a linguist and native speakers, automatically based on another dictionary (dictionary-based) or us-

4. "The film is really amazing."

Paper	Construction method	Size	Arabic variant	Assigned polarity
(El-Beltagy and Ali, 2013)	Corpus-based	4,392	Egyptian/MSA	Pos/Neg
(Abdulla *et al.*, 2014)	Manually Semi-automatic Corpus-based	4,815 9,100 8,618	MSA & DA	Pos/Neg
(Duwairi *et al.*, 2015)	Manually	2,376	MSA	Pos/Neg
(Assiri *et al.*, 2017)	Corpus-based + Dictionary-based	14,000	Saudi	Pos/Neg
(Abdul-Mageed *et al.*, 2014)	Manually	3,982	MSA /DA	Pos/Neg

Table 1. *The lexicons constructed and evaluated within the reviewed ASA studies.*

ing the corpus itself (corpus-based) or semi-automatically where manual interference is needed to normalize the automatically-built lexicon (Liu, 2012).

Considering the lexicons built within the works reviewed here (see Table 1), Abdulla *et al.* (2014) presented three sentiment lexicons built using manual, semi-automatic and automatic methods. The first lexicon has 4,815 entries. It was manually constructed through translating seed words from SentiStrength English lexicon using an English-Arabic dictionary with their polarity assigned manually. The translated seeds were then expanded by adding synonyms of each word under the same polarity in addition to the most common MSA words derived via Term Frequency (TF) weighting, emotions and dialectal terms from different Arabic dialects. The second one was built through the direct translation of SentiStrength using Google translate. Human interference was needed to normalize and clean the translated Arabic version yielding a lexicon of 9,100 entries. As for the third lexicon, it was compiled using a corpus-based automated approach in which the most common positive and negative terms were derived from the annotated corpus via TF weighting. For terms having both polarities, the polarity of the term whose TF is greater was adopted.

In Duwairi *et al.* (2015), the authors adopted the same scenario used in Abdulla *et al.* (2014) to manually build an MSA sentiment lexicon of 2,376 words. The lexicon was expanded through adding synonyms using Sakhr dictionary (Reyes and Rosso, 2014), stems and emotions.

Based on an assumption that sentiment terms often appear with other terms of same polarity, El-Beltagy and Ali (2013) presented an Egyptian lexicon built using a corpus-based method. As a first step, a list of 380 sentiment words seeds was used. Then following their hypothesis, the authors expanded this list by looking for patterns containing these seeds and their accompanied single terms.

Assiri *et al.* (2017) constructed a Saudi dialect lexicon by integrating the dictionary-based and corpus-based methods. A list of Saudi seed words was expanded using the method by El-Beltagy and Ali (2013), then terms from a pre-created lexicon by Badaro *et al.* (2014) were added to the Saudi lexicon after they were subjected to

normalization and cleaning processes, then a collection of Saudi terms were manually added resulting in a lexicon of 14,000 sentiment terms.

4.2. *Aspect-level Sentiment Analysis*

Aspect-level SA identifies sentiment targets crucial for applications such as question-answering and recommendation systems (Liu, 2012). Due to the complexity of the Arabic language, aspect-level SA was less tackled by ASA research. In Farra and McKeown (2017), the authors proved that handling the richness of the Arabic language through specific morphological representations makes important targets (entities) and the sentiment towards them better identified. This was done using a framework of two cascaded sequence labeling CRF models: target-specific model and sentiment-specific model. While the first model is responsible for recognizing the entities, the second one predicts the sentiments towards these entities. Both models were trained to provide a sequence of entity/sentiment labels for the input tokens. The merit provided by this system is that the training phase involves learning the syntactic relations between entities and sentiment-bearing words. For this purpose, MADAMIRA morphological analyzer (Pasha *et al.*, 2014) was exploited as it enables an advanced tokenization with which multiple morphological representations could be formulated. For instance, clitics such as the definition article "ال التعريف" that usually indicates an entity could be split off the word, combined with a detailed POS feature and fed into the entity-specific model leading to an improved recall of the recognized entities. The proposed system was evaluated using 1,177 online comments with annotated targets and a lack of punctuation obtained from Arabic Opinion Target (Farra *et al.*, 2015) which is a part of Qatar Arabic Language Bank (QALB) corpus (Zaghouani *et al.*, 2014). The experimental results concluded that both models have achieved better results compared to multiple lexical baselines.

4.3. *Document and Sentence-level Sentiment Analysis*

Document and sentence-level SA works form the majority of the recent ASA research. Therefore, we dedicate the next section to cover the important SA studies conducted at this level. The reviewed studies are organized according to the methods used to build the ASA models.

5. Document and Sentence-level Sentiment Analysis Approaches

Arabic Sentiment Analysis can be conducted using traditional machine learning approaches such as supervised/unsupervised and hybrid, deep learning approaches (supervised, unsupervised) or rule-based approaches (Lexicon-based). The following subsections introduce a detailed explanation of some of these methods in addition to the state-of-the-art research related to it.

5.1. *Supervised Learning-based Approaches*

Supervised learning methods require a labeled corpus to train the classifier how to predict the text polarity (Biltawi *et al.*, 2016). The learning process is carried out by inferring that a combination of a sentence's specific features yields a specific polarity class (Shoukry and Rafea, 2012). The most common features used are bag-of-words and bag-of-n-grams features in addition to various linguistic features extracted by morphological analyzers. Having the features extracted, sentiment classification is then performed using supervised learning classification algorithms such as Support Vector Machines (SVM), Naive Bayes (NB), Decision Tree (DT) and NEUNET (NN) (Biltawi *et al.*, 2016). Research works that adopted supervised approaches were concerned about which preprocessing tasks, features or classification algorithms can lead to a better classification performance either for MSA or DA.

Considering the wide spread of the Egyptian dialect across Twitter, enriching the Arabic sentiment resources with a pure Egyptian sentiment corpus along with Egyptian-specific preprocessing tools was the aim of Shoukry and Rafea (2012). They collected a dataset of 1,000 positive/negative Egyptian tweets to test their supervised SA model. The preprocessing included removing usernames, hashtags, URLs and non-Arabic letters. To classify a tweets' sentiment, SVM and NB were employed in two experiments. In the first one, stopwords were kept, while in the second they were omitted. Results revealed that SVM performed better than NB in both experiments achieving an accuracy of 72%, compared to 65% scored by NB.

To examine the impact of combining emotion icons in SA, Al-Osaimi and Badruddin (2014) introduced an SA model for multi-dialectal tweets. The collected corpus included 3,000 positive, negative and neutral tweets. Term Frequency-Inversed Document Frequency (TF-IDF) was used to extract the features. Sentiment classification was conducted using NB and KNN algorithms. Results showed that preserving emotion icons enhanced the model's performance as the best accuracy achieved by NB classifier increased from 58.28% to 63.79%.

The recently-emerged form of Arabic (Arabizi) was investigated in Duwairi *et al.* (2014). The study seeked to convert the dialectal and Arabizi content into MSA. A dataset of 1,000 positive/negative/neutral tweets written in Jordanian and Arabizi was collected. For preprocessing, stemming, tokenization, stopwords filtering tasks were applied in addition to the conversion of Jordanian and Arabizi to MSA. Morphological features, negations and emoji were included in the features set. The authors observed that, if stemming and stopwords removal are disabled, better performance can be achieved, while negation detection and conversion from Arabizi to MSA did not achieve a remarkable improvement in the evaluation measures. KNN, SVM and NB classifiers were used, where NB was the best with an accuracy of 76.78%.

In Salamah and Elkhlifi (2016), an under-represented Arabic dialect was investigated where a dataset of 340,000 Kuwaiti tweets were collected and manually annotated for positive and negative polarity. Tweet-related features and opinions-oriented ones were extracted. The opinion-oriented features were obtained from 22 manually-

Paper	Algorithm/features	Dataset	Evaluation
(Shoukry and Rafea, 2012)	SVM, NB unigrams+bigrams	1,000 tweets Egyptian pos/neg	Best: SVM acc=72%
(Al-Osaimi and Badruddin, 2014)	NB, KNN TF-IDF unigrams	3,000 tweets multi-dialects pos/neg/neut	Best: NB acc=58.28% (-emoji) acc=63.79% (+emoji)
(Duwairi *et al.*, 2014)	KNN, SVM, NB syntactic, negation emoji	1,000 tweets Jordanian/Arabizi pos/neg/neut	Best: NB acc=76.78%
(Salamah and Elkhlifi, 2016)	SVM, J48, RT, DT tweet-related emotion-bearing words	340,000 tweets Kuwaiti pos/neg	Best: SVM F1=71.5%
(Abdul-Mageed, 2015)	SVM, NB, IB1 POSs tokens	1,552 sentences MSA subj/obj	Best: SVM acc=85%

Table 2. *Summary of supervised learning-based ASA research works.*

built classes that combine emotions-bearing words. SVM, J48, Random Tree (RT) and decision Tree (DT) classifiers were used. SVM scored the best results with an F1-score of 71.5% compared to 42%, 48% and 51% achieved by J48, DT and RT respectively. A summary of the above mentioned research papers is listed in Table 2 where Best, acc and F1 indicate the best method, the scored accuracy and F1-score respectively.

One of the pioneering works about subjectivity detection in MSA was presented by Abdul-Mageed (2015). The author hypothesized that using specific tokens would favorably impact the subjectivity classification task. The proposed model was trained with a collection of words having certain POS tags such as ADJ, ADV and NOUN_PROP. The experiments were conducted using Penn Arabic Treebank dataset (Popescu and Etzioni, 2007) with several ML techniques applied: SVM, NB, and Instance-based learning. These techniques were compared against each other with two types of features' settings: frequency and presence vectors. In all experiments, the preprocessing step was essential as the study highlighted that the rich morphology of MSA imposes using the compressed form of words in order to obtain a better model generalization. The obtained results emphasized the positive impact of using certain tokens rather than all the words for training; moreover, similar to the SA task, SVM was found of the best performance for subjectivity classification, compared to other ML methods where it scored a high accuracy equals to 85%.

Hybrid approaches combine lexical and linguistic features together with lexicon-derived features to perform sentiment analysis. This involves incorporating the term's polarity score defined by a sentiment lexicon in the features set needed to train a supervised sentiment classifier (Biltawi *et al.*, 2016). Abdul-Mageed *et al.* (2014) studied the efficiency of standard and genre specific features when used to express MSA and

DA seeking for the best scheme to represent lexical information within SA context. To do that, the authors constructed an adjective sentiment lexicon to enrich the lexical features. Their SA system SAMAR has two classification stages: subjectivity and polarity classification. Four MSA/dialectal datasets were collected manually including reviews and tweets. Syntactic features, extracted via AMIRA morphological analyzer (Diab, 2009), were adopted in addition to an extra feature resulted from the matches between the input tokens and the adjectives contained in the manually-built lexicon. Moreover, a novel feature that distinguishes MSA from DA was added. The used lexicon includes 3,982 labeled adjectives. Experimental study showed that using SVM trained with the different features enabled beating the baselines for most datasets either for subjectivity classification with a best accuracy of 73% or for the sentiment classification with an accuracy of 70.30% for DAR dataset.

To compensate for the lack of publicly available resources, Salameh *et al.* (2015) suggested using publicly available English NLP tools and lexical resources. This study presented an ASA model that employs an English SA system with an English lexicon on a translated Arabic content. Four datasets of positive/negative/neutral tweets and social media posts written in MSA/dialectal were used. Preprocessing included normalization, tokenization and POS tagging to produce syntactic and stylistic features. An English SA model NRC-Canada designed by Mohammad *et al.* (2013) was modified to handle the Arabic text along with a translated version of NRC Hashtag Sentiment Lexicon. The Arabic content translated to English was targeted using the system developed in Kiritchenko *et al.* (2014). The obtained accuracy values for Levantine datasets were 78.65% for the Syrian dataset and 63.89% for BBN.

Baly *et al.* (2017) introduced a hybrid model OMAM whose features were inspired from the English SA model (Balikas and Amini, 2016). An equivalent set of surface, syntactic and semantic features were obtained with the assistance of MADAMIRA from Pasha *et al.* (2014) and SAMA by Maamouri *et al.* (2010) morphological analyzers. Additional features were provided by ArSenL (Badaro *et al.*, 2014), AraSenti (Al-Twairesh *et al.*, 2016) and ADHL (Mohammad *et al.*, 2016) lexicons. Preprocessing phase included replacing emotions, URLs and hashtags with special tokens. The model was applied on dialectal Arabic tweets provided by SemEval-2017 (Rosenthal *et al.*, 2017). Results indicated that SVM classifier trained with the previous features achieved an F1 score of 42.2%, a recall of 43.8% and an accuracy of 43%.

With the key role of the lexicon-derived features in improving the performance of hybrid SA systems, there was a crucial need for a large-scale, domain-independent, high-coverage and publicly-available Arabic lexicon. To meet that need, Al-Moslmi *et al.* (2017) introduced the Arabic senti-lexicon to assist in sentiment classification of multi-domain, multi-dialectal Arabic reviews. The quality of the constructed lexicon towards SA task was assessed through training the model with five types of feature sets most of which were lexicon-derived. Features included sentiment words' polarity-based, sentiment words' presence-based, frequency POS-based, sentence level-based and other features related to words and sentences statistics. SVM, NB, LLR, KNN and neural network (NEUNET) were employed. To evaluate the presented model, the au-

Paper	Hybrid features	Algorithm	Dataset	Evaluation
(Abdul-Mageed *et al.*, 2014)	Linguistic syntactic adjective polarity score from Adj-Lex	SVM several kernels	DAR: 2,798 TGRD: 3,015 THR: 3,008 MONT: 3,097 MSA/dialectal pos/neg/neut	Best: SVM linear kernel acc=70.3% (DAR)
(Salameh *et al.*, 2015)	linguistic word N-grams Char N-grams score from translated-Lex	SVM	1,111 dialectal tweets pos/neg 1,200 Levant comments pos/neg/neut 2,000 Syrian tweets pos/neg/neut 2,681 dialectal tweets pos/neg	acc=85.23% (dialects)
(Baly *et al.*, 2017)	linguistic syntactic emotion presence tweet-related score from MSA/dialects-Lex	SVM	3,355 dialectal tweets pos/neg/neut	acc=43%
(Al-Moslmi *et al.*, 2017)	N-grams sentence-level syntactic score from ArabicSenti-Lex	SVM, NB, LLR, KNN NN	8,861 reviews dialects pos/neg	Best: LLR, NN F1=97%

Table 3. *Summary of Hybrid ASA research works.*

thors created a dataset called Multi-domain Arabic Sentiment Corpus (MASC) including 8,861 positive/negative customer reviews written in several Arabic dialects. Data was first preprocessed in terms of tokenization, normalization, stemming and stopwords removal. The model was trained on each feature set solely, then on all of them combined in one set. Results indicated that, SVM achieved the best results when only POS-based features are included. However, when all features are used for training, LLR, NN and NB were of better performance where LLR and NN achieved an F1-score of roughly 97%, while NB achieved 96% compared to 82.07% and 77.97% F1-scores achieved by SVM and KNN respectively. A summary of the above-mentioned research papers of hybrid SA models is listed in Table 3 where Best, acc and F1 indicate the best method, the scored accuracy and F1-score respectively.

5.2. *Lexicon-based Approaches*

In lexicon-based methods, neither labeled data nor training step are required to design the sentiment classifier. The polarity of a sentence or a document is determined via the lexicon-derived sentiment scores of its constituent words (Liu, 2012). A sentiment lexicon combines a list of subjective words and phrases along with their positive or negative score which denotes the sentiment polarity and strength of a word/phrase (Piryani *et al.*, 2017). Sentiment lexicons can be general-purpose or domain-specific, built either manually or automatically (Abdulla *et al.*, 2013). For each entry in the lexicon, the sentiment weight or score is assigned by one of these weighting algorithms:

– Straight Forward Sum (SFS) method: adopts the constant weight strategy to assign weights to the lexicon's entries such that negative words have the weight of –1 while positive ones have the weight of 1. The polarity of a given text is thus calculated by accumulating the weights of negative and positive terms and the total polarity is determined by the sign of the resulted value. Thus, for a tweet such "غوغل مبدعة شي خرافي"[5], the polarity is calculated as follows: google+incredibly+creative= 0+1+1=+2. The tweet has a positive polarity;

– Double Polarity (DP) method: assigns both a positive and a negative weight for each term in the lexicon. For example, if a positive term in the lexicon has a weight of 0.6, then its negative weight will be: – (1–0.6) =–0.4. Similarly, a negative term of a weight equals to –0.9 would have a 0.1 positive weight. Polarity is calculated by summing all the positive weights and all the negative weights in the input text. Consequently, the final polarity is determined according to the greater absolute value of the resulted sum (Piryani *et al.*, 2017). Thus, the positive score of the previous tweet is [0+0.5+0.8]=1.3 while the negative score=[0+(–0.5)+(–0.2)]=–0.7. Since the positive score is greater than the negative one, this indicates the positive polarity of the tweet assuming that "خرافي"[6] has a positive score of +0.8 and "مبدعة"[7] is of +0.5 positive score.

Lexicons of uniform weight along with the SFS method have been commonly used in most lexicon-based SA research. However, since SFS depends only on the counts of positive and negative words of a sentence to determine its polarity, it might lead to miss-classified instances under the label "neutral". This is encountered when the number of positive words in a sentence equals that of the negative words (Liu, 2012). For example, if a negative word such as "terribly" and a positive one like "exciting" are contained in the same sentence "the movie is terribly exciting", then the sentence's polarity score computed via SFS and uniform weight strategy would be: movie+terribly+exciting=0+(–1)+1= 0 which refers to a neutral polarity, while the previous sentence is obviously bearing a positive sentiment.

5. "Google is incredibly creative."
6. "Incredibly."
7. "Creative."

Several attempts were introduced to develop a novel weighting algorithm with which a better sentiment classification can be achieved. One of these attempts was presented by El-Beltagy and Ali (2013) where the authors noticed that sentiment terms often appear with other terms of same polarity. Based on this theory, they constructed a corpus-based lexicon (details in Section 4.1). Using the resulted lexicon, SFS and DP methods were adopted to determine the positive/negative/neutral sentiment of the input text. The model was also tested with the uniform weight scheme with negation switch policy, intensification words weighting and person names removal applied. Two manually-collected and annotated Egyptian datasets were used. The first one, called Dostour, combines 100 comments, while the second represents a Twitter dataset of 500 tweets. The best performance was achieved by the second weighting strategy with DP method where an accuracy of 83.3% was scored for Twitter dataset while for Dostour dataset, an accuracy of 63% was achieved.

Aiming to evaluate manually-built against the automatically-built lexicons for the SA task, Abdulla *et al.* (2014) examined performing sentiment analysis of MSA/dialectal Arabic using three lexicon variants built via different construction methods (see Section 4.1). In addition, an integrated lexicon resulted from merging the three constructed ones was also utilized for the final system evaluation. Two datasets were used in the experiments, the first contains 2,400 positive/negative comments from Maktoob collected by Al-Kabi *et al.* (2013), while the second combines 2,000 positive/negative/neutral tweets obtained from Abdulla *et al.* (2013) . Data was normalized and a light stemming was applied. Light stemming removes common affixes from words without reducing them to their stems or roots and thus retains the variety of words having same root and different meanings. Sentiment classification was then performed using the four lexicons one by one with SFS method and switch negation policy applied. Experiments showed that the stemming degraded the performance with manually-built and dictionary-based lexicons. In contrast, the accuracy was improved when stemming was applied on the corpus-based lexicon which forms more than half the size of the integrated lexicon. Hence, the best results were achieved with the integrated lexicon achieved since for Maktoob dataset, an accuracy of 74.6% was scored, compared to 70.2% with non-stemming option while for Twitter dataset, the scored accuracy was 70.2%, against 60.75% with non-stemming option.

In Duwairi *et al.* (2015), the authors claimed that when dealing with MSA data, the likelihood of finding a stem in the sentiment lexicon is higher than that of finding the original word. This has been investigated using an MSA sentiment lexicon constructed manually as it was explained in Section 4.1. A dataset of 4,400 positive/negative tweets was manually collected and annotated to evaluate the model. The data was preprocessed such that stopwords were removed while negations were kept. Stemming of the input data was conducted by MSA Khoja stemmer [8]. To investigate the stemming impact, experiments were conducted with/without stemming. SFS method with switch negation policy were employed to calculate the sentiment score of the input tweets. The results revealed that, for such MSA data, stemming has im-

8. http://zeus.cs.pacificu.edu/shereen/ArabicStemmerCode.zip.

proved the sentiment classification performance where the accuracy improved from 23% to 46%, while F1-score increased from 31.3% to 55.51%.

Paper	Scoring method	Lexicon/Features	Dataset	Evaluation
(El-Beltagy and Ali, 2013)	SFS, DP	Egyptian size:4,392 unigrams	1: 100 comments 2: 500 tweets Egyptian pos/neg/neut	Best: DP 1: acc=83.3% 2: acc=63%
(Abdulla *et al.*, 2014)	SFS	MSA/dialectal size: 19,800 unigrams	1: 2,400 comments 2: 2,000 tweets MSA/dialectal pos/neg/neut	+stemming 1: acc=74.6% 2: acc=70.2%
(Duwairi *et al.*, 2015)	SFS	MSA unigrams	4,400 tweets MSA pos/neg	+stemming F1 =55.51%
(Assiri *et al.*, 2017)	WLBA, SFS DP	Saudi/dialects size:14,000 lexicon term length, negation and supplication	1: 4,700 Saudi tweets pos/neg 2: 500 Egyptian tweets pos/neg/neut	Best: WLBA 1: acc=81% 2: acc=76%

Table 4. *Summary of lexicon-based ASA research works.*

Unlike the above-mentioned methods, which employed pre-weighted lexicons to determine the sentiment score, Assiri *et al.* (2017) introduced a polarity weighting algorithm called WLBA which assigns weights to the polarity words by learning from the data itself. This algorithm considers the polarity words' context as it explores and counts how frequently a pair of (polarity, non-polarity) words co-occurs. Later, it assigns a weight to the polarity word due to its associations' count with the non-polarity word in the whole corpus. A Saudi lexicon was built using corpus-based and dictionary-based approaches (see Section 4.1). Upon applying the model on Egyptian dataset from (El-Beltagy and Ali, 2013) and a manually-collected Saudi dataset of 4,700 tweets, results showed that WLBA achieved a poor performance compared to SFS and DP for both datasets due to ignoring complex structural and lexical specifications of the Saudi corpus. However, when features like negation and supplication were accurately handled via rule-based methods, WLBA outperformed other methods with an accuracy of 81%, compared to 72% and 43% scored by SFS and DP methods respectively. Additionally, for the Egyptian dataset, the achieved accuracy was 76%, compared to 71% and 68% scored by SFS and DP method respectively. Table 4 lists a summary of the above-mentioned lexicon-based research papers where Best, acc and F1 indicate the best method, the scored accuracy and F1-score respectively.

5.3. *Deep Learning Approaches*

Deep learning approaches are representation learning methods which learn discriminative features automatically from the data through either an unsupervised manner or via a supervised strategy, more specifically, self-supervised learning which is an instance of supervised learning whereby the training labels are determined by the input data (here document statistics such as the usage of words) (Gomez *et al.*, 2017). In addition, for a specific task like document classification, such embeddings can be learned within a neural network model trained on annotated data (Mikolov *et al.*, 2013). These methods can learn continuous and real-valued multiple levels of text representation using multi-layer nonlinear neural networks where each layer transforms the representation at one level into a representation at a higher and more abstract level (Mikolov *et al.*, 2013). The learned representations can be divided into two types:

– Word embeddings: where every word in the corpus is mapped to a real-valued low-dimensional vector in the embedding space using one of the word mapping algorithms such as word2vec (Mikolov *et al.*, 2013) and GloVe (Pennington *et al.*, 2014);

– Document embeddings: generate continuous representations of larger blocks of text such as sentences, paragraphs or whole documents using a document mapping algorithm such as doc2vec (Le and Mikolov, 2014).

Both representation types can be used as features for further classification tasks such as sentiment classification. Indeed, text embeddings features have been successfully applied in recent ASA research as they could capture the fine-grained semantic and syntactic regularities within the input text (Le and Mikolov, 2014). In addition, the automatic feature extraction, the low-dimensionality and less data sparsity of the embedding vectors have made deep learning-based SA models competitive to hand-crafted-based ones (Mikolov *et al.*, 2013).

In Altowayn and Tao (2016), the authors replaced the hand-crafted features with efficient features produced without much effort to be adopted for the sentiment analysis task. With the application of minor preprocessing, word embeddings features were used as discriminative features to train several supervised classifiers. The used embeddings were generated using Continuous bag of words (CBOW) learning algorithm (Mikolov *et al.*, 2013) and an MSA/DA training corpus of 190 million words. The authors indicated that their embeddings model could handle dialects efficiently as it mapped different writing shapes of dialectal words close to each other in the embedding space. To perform the SA task, fixed-sized embedding vectors were learned for a combination of three datasets of multi-dialectal tweets: ASTD (Nabil *et al.*, 2015), ArTwitter (Abdulla *et al.*, 2013) and QCRI (Mourad and Darwish, 2013), in addition to other two datasets representing book reviews: LABR (Aly and Atiya, 2013) and MSA news articles derived from the translated MPQA corpus (Banea *et al.*, 2010). Results showed that, for subjectivity classification of the MPQA dataset, the presented model has slightly improved the performance compared to hand-crafted features-based systems of Banea *et al.* (2010) and Mourad and Darwich (2013) where it achieved an accuracy of 77.87% and F-score of 76.14%. As for the polarity classification, best

metrics values were scored by the logistic regression algorithm with an accuracy of 81.88% and F-measure of 81.58%.

Arabic word embeddings are usually learned using large-scale training corpora so that they could cover the vocabulary of the dataset to be sentimentally classified. Thus, the learning process is considered. costly in terms of the time needed for training. This could be avoided if pretrained Arabic word embeddings were included in a neural SA model. Gridach *et al.* (2017) have investigated this idea where an ASA model was developed using word embeddings provided by Zahran *et al.* (2015) and previously trained with MSA/dialectal corpora using three word representations: Glove, SG and CBOW. These representations were examined as initializing vectors of the input words fed to a deep learning SA model built using Convolutional Neural Networks (CNNs). The proposed model CNN-ASAWR was developed as a variant of Collobert *et al.* (2011) system. The trained model was applied on two MSA/dialectal datasets: ASTD (Nabil *et al.*, 2015) and SemEval-2017 (Rosenthal *et al.*, 2017). Results showed that Arabic pre-trained word representations can be considered as universal feature extractors used for the sentiment classification task as better performances were achieved. In ASTD dataset for instance, the best F-measure scored by CNN-ASAWR was 72.14%, compared to 62.60% achieved by Nabil *et al.* (2015) while for SemEval-2017, an F-measure of 63% was achieved against 61% scored by the system of El-Beltagy *et al.* (2017) which ranked first in SemEval competition.

With the lack of lexical and semantic resources especially for under-represented Arabic dialects, paragrah embeddings represent an alternative expressive features for DA. Based on that, the authors in Mdhaffar *et al.* (2017) investigated representing Tunisian comments by distributed paragraph representations to be used as features in a Tunisian SA model. Their model was evaluated using a combination of publicly available MSA/multi-dialectal datasets: OCA (Rushdi-Saleh *et al.*, 2011), LABR (Aly and Atiya, 2013) and a manually annotated Tunisian Sentiment Analysis Corpus (TSAC) obtained from Facebook comments. Doc2vec algorithm by Le and Mikolov (2014) was applied to generate document vectors of each comment. The produced vectors were then fed SVM, Bernoulli NB (BNB) and Multilayer Perceptron (MLP) classifiers with various combinations of MSA, dialects and Tunisian used as training sets. The best results were scored by MLP classifier when TSAC corpus was solely used as a training set where it achieved an accuracy equals to 78% and an F1-score of 78%.

Each deep learning architecture has specific merits which are usually related to its building unit. Baniata and Park (2016) investigated the impact of using a combination of CNN and Bidirectional-Long Short Term Memory (BiLSTM) on SA of MSA/dialectal tweets. They relied on the fact that the phrase representation of every sentence captured by CNN can be further enhanced by using BiLSTM network which can capture the contextual information and thus yields an improved performance. Two configurations were examined: CNN-BiLSTM, which involves generating the sentence representation to be improved later by the context information derived from both direction, and BiLSTM-CNN, where contextual information is first captured then fed to CNN to assist in generating the sentence representation. The used CNN model

Paper	Embedding	Dataset	Classifier	Evaluation
(Al Sallab *et al.*, 2015)	Recursive parsing tree	LDC-ATB MSA pos/neg	DNN, DBN DAE, RAE Linear-SVM	Best: RAE acc=74.3%
(Altowayan and Tao, 2016)	word2vec (CBOW)	LDC-ATB ASTD, ArTwitter QCRI, LABR MPQA MSA/dialects pos/neg	LR, SGD GNB, RF Linear-SVM Nu-SVM	Best (MSA): Linear-SVM acc=77.87% Best (dialects): LR acc=81.88%
(Gridach *et al.*, 2017)	word2vec (skip-gram) (CBOW) Glove	ASTD SemEval 2017 MSA/dialects pos/neg/neut	CNN	F-score=72.14% (ASTD) F-score=61% (SemEval2017)
(Baniata and Park, 2016)	word2vec pretrained vectors	LABR MSA/Dialects pos/neg	CNN-BiLSTM BiLSTM-CNN	Best: CNN-BiLSTM acc=86.43%
(Mdhaffar *et al.*, 2017)	Doc2vec	113,196 tweets OCA, LABR Tunisian/dialects pos/neg	SVM MLP BNB	Best: MLP prec=78% recall=78%
(Al Sallab *et al.*, 2017)	Recursive syntactic parsing tree	Tweets QALB ATB MSA/dialects pos/neg	AROMA (modified RAE) DNN, DBN DAE-DBN, RAE NB, Linear-SVM	Best: AROMA acc=86.5%

Table 5. *Summary of Deep learning-based ASA research works.*

contained layers of filter sizes 3, 4 and 5 with the activation function ReLu used in both configurations. Both ensembles were evaluated using LABR dataset (Aly and Atiya, 2013). The data was normalized first through removing diacritics, punctuations and non-Arabic characters, and the vocabulary size was reduced by keeping words of frequency greater than 10. Word embeddings were then obtained based on pre-trained word vectors by Al-Rfou *et al.* (2013). It was noted that CNN-BiLSTM architecture achieved an accuracy of 86.43%, whereas BiLSTM–CNN architecture has suffered from of overfitting after the fifth epoch yielding an accuracy of 66.26%.

The variety of deep learning architectures has evoked the question about which architecture can perform better for ASA analysis. Therefore, Al Sallab *et al.* (2015) explored four deep learning models of different architectures and compared their performances within the context of ASA. The first three models are: Deep Neural Network (DNN), Deep Belief Network (DBN) and Deep Auto Encoders (DAE). While DNN model employs the back propagation in a conventional neural network with several

layers, DBN avoids overfitting through a pretraining phase before feeding a discriminative fine tuning step whereas DAE provides a compact representation of the input sentence with a reduced dimensionality. These models were trained using the ordinary Bag-of-Words features along with lexicon features derived from ArSenL lexicon (Badaro *et al.*, 2014). As for the fourth model, Recursive Auto Encoder (RAE), it was suggested to address the lack of context handling procedures issue found in the previous three models. RAE can parse raw sentence words in the best order for which the error of recreating the same sentence words in the same order is as minimum as possible. This is done via a recursive parse tree where the sentence words are parsed recursively till finding the best words' order. The evaluation was performed using Linguistic Data Consortium Arabic Tree Bank [9]. Upon comparing the performances of the four models in positive/negative sentiment classification against an SVM model with hand-crafted features, it was noted that the performance of DNN, DBN and DAE was close to SVM's, while DAE provided a better representation for the input sparse sentence vector. The RAE model outperformed all the other models achieving an accuracy of 74.3% and F1-score of 73.5%, compared to an accuracy of 45.2% and F1-score of 44.1% scored by linear SVM. This indicates the privilege of recursive models compared to one-shot models in terms of learning accurate semantic representations.

According to Al Sallab *et al.* (2015), the efficiency of RAE-based models was attributed to their ability to perform SA without the need for opinion resources or extensive NLP. However, standard RAE models become insufficient to handle Arabic lexical sparsity and ambiguity which limit the model's ability to generalize and causes over-fitting. These issues were addressed in Al Sallab *et al.* (2017) where A Recursive Deep Learning Model for Opinion Mining in Arabic (AROMA) was developed. To enable modeling the semantic interactions at the morpheme level and to reduce the lexical sparsity and ambiguity, the training data was subjected to morphological tokenization using MADAMIRA (Pasha *et al.*, 2014) before it was fed to AROMA. In addition, semantic embedding with/without unsupervised pre-training alongside sentiment embedding were used to provide improved word distributed representations. Furthermore, instead of using the greedy algorithm to define the order of the model's recursion, AROMA employed phrase structures to automatically generate syntactic parse trees with which a better modeling of composition was achieved. The presented model was evaluated using three datasets annotated for pos/neg polarities: an MSA dataset from Abdul-Mageed *et al.* (2011) called ATB, dialectal Tweets dataset by Refaee and Reeser (2014) an MSA/DA comments derived from Farra *et al.* (2015) and referred to as QALB. The experiments involved using different combinations of the contributions augmented to the standard RAE. The results indicated that compared to the standard RAE, AROMA with all the contributions combined could improve the classification accuracy significantly by 12.2%, 8.4% and 7.2% for the ATB, QALB and Tweets datasets, respectively. Moreover, AROMA was evaluated against several ML and DL models where it overcome all of them as it scored an accuracy increment of 7.3%, 1.7% and 7.6% for the same previous datasets respectively.

9. http://www.ldc.upenn.edu/Catalog/catalogEntry.jsp?catalogId=2005T20.

6. Discussion and Conclusion

The complex morphological nature of the Arabic language along with the wide usage of dialects require a careful design of the SA models such that inflected words, various writing styles, typos and varying grammatical nature are handled efficiently. Considering these issues, supervised learning-based methods adopted special pre-processing procedures such as stemming, lemmatization and tokenization. Another hand-crafted features such as n-grams and bag-of-words in addition to syntactic and stylistic features were used by these models. With such a variety of features, feature vectors tend to be of high dimensionality and sparsity which may drown the classifier with noisy features or lead to memory issues as it has been reported in Duwairi *et al.* (2014). To reduce the features' size, feature selection methods like TF-IDF weighting with a specific threshold was adopted as in Shoukry and Rafea (2012) and Al-Osaimi and Badruddin (2014). In the same context, some works suggested reducing the text size through stemming or stopwords removel. However, it has been proved that stemming has no impact on the classification performance, especially for DA since available stemming tools mostly target MSA (Duwairi *et al.*, 2014). Moreover, with the lack of reliable stopwords lists for dialects, keeping stopwords was proved to be better than eliminating them as they can assist in capturing the sentiment (Shoukry and Rafea, 2012). Supervised SA systems are generally robust and accurate. However, performances may vary from one system to another due to the used classifier. It has been noted that SVM usually outperforms other classifiers (Shoukry and Rafea, 2012; Salamah and Elkhlifi, 2016); this can be attributed to the fact that SVM can efficiently handle feature vectors of high dimensions and sparsity through its overfitting protection property.

Supervised SA models provide an accurate reliable performance, yet the labor-intensive task of preparing a sentimentally annotated corpus along with the training overhead and memory issues are important disadvantages that cannot be ignored. In contrast, lexicon-based SA models are easy to design since they do not need a labeled input data. Moreover, the training overhead is avoided by using a sentiment lexicon that acts as a rule-based classifier. However, one major drawback of these models is that they are not aware of language subtleties such as sarcasm, negations, etc. Because uniform weight scheme lexicons ignore the contextual-related information since a sentence's polarity is recognized by the polarity scores of its constituent words. To enhance the SA performance of these methods, new weighting schemes were developed based on the word's co-occurrence information (context) as in El-Beltagy and Ali (2013) and Assiri *et al.* (2017). On the other hand, the used lexicons also suffer from low-coverage and Out Of Vocabulary (OOV) issues especially for dialects which degrades the classification performance remarkably. Increasing the lexicon-coverage has been tackled in Abdulla *et al.* (2014) and Duwairi *et al.* (2015) through creating a large-sized lexicon initially constructed using seed words derived from publicly available lexicons or from the corpus itself then enriched with stems, synonyms and dialectal terms. Nevertheless, these solutions were insufficient to overcome the lexi-

con's dialect- and domain-dependency problems unless a very large-sized lexicon is built which is considered a difficult task.

To exploit the merits of the two previous methods, hybrid models have emerged. Lexicon-derived features are combined with linguistic ones to obtain a better sentiment classification performance. What makes these models better than both supervised and lexicon-based models is its ability to involve external semantic resources and datasets as in Salameh *et al.* (2015) and Baly *et al.* (2017) Furthermore, the wide variety of the hand-crafted features provide a coherent representation of the contextual information (Abdul-Mageed *et al.*, 2014; Al-Moslmi *et al.*, 2017).

The good performance of hybrid methods has been achieved at the cost of the laborious tasks of designing the features and building the lexicons. Deep learning-based methods alleviate such efforts through learning the features automatically from the data itself using deep neural networks. Text embedding features such as word/document embeddings generated via word2vec/doc2vec methods have proved their efficiency for SA when they were used to train ML classifiers (Altowayan and Tao, 2016; Mdhaffar *et al.*, 2017). Moreover, a better classification performance can be obtained if various architectures of deep neural networks, whose units adopt the compositional manner to represent the input text, are used to design the classifier as in Baniata and Park (2016), Al Sallab *et al.* (2015) and Al Sallab *et al.* (2017). It is obvious that DL methods are superior to traditional ML methods in terms of SA performance and features extraction cost. Nevertheless, SA using deep neural networks architectures requires more training time (Joulin *et al.*, 2016). This can be handled by applying an appropriate tuning of hyperparameters in addition to specific preprocessing and postprocessing procedures.

The previous comparison analysis of methods could be supported by tracking the classification performance of a specific dataset using traditional ML, rule-based and DL methods. For instance, given the Jordanian ArTwitter dataset, it could be noted that performing SA using a lexicon-based method (Abdulla *et al.*, 2014) resulted in a classification accuracy of 70%, while adopting distributed representations extracted via word2vec as in Altowayan and Tao (2016) has increased the accuracy by 11.88%. On the other hand, within the same category of methods such as the DL methods, it could be deduced that the effective handling of the special properties of the Arabic language has a positive impact on the SA performance as it can be seen in Al Sallab *et al.* (2015) and Al Sallab *et al.* (2017) where the accuracy improved from 74.3% using standard RAE to 86.5% when this model was equipped with Arabic-specific modifications.

The development of ASA models has involved the provision of annotated corpora (see Table 6), semantic resources and pretrained word vectors. This enriched the repository of NLP Arabic tools and resources. Regarding the research reviewed in this paper, most of the proposed datasets were of informal dialectal content as they were harvested from social media platforms. For single-dialect datasets such as Saudi (Assiri *et al.*, 2017), Kuwaiti (Salamah and Elkhlifi, 2016) or Jordanian (Duwairi *et al.*, 2014), they are rarely reused by other studies. However, pure Egyptian or

Paper	Dataset name/type	Size	Arabic Variant	Polarity	Publicly Available
(Shoukry and Rafea, 2012)	tweets	1,000	Egyptian	Pos/neg	No
(Al-Osaimi and Badruddin, 2014)	tweets	3,000	DA	pos/neg /neut	No
(Salamah and Elkhlifi, 2016)	tweets	340,000	Kuwaiti	pos/neg	No
(Abdul-Mageed *et al.*, 2014)	TGRD THR MONT DAR	3,015 3,008 3,097 2,798	MSA & DA	pos/neg /neut	No
(Salameh *et al.*, 2015)	Syr BBN	2,000 1,200	Syrian Levantine	pos/neg /neut	Yes Yes
(Al-Moslmi *et al.*, 2017)	reviews	8,861	DA	pos/neg	Yes
(El-Beltagy and Ali, 2013)	tweets comments	100 500	Egyptian	pos/neg /neut	No
(Abdulla *et al.*, 2014)	tweets comments	2,000 2,400	MSA/Jordanian MSA/DA	pos/neg pos/neg	Yes No
(Duwairi *et al.*, 2015)	tweets	4,400	MSA	pos/neg	No
(Assiri *et al.*, 2017)	tweets	4,700	Saudi	pos/neg	No
(Mdhaffar *et al.*, 2017)	TSAC	16,970	Tunisian	pos/neg	Yes

Table 6. *The datasets constructed and/or evaluated within the reviewed ASA studies.*

Egyptian-dominated datasets as El-Beltagy and Ali (2013), QCRI (Mourad and Darwish, 2013) and ASTD (Nabil *et al.*, 2015) have been used as a baseline in many other studies. This is due to the fact that Egyptian dialect forms the majority of the textual content on social media which makes it a preferable dialect to investigate by many research works. Yet, recent studies have shed light on other dialects spoken by the "Arab spring" countries such as Syrian and Tunisian (Salameh *et al.*, 2015; Mdhaffar *et al.*, 2017). Regarding MSA/multi-dialectal datasets such as Rushdi-Saleh *et al.* (2011), Aly and Atiya (2013), Abdulla *et al.* (2014), Abdul-Mageed *et al.* (2014) and Al-Moslmi *et al.* (2017), they are widely reused especially that modern ASA systems are designed with the objective of being dialect/domain-independent systems. On the other hand, the presented ASA systems have provided several Arabic sentiment lexicons. Most of which support MSA/multi-dialects such as lexicons in Abdulla *et al.* (2014), Abdul-Mageed *et al.* (2014) and Al-Moslmi *et al.* (2017) or Egyptian (El-Beltagy and Ali, 2013) which is publicly available.

Finally, some of the reviewed deep learning-based models produced Arabic word vectors trained either on MSA/multi-dialectal corpora as in Al-Rfou *et al.* (2013) and Altowayan and Tao (2016) or with Tunisian corpus (Mdhaffar *et al.*, 2017) or using an Egyptian corpus as in Zahran *et al.* (2015). According to Gridach *et al.* (2017), involving pretrained word vectors in a deep learning model enhances the quality of the model's embeddings vectors and thus improves further classification tasks. Based on

that fact, the performance of Arabic deep learning-based SA models can be improved by exploiting pretrained Arabic word vectors trained on external corpora (Baniata and Park, 2016; Gridach *et al.*, 2017). This can be facilitated if the produced Arabic word vectors were made publicly available as those from Al-Rfou *et al.* (2013).

All the presented studies tried to address one or more challenging issues of ASA. Although valuable efforts were spent, there is a lot to do towards developing new tools and resources able to support MSA and DA more efficiently. For instance, Named Entities (NEs), especially person names, either in MSA or DA form a dilemma to any ASA system as they might considered as an adjective of a specific sentiment. Instead of excluding NEs (El-Beltagy and Ali, 2013; Duwairi *et al.*, 2014), they could be used as sentiment indicatives through assigning polarities to them. Thus, the polarity of a sentence could be predicted given the polarity of an NE contained in it. This idea is applicable for data collected during a short period of time in which opinions towards an NE are rather fixed. For instance, the location name "حلب" referring to Aleppo city in Syria has been often mentioned within negative contexts during December 2016 when Eastern Aleppo was under siege. Another difficult and interesting topic to investigate is SA of DA in terms of providing a universal system through which dialects' variances are ignored and common words between dialects along with their synonymous relations are considered. One possible way to perform that is by using word/phrase embeddings composed using syntactic-ignorant compositional functions and learned within a deep neural model.

7. References

Abdul-Mageed M., Subjectivity and sentiment analysis of Arabic as a morophologically-rich language, PhD thesis, Indiana University, 2015.

Abdul-Mageed M., Diab M., Korayem M., "Subjectivity and sentiment analysis of modern standard Arabic", *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2*, Association for Computational Linguistics, p. 587-591, 2011.

Abdul-Mageed M., Diab M., Kübler S., "SAMAR: Subjectivity and sentiment analysis for Arabic social media", *Computer Speech & Language*, vol. 28, n° 1, p. 20-37, 2014.

Abdulla N. A., Ahmed N. A., Shehab M. A., Al-Ayyoub M., "Arabic sentiment analysis: Lexicon-based and corpus-based", *2013 IEEE Jordan Conference on Applied Electrical Engineering and Computing Technologies (AEECT)*, IEEE, p. 1-6, 2013.

Abdulla N., Mohammed S., Al-Ayyoub M., Al-Kabi M. *et al.*, "Automatic lexicon construction for arabic sentiment analysis", *2014 International Conference on Future Internet of Things and Cloud (FiCloud)*, IEEE, p. 547-552, 2014.

Al-Kabi M. N., Abdulla N. A., Al-Ayyoub M., "An analytical study of arabic sentiments: Maktoob case study", *8th International Conference for Internet Technology and Secured Transactions (ICITST)*, IEEE, p. 89-94, 2013.

Al-Moslmi T., Albared M., Al-Shabi A., Omar N., Abdullah S., "Arabic senti-lexicon: Constructing publicly available language resources for Arabic sentiment analysis", *Journal of Information Science*, 2017.

Al-Osaimi S., Badruddin K. M., "Role of Emotion icons in Sentiment classification of Arabic Tweets", *Proceedings of the 6th international conference on management of emergent digital ecosystems*, ACM, p. 167-171, 2014.

Al-Rfou R., Perozzi B., Skiena S., "Polyglot: Distributed word representations for multilingual nlp", *arXiv preprint arXiv:1307.1662*, 2013.

Al Sallab A. A., Baly R., Badaro G., Hajj H., El Hajj W., Shaban K. B., "Deep learning models for sentiment analysis in arabic", *ANLP Workshop*, vol. 9, 2015.

Al Sallab A. A., Baly R., Hajj H., Shaban K. B., El-Hajj W., Badaro G., "AROMA: A Recursive Deep Learning Model for Opinion Mining in Arabic As a Low Resource Language", *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, vol. 16, n° 4, p. 25:1-25:20, 2017.

Al-Twairesh N., Al-Khalifa H. S., Alsalman A., "AraSenTi: Large-Scale Twitter-Specific Arabic Sentiment Lexicons", *ACL (1)*, 2016.

Altowayan A. A., Tao L., "Word embeddings for Arabic sentiment analysis", *2016 IEEE International Conference on Big Data (Big Data)*, IEEE, p. 3820-3825, 2016.

Aly M. A., Atiya A. F., "LABR: A Large Scale Arabic Book Reviews Dataset.", *ACL (2)*, p. 494-498, 2013.

Assiri A., Emam A., Al-Dossari H., "Towards enhancement of a lexicon-based approach for Saudi dialect sentiment analysis", *Journal of Information Science*, 2017.

Assiri A., Emam A., Aldossari H., "Arabic sentiment analysis: a survey", *International Journal of Advanced Computer Science and Applications*, vol. 6, n° 12, p. 75-85, 2015.

Badaro G., Baly R., Hajj H., Habash N., El-Hajj W., "A large scale Arabic sentiment lexicon for Arabic opinion mining", *ANLP 2014*, 2014.

Balikas G., Amini M.-R., "TwiSE at SemEval-2016 Task 4: Twitter Sentiment Classification", *arXiv preprint arXiv:1606.04351*, 2016.

Baly R., Badaro G., Hamdi A., Moukalled R., Aoun R., El-Khoury G., Al Sallab A. A., Hajj H., Habash N., Shaban K., El-Hajj W., "OMAM at SemEval-2017 Task 4: Evaluation of English State-of-the-Art Sentiment Analysis Models for Arabic and a New Topic-based Model", *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, Association for Computational Linguistics, Vancouver, Canada, p. 603-610, August, 2017.

Banea C., Mihalcea R., Wiebe J., "Multilingual subjectivity: Are more languages better?", *Proceedings of the 23rd international conference on computational linguistics*, Association for Computational Linguistics, p. 28-36, 2010.

Baniata L. H., Park S.-B., "Sentence Representation Network for Arabic Sentiment Analysis", *Proceedings of the Korean Information Science Society*, 2016.

Biltawi M., Etaiwi W., Tedmori S., Hudaib A., Awajan A., "Sentiment classification techniques for Arabic language: A survey", *7th International Conference on Information and Communication Systems (ICICS)*, IEEE, p. 339-346, 2016.

Collobert R., Weston J., Bottou L., Karlen M., Kavukcuoglu K., Kuksa P., "Natural language processing (almost) from scratch", *Journal of Machine Learning Research*, vol. 12, p. 2493-2537, 2011.

Diab M., "Second generation AMIRA tools for Arabic processing: Fast and robust tokenization, POS tagging, and base phrase chunking", *2nd International Conference on Arabic Language Resources and Tools*, vol. 110, 2009.

Duwairi R., Ahmed N. A., Al-Rifai S. Y., "Detecting sentiment embedded in Arabic social media–a lexicon-based approach", *Journal of Intelligent & Fuzzy Systems*, vol. 29, n° 1, p. 107-117, 2015.

Duwairi R., Marji R., Sha'ban N., Rushaidat S., "Sentiment analysis in arabic tweets", *5th international conference on Information and communication systems (icics)*, IEEE, p. 1-6, 2014.

El-Beltagy S. R., Ali A., "Open issues in the sentiment analysis of Arabic social media: A case study", *9th international conference onInnovations in information technology (iit)*, IEEE, p. 215-220, 2013.

El-Beltagy S. R., El kalamawy M., Soliman A. B., "NileTMRG at SemEval-2017 Task 4: Arabic Sentiment Analysis", *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, Association for Computational Linguistics, Vancouver, Canada, p. 790-795, August, 2017.

Farra N., McKeown K., "SMARTies: Sentiment Models for Arabic Target entities", *arXiv preprint arXiv:1701.03434*, 2017.

Farra N., McKeown K., Habash N., "Annotating targets of opinions in arabic using crowdsourcing", *Proceedings of the Second Workshop on Arabic Natural Language Processing*, p. 89-98, 2015.

Gomez L., Patel Y., Rusiñol M., Karatzas D., Jawahar C., "Self-supervised learning of visual features through embedding images into text topic spaces", *arXiv preprint arXiv:1705.08631*, 2017.

Gridach M., Haddad H., Mulki H., "Empirical Evaluation of Word Representations on Arabic Sentiment Analysis", *International Conference on Arabic Language Processing*, Springer, p. 147-158, 2017.

Habash N. Y., "Introduction to Arabic natural language processing", *Synthesis Lectures on Human Language Technologies*, vol. 3, n° 1, p. 1-187, 2010.

Joulin A., Grave E., Bojanowski P., Mikolov T., "Bag of tricks for efficient text classification", *arXiv preprint arXiv:1607.01759*, 2016.

Kiritchenko S., Zhu X., Mohammad S. M., "Sentiment analysis of short informal texts", *Journal of Artificial Intelligence Research*, vol. 50, p. 723-762, 2014.

Le Q., Mikolov T., "Distributed representations of sentences and documents", *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, p. 1188-1196, 2014.

Liu B., "Sentiment analysis and opinion mining", *Synthesis lectures on human language technologies*, vol. 5, n° 1, p. 1-167, 2012.

Maamouri M., Graff D., Bouziri B., Krouna S., Bies A., Kulick S., "Standard Arabic morphological analyzer (SAMA) version 3.1", *Linguistic Data Consortium, Catalog No.: LDC2010L01*, 2010.

Mdhaffar S., Bougares F., Esteve Y., Hadrich-Belguith L., "Sentiment Analysis of Tunisian Dialect: Linguistic Resources and Experiments", *WANLP 2017 (*co-located with *EACL 2017)*, 2017.

Mikolov T., Sutskever I., Chen K., Corrado G. S., Dean J., "Distributed representations of words and phrases and their compositionality", *Advances in neural information processing systems*, p. 3111-3119, 2013.

Mohammad S. M., Kiritchenko S., Zhu X., "NRC-Canada: Building the state-of-the-art in sentiment analysis of tweets", *arXiv preprint arXiv:1308.6242*, 2013.

Mohammad S. M., Salameh M., Kiritchenko S., "How Translation Alters Sentiment.", *J. Artif. Intell. Res. (JAIR)*, vol. 55, p. 95-130, 2016.

Mourad A., Darwish K., "Subjectivity and Sentiment Analysis of Modern Standard Arabic and Arabic Microblogs", *WASSA@ NAACL-HLT*, p. 55-64, 2013.

Nabil M., Aly M. A., Atiya A. F., "ASTD: Arabic Sentiment Tweets Dataset", *EMNLP*, p. 2515-2519, 2015.

Pasha A., Al-Badrashiny M., Diab M. T., El Kholy A., Eskander R., Habash N., Pooleery M., Rambow O., Roth R., "MADAMIRA: A Fast, Comprehensive Tool for Morphological Analysis and Disambiguation of Arabic", *LREC*, vol. 14, p. 1094-1101, 2014.

Pennington J., Socher R., Manning C. D., "Glove: Global vectors for word representation", *EMNLP*, vol. 14, p. 1532-1543, 2014.

Piryani R., Madhavi D., Singh V. K., "Analytical mapping of opinion mining and sentiment analysis research during 2000–2015", *Information Processing & Management*, vol. 53, n° 1, p. 122-150, 2017.

Popescu A.-M., Etzioni O., "Extracting product features and opinions from reviews", *Natural language processing and text mining*, Springer, p. 9-28, 2007.

Refaee E., Rieser V., "An Arabic Twitter Corpus for Subjectivity and Sentiment Analysis", *LREC*, p. 2268-2273, 2014.

Reyes A., Rosso P., "On the difficulty of automatically detecting irony: beyond a simple case of negation", *Knowledge and Information Systems*, vol. 40, n° 3, p. 595-614, 2014.

Rosenthal S., Farra N., Nakov P., "SemEval-2017 Task 4: Sentiment Analysis in Twitter", *Proceedings of the 11th International Workshop on Semantic Evaluation*, SemEval '17, Association for Computational Linguistics, Vancouver, Canada, August, 2017.

Rushdi-Saleh M., Martín-Valdivia M. T., Ureña-López L. A., Perea-Ortega J. M., "OCA: Opinion corpus for Arabic", *Journal of the Association for Information Science and Technology*, vol. 62, n° 10, p. 2045-2054, 2011.

Salamah J. B., Elkhlifi A., "Microblogging opinion mining approach for kuwaiti dialect", *Computing Technology and Information Management*, vol. 1, n° 1, p. 9, 2016.

Salameh M., Mohammad S., Kiritchenko S., "Sentiment after Translation: A Case-Study on Arabic Social Media Posts", *HLT-NAACL*, p. 767-777, 2015.

Shoukry A., Rafea A., "Sentence-level Arabic sentiment analysis", *2012 International Conference on Collaboration Technologies and Systems (CTS)*, IEEE, p. 546-550, 2012.

Turney P. D., "Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews", *Proceedings of the 40th annual meeting on association for computational linguistics*, Association for Computational Linguistics, p. 417-424, 2002.

Zaghouani W., Habash N., Mohit B., The Qatar Arabic language bank guidelines, Technical report, Technical Report CMU-CS-QTR-124, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, September, 2014.

Zahran M. A., Magooda A., Mahgoub A. Y., Raafat H. M., Rashwan M., Atyia A., "Word Representations in Vector Space and their Applications for Arabic", *CICLing (1)*, p. 430-443, 2015.

Une approche fondée sur les lexiques d'analyse de sentiments du dialecte algérien

Imane Guellil[*,**]**— Faical Azouaou**[*]**— Houda Saâdane**[***]**— Nasredine Semmar**[****]

** Laboratoire des Méthodes de Conception des Systèmes. École nationale Supérieure d'Informatique, BP 68M, 16309, Oued-Smar, Alger, Algérie. http://www.esi.dz*
*** École Supérieure des Sciences Appliquées d'Alger ESSA-Alger*
**** GEOLSemantics, 12 Avenue Raspail, 94250 Gentilly, France*
***** CEA, LIST, Laboratoire Vision et Ingénierie des Contenus, 91191 Gif-sur-Yvette, France*
i_guellil@esi.dz ; i.guellil@essa-alger.dz ; f_azouaou@esi.dz ; houda.saadane@geolsemantics.com ; nasredine.semmar@cea.fr

RÉSUMÉ. *La plupart des outils d'analyse de sentiments traitent essentiellement l'arabe standard moderne (ASM), et peu d'entre eux ne prennent en considération les dialectes. À notre connaissance, aucun outil en libre accès n'est disponible concernant l'analyse de sentiments de textes écrits en dialecte algérien. Cet article présente un outil d'analyse de sentiments des messages écrits en dialecte algérien. Cet outil est fondé sur une approche combinant l'utilisation de lexiques ainsi qu'un traitement spécifique de l'agglutination. Nous avons évalué notre approche en utilisant deux lexiques annotés en sentiments et un corpus de test contenant 749 messages. Les résultats obtenus sont encourageants et montrent une amélioration continue après l'exécution de chaque étape de notre approche.*

ABSTRACT. *Most of the sentiment analysis tools process only Modern Standard Arabic (MSA). Indeed, few dialects are considered by the actual tools, in particular Algerian dialect where we do not identify any free tool carrying texts of this dialect. In this article we present a tool for sentiment analysis of messages written in Algerian dialect. This tool is based on an approach which uses both lexicons and specific treatment of agglutination. This approach was experimented using two sentiment lexicons and a test corpus containing 749 messages. The obtained results were encouraging and showing continuous improvement after each step of the considered approach.*

MOTS-CLÉS : *analyse de sentiments, dialecte algérien, lexique de sentiments, agglutination.*

KEYWORDS: *sentiment analysis, algerian dialect, sentiment lexicon, agglutination.*

TAL. Volume 58 – n°3/2017, pages 41 à 65

1. Introduction

L'arabe est une langue riche et complexe utilisée par plus de 400 millions de locuteurs dans le monde (Siddiqui *et al.*, 2016). Cette langue est cependant dans un état de diglossie[1] dans les pays où elle est utilisée car elle coexiste avec vingts-deux dialectes (Sadat *et al.*, 2014). L'intérêt porté à l'arabe et ses dialectes a largement augmenté au cours de ces dernières années. Un intérêt principalement dû à la proportion des locuteurs arabes au sein des médias sociaux. Au cours de la dernière décennie, plusieurs travaux ont été menés sur l'arabe et ses dialectes. Une revue des méthodes et résultats pour le traitement de l'arabe dialectal a été réalisée par Shoufan et Alameri (2015). Quatre types de tâches y sont présentés : 1) l'analyse basique, 2) la construction de ressources, 3) l'identification du dialecte utilisé, et 4) l'analyse sémantique.

La richesse des médias sociaux en termes d'opinions, d'émotions et de sentiments a suscité l'intérêt de la communauté de recherche à se pencher beaucoup plus sur les problématiques reliées à l'analyse sémantique et plus particulièrement sur l'analyse de sentiments de l'arabe et ses dialectes. L'analyse de sentiments consiste à déterminer la valence (positive, négative ou neutre) d'un message donné. Plusieurs approches d'analyse ou de classification de sentiments ont vu le jour : 1) *l'approche supervisée :* utilisant les techniques d'apprentissage, elle est fondée sur les corpus annotés, 2) *l'approche non supervisée :* fondée sur l'existence d'un lexique de sentiments contenant un ensemble de termes, leur valence et leur intensité pouvant aller de -1 à $+1$ ou encore de -5 à $+5$, et 3) *l'approche hybride :* combinant les deux approches précédentes. Toutes ces approches ont été adoptées dans le cas de l'arabe et de ses dialectes. Cette adoption, ou adaptation, véhicule des problèmes liés à l'approche à laquelle sont ajoutés les problématiques de l'arabe et ses dialectes. Une présentation de ces problèmes est donnée dans (Guellil et Boukhalfa, 2015). Ces problèmes sont principalement liés au manque de corpus annotés et aux problèmes quantitatif et qualitatif des lexiques de sentiments.

Ces problématiques sont recensées dans la littérature comme suit : 1) les problématiques orthographiques liées à la diacritisation ainsi qu'aux différentes manières d'écrire chaque lettre arabe, et 2) les problématiques morphologiques liées à la dérivation, la flexion et l'agglutination. En plus de ces problématiques, chaque dialecte rajoute des problématiques supplémentaires. Prenons par exemple le dialecte algérien (qui est l'objet de ce travail), il s'agit d'un dialecte maghrébin utilisé par plus de 40 millions de locuteurs (ce qui représente 10 % de la population s'exprimant en arabe). Il souffre cependant d'un manque considérable de travaux, d'outils et de ressources. En plus des problématiques de ce dialecte partagées avec l'arabe standard, il dispose d'une très forte richesse du vocabulaire provenant de plusieurs autres langues. Dans (Meftouh *et al.*, 2012), les auteurs illustrent le fait que le dialecte algérien est composé de 65 % d'arabe, de 19 % de français et de 16 % de turque et de

1. Situation où sont en usage deux langues apparentées génétiquement et structurellement et dont les distributions fonctionnelles sont complémentaires (Fishman, 1967).

berbère. Ce dialecte dispose également d'une très grande richesse morphologique, par exemple, l'agglutination aux mots des pronoms personnels et les pronoms de compléments d'objet direct (COD), que nous trouvons aussi dans l'arabe standard, est étendue aux pronoms de compléments d'objets indirect (COI) ainsi que la négation.

Dans ce travail, nous présentons et implémentons une approche d'analyse de sentiments (AS) du dialecte algérien (DALG) combinant l'utilisation de lexiques de sentiments et le traitement de l'agglutination. Cette dernière se compose de deux principales étapes : 1) construction d'un lexique de sentiments, et 2) calcul de la valence et intensité du sentiment d'un message donné. La première étape consiste à construire un lexique de sentiments en dialecte algérien en disposant d'un lexique en anglais. Dans la seconde étape, nous procédons à tous les traitements morphologiques nécessaires dans le cadre de l'analyse de sentiments. Pour évaluer notre approche, nous utilisons deux lexiques de sentiments en anglais : SentiWordNet[2] et SOCAL[3]. Nous construisons par la suite SentiALG et SOCALALG (représentant la version algérienne de ces lexiques). Nous utilisons deux corpus de test : 1) une partie annotée du corpus multidialectal PADIC (Meftouh *et al.*, 2015), et 2) une partie contenant des messages (posts et commentaires) extraits du média social Facebook.

La suite de l'article est organisée comme suit : nous présentons d'abord dans la section 2 les principales caractéristiques de l'arabe et de ses dialectes, ensuite nous exposons dans la section 3 les principaux travaux connexes menés sur l'analyse de sentiments de cette langue et de ses dialectes. La section 4 décrit l'approche que nous proposons pour l'analyse. Nous consacrons la section 5 aux expériences menées ainsi qu'à la présentation des résultats obtenus. La section 6 conclut notre étude et présente nos travaux futurs.

2. Spécificités de l'arabe et ses dialectes : zoom sur le dialecte algérien

La langue arabe est une des langues les plus parlées et utilisées dans le monde. Elle est la langue officielle de plus de vingt-deux pays parlée par plus de 400 millions de locuteurs et elle est utilisée comme vecteur de transmission religieux pour tous les musulmans au nombre de un milliard et demi (Saâdane, 2015) à travers le monde. Elle constitue ainsi un élément principal dans la culture et la pensée d'une partie importante de l'humanité et du patrimoine mondial (Saâdane *et al.*, 2013). Elle est également la quatrième langue la plus utilisée d'Internet (Siddiqui *et al.*, 2016). L'arabe dispose de trois principales variétés qui coexistent côte à côte à savoir : 1) l'arabe classique utilisé dans le Coran, livre sacré des musulmans, 2) l'arabe standard moderne (ASM) utilisé par les locuteurs arabes instruits dans leurs écrits et dans les conversations formelles à savoir dans le système éducatif et littéraire. La plupart des travaux de recherche s'appuient sur cette variante, 3) l'arabe dialectal qui constitue le moyen de

2. http ://sentiwordnet.isti.cnr.it/
3. https ://github.com/sfu-discourse-lab/SO-CAL

communication de la vie quotidienne, employé dans les conversations informelles, interviews et la littérature orale.

2.1. *Spécificités de l'ASM*

L'ASM est classé dans le groupe des langues sémitiques contemporaines qui s'écrit de droite à gauche. L'alphabet arabe contient vingt-huit lettres dont vingt-cinq consonnes et trois voyelles. En plus des voyelles, l'arabe utilise également des marques diacritiques correspondant à des voyelles courtes. Prenons par exemple la lettre ب qui se prononce (b). Si nous mettons au-dessus de cette lettre la diacritique fatha, la lettre devient : بَ se prononçant : 'ba'. Si nous mettons la même diacritique mais au-dessous de la lettre (correspondant à la kasra), la lettre devient بِ et se prononce 'bi'. Mise à part la diacritisation, les lettres arabes ont en général quatre manières de s'écrire : 1) au début du mot, elle s'écrit بـ, prenons l'exemple du mot بئر *biŷr* [4] 'un puits', 2) au milieu du mot, par exemple ـبـ(حقيبة *Haqiybaħ* 'une valise'), 3) à la fin du mot en étant attachée à lettres précédente, par exemple ـب (قلب *qalb* 'un cœur'), et 4) à la fin du mot sans être attachée à la lettre précédente, par exemple ب (باب *baAb* 'une porte').

Un mot en ASM peut également présenter plusieurs aspects morphologiques dont la dérivation, la flexion et l'agglutination. La dérivation consiste à représenter chaque mot sous la forme de « lemme-schéma ». Par exemple, les trois lettres « ktb » est un lemme lié à « l'écriture ». Dans les schémas que nous utilisons, les lettres du lemme sont remplacées par les chiffres 1, 2 et 3 dans l'ordre. Si nous appliquons par exemple le schéma « 1a2a3a » au lemme, nous obtenons le mot كَتَبَ *kataba* 'Il a écrit'. La flexion représente les différentes variations grammaticales d'un mot pouvant être reliées à sa conjugaison, son passage au féminin, pluriel, etc. Le lemme « ktb » est donc conjugué au présent de la sorte : أَكْتُبُ *Âktubu* 'j'écris', تَكْتُبِين *taktubiyn* 'tu écris'/ féminin, etc. Ce verbe est conjugué différemment au passé. Par exemple la traduction de 'j'ai écrit' en arabe est كَتَبْتُ *katabtu*. L'agglutination consiste à rassembler un ensemble de mots, de pronoms (préfixes et suffixes) et de clitiques entre eux. Par exemple la forme agglutinée سيكتبونها *sayaktubuwnahaA* 'ils l'écriront' peut être séparée de la sorte : س+ ي + كتب + ون + ها. Le lemme étant كتب *(ktb)*. La lettre س étant la marque du futur. Les deux lettres وني séparées par le lemme représentent le pronom personnel : « Ils ». Le pronom ها représente le complément d'objet direct (COD). Il est à signaler que les spécificités que nous venons de décrire pour l'ASM

4. Translittération arabe présentée dans schème Habash-Soudi-Buckwalter (HSB) (Habash *et al.*, 2007).

sont aussi présentes dans ses dialectes. Les différences entre l'ASM et ses dialectes résident principalement dans 1) la richesse du vocabulaire des dialectes par rapport à l'ASM et 2) le changement des affixes (préfixes et suffixes) utilisés dans les dialectes. Pour illustrer ces différences, nous nous penchons sur le DALG. Ce dialecte souffre d'un manque considérable de ressources, d'outils et de travaux de recherche le traitant en le comparant aux autres dialectes arabes.

2.2. *Spécificités du dialecte algérien*

Le DALG est utilisé principalement pour la communication orale de tous les jours (dans la vie quotidienne, les séries télévisées en Algérie, etc.). Il n'est pas enseigné dans les écoles, et reste absent des communications écrites officielles. Néanmoins ces dernières années, ce dialecte prend une place plus importante à l'écrit avec les médias sociaux (Harrat *et al.*, 2017). Comme nous avons exposé les principales caractéristiques orthographiques et morphologiques du DALG partagées avec l'ASM au sein de la section 2.1, nous nous concentrons dans cette partie sur les caractéristiques propres au DALG.

2.2.1. *Spécificités orthographiques du dialecte algérien*

Le DALG fait appel à toutes les voyelles et consonnes utilisées par l'ASM. En plus de ces dernières, il fait appel aux trois lettres ڤ، پ, ڤ, se prononçant respectivement p, g, v (Meftouh *et al.*, 2015). Le DALG est enrichi par les langues des groupes ayant colonisé ou géré la population algérienne au cours de l'histoire du pays. Parmi les langues de ces groupes, citons le turc, l'espagnol, l'italien et plus récemment le français (Saâdane et Habash, 2015 ; Saâdane, 2015 ; Meftouh *et al.*, 2012). De ce fait, au sein du DALG, nous trouvons des mots tels que سلم *sallam* 'saluer' et ayant comme origine l'ASM (Harrat *et al.*, 2017). Nous pouvons également trouver un mot comme فرملي *Farmliy* 'infirmier' étant originaire du français, ou encore le mot بابور *baAbuwr* 'bateau', issu du turc, شلاغم *šlaAγam* 'moustaches' emprunté du berbère, زبلة *zablaħ* 'faute', issu de la langue italienne et سيمانة *siymaAnaħ* 'une semaine' emprunté de l'espagnol.

2.2.2. *Spécificités morphologiques du dialecte algérien*

Nous abordons au cours de cette partie quatre aspects importants reliés à la morphologie (agglutination) du DALG, à savoir : 1) la conjugaison, 2) la négation, 3) les noms et adjectifs, et 4) les compléments d'objet direct (COD) et les compléments d'objet indirect (COI).

2.2.2.1. Conjugaison en dialecte algérien

Comme au sein de n'importe quel langage, la conjugaison inclut l'ajout d'un ensemble de préfixes et de suffixes à un lemme donné. Ces affixes varient selon le

Verbe	Présent	Traduction	Passé	Traduction	Impératif	Traduction
حب	نحب	J'aime	حبيت	J'ai aimé	-------	-------
	تحب	Tu aimes	حبيت	Tu as aimé	حب	Aime
	تحبي	Tu aimes	حبيتي	Tu as aimé	حبي	Aime
	يحب	Il aime	حب	Il a aimé	-------	-------
	تحب	Elle aime	حبت	Elle a aimé	-------	-------
	نحبو	Nous aimons	حبينا	Nous avons aimé	نحبو	Aimons
	تحبو	Vous aimez	حبيتو	Vous avez aimé	حبو	Aimez
	يحبو	Ils aiment	حبو	Ils ont aimé	-------	-------

Tableau 1. *Conjugaison du verbe 'aimer' au présent, passé et impératif*

pronom utilisé. Ils sont généralement les mêmes pour tous les verbes. Nous faisons figurer dans le tableau 1 la conjugaison du verbe 'aimer' en DALG aux temps les plus utilisés, c'est-à-dire le présent et le passé composé de l'indicatif ainsi qu'à l'impératif. Notons qu'en arabe il existe deux formes pour la deuxième personne du singulier, selon le sexe de la personne à qui l'on s'adresse.

À partir du tableau 1, nous pouvons déjà conclure que les lettres ن ، ت ، ي représentent des préfixes et les lettres ي ، و ، ت ، يت ، ينا ، يتي ، بيتو représentent des suffixes pour le DALG. Il est également à noter que la conjugaison au futur ne figure pas sur ce tableau car cette dernière est la même que celle au présent associé à des mots du futur tels que اومبعد *Awmbʕd*, راح *raAH*, غدوة *γadwaħ*, etc, voulant respectivement dire (demain, aller, après) (Harrat *et al.*, 2016).

2.2.2.2. Négation en dialecte algérien

La négation du DALG peut se faire de deux manières principales, soit 1) à l'aide des lettres ما ش *maA…š* ou encore des lettres م ش *maA…š*, 2) à l'aide du mot ماشي *maAšiy*. Prenons l'exemple de la phrase « Je l'aime » qui devient en DALG نحبو *nHabuw* pour le masculin et نحبها *nHabhaA* pour le féminin. La négation de cette phrase étant « Je ne l'aime pas » qui devient مانحبوش *mAnHabuwš*. Donc pour faire une analogie avec le français, le ما *maA* joue du rôle du 'ne' et le «ش» *š* joue le rôle de 'pas'. Prenons un autre exemple avec la phrase : « Cette fille est bien » qui

COD	Exemple	Traduction	COI	Exemple	Traduction
ني	تحبيني	Tu m'aimes	لي	قولي	Tu me le dis
ك	كرهتك	Je t'ai détesté	لك	قولتلك	Je te l'ai dit
ه هو و	كرهتو حبيته	Je l'ai détesté Je l'ai aimé	لو	نقولو	Je lui dis
ها	كرهتها	Je l'ai détesté	لها	قولتولها	Je le lui ai dit
نا	كرهتونا	Vous nous avez détestés	لنا نا	قولتونا	Tu nous l'as dit
كم	حبيناكم	Nous vous avons aimés	لكم	قولتلكم	Je vous l'ai dit
هم	كرهتوهم	Vous les avez détestés	لهم	قوللهم	Dis-leur

Tableau 2. *Les pronoms COD et COI du dialecte algérien*

devient en DALG «هاد الطفلة مليحة» *haAd AlTuflaħ mliyHaħ*. Sa négation étant « Cette fille n'est pas bien » qui devient en DALG «هاد الطفلة ماشي مليحة» *haAd AlTuflaħ maAšiy mliyHaħ*. Nous constatons maintenant que pour exprimer la négation de مليحة *mliyHaħ*'bien', nous employons le terme ماشي *maAšiy*. Donc, pour résumer, nous observons que la séquence ما ش est utilisée avec les verbes et le terme ماشي est utilisé avec les noms et les adjectifs.

2.2.2.3. COD et COI du dialecte algérien (clitiques pronominaux)

Les COD et COI sont également agglutinés aux verbes conjugués en DALG, au même titre que les pronoms personnels et la négation. Les pronoms COD et COI représentent des suffixes du verbe conjugué. Ces pronoms ont déjà été étudiés (Guellil et Azouaou, 2017 ; Saâdane et Habash, 2015 ; Harrat *et al.*, 2016). Il existe cependant un nombre plus important de pronoms que ceux cités dans ces travaux car l'agglutination des pronoms de base entre eux donne naissance à de nouveaux suffixes. Nous récapitulons dans le tableau 2, l'ensemble des COD et COI de base.

2.2.2.4. Noms et adjectifs dans le dialecte algérien

Comme pour toutes les langues, les noms et les adjectifs ont un genre et un nombre. Plusieurs auteurs (Harrat *et al.*, 2016 ; Harrat *et al.*, 2017 ; Guellil et Azouaou, 2016) attestent que pour former le féminin des noms et adjectifs en DALG la lettre ة doit être ajoutée comme suffixe. Par exemple le féminin de l'adjectif مليح *mliyH* 'bien' est مليحة *mliyHaħ*. Concernant le pluriel, tous les travaux étudiés s'accordent sur le fait que le masculin pluriel et le féminin pluriel sont formés à partir du nom ou de l'adjectif auxquels sont respectivement ajoutés les suffixes ين *yn* et ات *At*. Par exemple, le pluriel de l'adjectif فنيان *fanyaAn* 'fainéant' est فنيانين *fanyaAniyn* et le pluriel du nom شيخة *šiyxaħ* 'enseignante' est شيخات *šiyxaAt*.

3. Analyse de sentiments de l'arabe et ses dialectes : état de l'art

L'analyse de sentiments (AS) est un domaine interdisciplinaire se trouvant entre les domaines de traitement du langage naturel, de l'intelligence artificielle et la fouille de texte (Medhat *et al.*, 2014). L'AS s'effectue sur trois niveaux : documents, phrases et aspects. La richesse des médias sociaux en termes d'opinion et de sentiment a suscité l'intérêt de la communauté de recherche (Guellil et Boukhalfa, 2015). Cet intérêt est aussi important pour la langue arabe compte tenu du nombre massif des utilisateurs s'exprimant en arabe et ses dialectes sur Internet : 156 millions d'utilisateurs selon Siddiqui *et al.* (2016), soit 18,8 % de la population globale d'Internet (Korayem *et al.*, 2012). En nous fondons sur les caractéristiques de l'arabe et de ses dialectes présentées au sein de la section 2, nous concluons que les approches dédiées aux autres langues ne pourraient être appliquées dans notre cas (sauf avec modifications majeures). Au sein du présent travail, nous nous concentrons donc sur les travaux menés sur l'arabe et ses dialectes où nous nous fondons sur six états de l'art regroupant et analysant les travaux menés sur cette langue ainsi que ses dialectes (Kaseb et Ahmed, 2016 ; Biltawi *et al.*, 2016a ; Korayem *et al.*, 2012 ; Harrag, 2014 ; Assiri *et al.*, 2015 ; Alhumoud *et al.*, 2015). Après une analyse approfondie de ces études, nous concluons cependant que l'AS de l'arabe et de ses dialectes peut s'effectuer en suivant trois approches (comme pour toutes les autres langues) : supervisées, non supervisées et hybrides. Nous présentons, l'ensemble des travaux menés sur l'AS de l'arabe et ses dialectes tout en les regroupant par le type d'approche utilisé.

3.1. *Approches supervisées*

L'approche supervisée dépend de l'existence des données (documents, phrases, etc.) annotées comme positives, négatives ou neutres (Biltawi *et al.*, 2016b). L'approche supervisée, reconnue également par la classification supervisée, peut se faire en faisant appel à plusieurs algorithmes de classification tels les machines à vecteurs support (MVS), les classifieurs bayésiens naïfs (BN), les arbres de décision

(AD), etc. De nombreux travaux ont été réalisés pour analyser et classer les sentiments de l'arabe et ses dialectes en utilisant des approches supervisées. Citons notamment le travail de Cherif *et al.* (2015a) qui a fait appel à la technique MVS pour classer les sentiments de message écrit en arabe (ASM) en cinq classes allant de très bien à pas bien du tout. Pour réaliser cette tâche, les auteurs ont commencé par le prétraitement des phrases. Ils font également appel à un extracteur de lemmes présenté dans un travail précédent (Cherif *et al.*, 2015b) afin de supprimer les préfixes et suffixes des mots pour obtenir leurs radicaux. Il faut cependant noter que ces auteurs suppriment les préfixes et suffixes reliés à la conjugaison, au pluriel et aux pronoms. Il ne supprime cependant pas les affixes reliés à la négation qui pourraient affecter la qualité de l'analyse de sentiments.

Dans (Hadi, 2015) les auteurs ont utilisé les deux méthodes MVS et BN pour classifier un ensemble de messages en positif, négatif ou neutre. Pour ce faire, ils construisent un corpus arabe (ASM) contenant 3 700 messages extraits de Twitter. Chaque message a été annoté par trois locuteurs arabes natifs en positif, négatif et neutre. Nous enchaînons avec le système « SAMAR » analysant en parallèle la subjectivité d'un texte ainsi que ses sentiments (Abdul-Mageed *et al.*, 2014). Ce travail se focalise principalement sur le ASM ainsi que le dialecte égyptien. Les auteurs utilisent plusieurs corpus dont certains extraits des médias sociaux et d'autres ayant été utilisés dans d'autres travaux tels que (Diab *et al.*, 2010). Les auteurs de ce travail font appel à une variante de la MVS *« light »* proposée dans (Joachims, 2002). Ils se fondent cependant sur beaucoup de caractéristiques (*features*) dont l'analyse morphologique, la recherche des parties du discours, le lexique annoté, etc. Dans (Itani *et al.*, 2012), les auteurs ont exploité un modèle BN pour classifier automatiquement les sentiments des posts Facebook écrits en plusieurs dialectes arabes. Ils se concentrent sur les dialectes syriens, égyptiens, irakiens et libanais. Nous finissons cette partie avec le travail de Mdhaffar *et al.* (2017) qui combinent plusieurs classificateurs pour analyser le sentiment de messages écrits en dialecte tunisien. Parmi les principaux classificateurs utilisés la MVS et le BN. Dans ce travail, les auteurs présentent également la construction du corpus TSAC qui est un corpus tunisien dédié à l'analyse de sentiments.

Tous les travaux présentés au sein de cette catégorie se fondent sur un corpus annoté pour pouvoir effectuer la classification des sentiments. La construction de ce corpus est, dans la majorité des cas, manuelle, ce qui est très consommateur de temps et d'effort, et amène les auteurs à construire souvent des corpus réduits influant négativement les résultats. Du fait que notre approche s'appuie sur un lexique de sentiments, nous avons fondé les travaux présentés dans notre cas sur les différents prétraitements proposés.

3.2. *Approches non supervisées*

L'approche non supervisée est une approche qui se fond sur un lexique de sentiments. Plusieurs travaux ont également été menés en faisant appel à cette

approche. Nous commençons par le travail de Al-Ayyoub *et al.* (2015) qui a permis de construire un lexique de 120 000 termes arabes (ASM). Pour aboutir à ce dernier, les auteurs ont commencé par collecter des lemmes en arabe. Ils les ont traduits en anglais en utilisant Google traduction. Ils ont supprimé ensuite les mots répétés. Ces auteurs ne prennent pas en considération le contexte du lemme dans le processus de traduction. Ils utilisent ensuite un lexique de sentiments anglais pour déterminer leur valence et intensité.

Dans la même perspective, un autre lexique contenant 157 969 synonymes et 28 760 lemmes a été construit dans (Badaro *et al.*, 2014). Pour aboutir à ce dernier, les auteurs ont dû combiner plusieurs ressources de l'arabe. Dans (Mohammad et Turney, 2013), les auteurs ont développé un lexique de sentiments contenant 14 182 unigrammes anglais classés en positif ou négatif à l'aide du Amazon Mechanical Turk[5]. Ce lexique a ensuite été traduit en quarante langues dont l'ASM. L'auteur dans (AL-Khawaldeh, 2015) a construit également un lexique de sentiments, mais en traitant aussi de la négation. Ces auteurs se consacrent sur l'arabe (ASM) et ont également défini un ensemble de règles pour capturer la morphologie de la négation. Abdulla *et al.* (2014a) ont commencé par la construction d'un corpus contenant 4 000 commentaires textuels collectés à partir de Twitter et Yahoo Maktoob[6]. Un lexique a alors été construit, il ne contient que 300 mots.

Nous enchaînons avec les travaux de Abdulla *et al.* (2014b) qui se focalisent sur trois techniques de construction de lexiques avec une manuelle et deux autres automatiques. Pour la partie automatique, les auteurs se focalisent sur la traduction du lexique de sentiments anglais SentiStrength[7] en utilisant Google traduction. Nous terminons avec le travail de Mataoui *et al.* (2016) qui est le seul à étudier l'AS du DALG. Au sein de ce travail, les auteurs ont construit manuellement un lexique de sentiments en commençant par un lexique arabe et égyptien existant. Pour répondre aux caractéristiques morphologiques de cette langue et de ce dialecte, les auteurs utilisent l'outil de lemmatisation nommé « Khoja »[8] (Khoja et Garside, 1999).

Nous constatons donc que l'approche non supervisée est essentiellement fondée sur la construction de lexiques. Pour aboutir à cette construction, les auteurs tendent vers trois techniques : 1) construction manuelle (dans ce cas-là ce n'est pas purement supervisé car les mots sont manuellement annotés), 2) combinaison entre plusieurs ressources existantes, 3) traduction d'une ressource existante. Pour notre cas et vu que l'approche de construction manuelle est très consommatrice de temps (en plus, elle a déjà été présentée dans (Mataoui *et al.*, 2016)), que la combinaison de plusieurs ressources est impossible dans le cas du dialecte algérien, souffrant d'un manque considérable de ressources, nous optons pour une construction à base de traduction. Pour ce faire, nous nous appuyons sur les différents travaux de Abdulla *et al.* (2014a)

5. https ://www.mturk.com/mturk/welcome
6. L'édition arabe de yahoo.
7. http ://sentistrength.wlv.ac.uk/
8. https ://github.com/motazsaad/khoja-stemmer-command-line

et de Abdulla *et al.* (2014b), qui se focalisent sur des lexiques de sentiments existants pour faire la traduction vers l'ASM. En ce qui concerne les travaux sur le dialecte algérien, l'unique travail recensé est celui de Mataoui *et al.* (2016). Néanmoins, ce travail fait appel à l'outil « Khoja » pour la phase de lemmatisation. Cependant, cet outil est dédié à l'ASM et ne peut donc pas être utilisé pour le DALG. L'une des problématiques majeures reliées aux dialectes arabes est que les outils dédiés à l'ASM ne donnent pas de bons résultats pour ses dialectes (Harrat *et al.*, 2014). En plus, les auteurs utilisent cet outil sans y apporter aucune modification. Pour illustrer l'incapacité de cet outil à traiter le DALG, nous avons lemmatisé un ensemble de phrases en DALG et nous avons constaté que cet outil ne traitait en aucun cas les mots agglutinés, tels que مانساهاش *mAnnsAhAš* 'je ne l'oublierai pas'. Il change aussi le sens de certains mots, comme le mot مليح *mliyH* 'bien' qu'il lemmatise en ملح *mlH* 'sel'. Ceci est principalement dû aux spécificités et suffixes du DALG qui ne sont pas partagés avec l'ASM. Un tel outil ne peut donner de bons résultats que pour des messages pouvant être classés comme des messages partagés entre l'ASM et le DALG, par exemple كرهت حياتي *kraht HyaAtiy* 'j'en ai marre de ma vie'.

3.3. *Approches hybrides*

L'approche hybride consiste à combiner les méthodes utilisées dans l'approche supervisée et non supervisée. Par exemple, dans le travail de Hedar et Doss (2013), les auteurs ont utilisé un classificateur MVS. Pour faire cette classification, les auteurs utilisent un lexique contenant 1 300 mots dont 600 sont positifs et 700 négatifs. Ce travail s'appuie sur l'argot égyptien. Les résultats expérimentaux ont montré que l'utilisation du lexique améliore considérablement les résultats. Dans (Khalifa et Omar, 2014), les auteurs ont proposé une méthode hybride fondée sur un lexique et un classificateur BN en même temps. La méthode proposée est précédée d'une phase de prétraitement (normalisation, segmentation, etc.). Le lexique intervient pour remplacer les mots avec leurs synonymes. Ces auteurs se focalisent sur l'arabe standard (ASM).

L'approche hybride est une approche qui donne de très bons résultats, mais nous ne pouvons l'appliquer à ce stade puisque nous ne disposons pas de données annotées.

Dans le but de synthétiser tous les travaux présentés, nous les classifions dans le tableau 3, par rapport à la langue étudiée (ASM, dialecte arabe en donnant le type du dialecte), ainsi que l'approche de classification et la méthode utilisée pour le travail présenté. En ce focalisant sur le tableau 3, nous constatons qu'un seul travail a été mené sur l'AS du DALG. Nous précisons cependant que le DALG est un dialecte peu étudié en général. Peu de travaux se sont focalisés sur ce dernier (Meftouh *et al.*, 2015 ; Harrat *et al.*, 2017 ; Guellil *et al.*, 2017b ; Saâdane et Habash, 2015 ; Guellil *et al.*, 2017a ; Harrat *et al.*, 2016)

Approches et méthodes utilisées		ASM	Dialecte arabes	
			Les travaux	Le dialecte étudié
Approche supervisée	MVS	(Cherif *et al.*, 2015)	(Abdul-Mageed *et al.*, 2014)	Égyptien
	BN		(Itani *et al.*, 2012)	Syrien Égyptien Irakien libanais
	MVS + BN	(Hadi, 2015)	(Mdhaffar *et al.*, 2017)	Tunisien
Approche non supervisée		(Al-Ayyoub *et al.*, 2015) (Abdulla *et al.*, 2014b) (Abdulla *et al.*, 2014a) (Badaro *et al.*, 2014)	(Abdulla *et al.*, 2014a)	Jordanien
		(Mohammad *et al.*, 2013) (AL-Khawaldeh, 2015) (Mataoui *et al.*, 2016)	(Mataoui *et al.*, 2016)	Algérien
Approche hybride	MVS + lexique		(Hedar *et al.*, 2013)	L'argot égyptien
	NB + lexique	(Khalifa *et al.*, 2014)		

Tableau 3. *Classification et synthèse des travaux étudiés*

4. Analyse de sentiments de textes écrits en dialecte algérien

Dans cette partie, nous définissons et implémentons une approche non supervisée, fondée sur un lexique de sentiments, pour déterminer la valence (positive, négative) et l'intensité (1,54, $-\ 2,87$, etc.) d'un message donné écrit en DALG. Notre approche reçoit en entrée un message écrit en DALG (en caractères arabes) ainsi qu'un lexique de sentiments en DALG préalablement construit en traduisant un lexique de sentiments anglais existant. Elle retourne en sortie la valence du message ainsi que son intensité. Pour ce faire, notre approche est constituée de deux étapes principales 1) construction d'un lexique de sentiments en DALG, 2) calcul de la valence et de l'intensité d'un message (en appelant le lexique créée à l'étape 1). La figure 1 illustre l'architecture générale de notre approche.

4.1. *Étape 1 : construction d'un lexique de sentiments en dialecte algérien*

Cette étape reçoit en entrée un lexique de sentiments en anglais (nous choisissons l'anglais car c'est la langue qui a bénéficié de plus de travaux sur l'AS (Guellil et

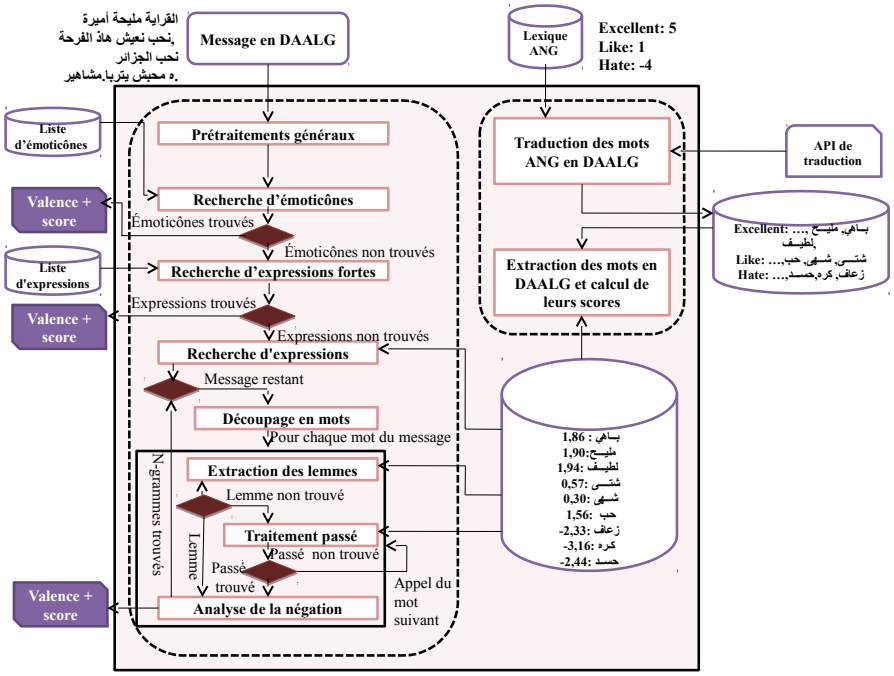


Figure 1. *Architecture générale de notre approche*

Boukhalfa, 2015)). Chaque mot de ce lexique est traduit en appelant une API de traduction. Après la phase de traduction, un lexique de sentiments est construit en extrayant chaque terme en DALG et calculant son score. Cette étape contient donc deux sous-étapes : 1) traduction des mots du lexique anglais en DALG et 2) extraction des mots en DALG et calcul de leurs scores.

4.1.1. *Traduction des mots du lexique anglais en dialecte algérien*

Pour faire cette traduction, nous faisons appel à l'API glosbe. Cette dernière prend en entrée un mot en anglais et retourne un ensemble de mots en DALG. La spécificité de cette API est que la traduction est faite par des utilisateurs ordinaires natifs du DALG. Nous traduisons chacun des mots de notre lexique de sentiments anglais en faisant appel à cette API. Nous affectons à tous les mots récoltés le même score que le mot en anglais. Prenons par exemple le mot anglais 'excellent'[9] voulant dire 'excellent' et ayant un score égal à $+5$. Sa traduction en DALG donne les mots :

9. https ://glosbe.com/en/arq/excellent

باهي *baAhiy*, لطيف *lTiyf*, مليح *mliyH*, etc. Tous ces mots ont donc un score égal à $+5$, de même que le mot 'excellent'. Nous nous inspirons dans cette partie de plusieurs travaux : 1) le travail de Elarnaoty *et al.* (2012) qui fait la traduction de la version anglaise du lexique MPQA (Multi-Perspective Question Answering) en arabe, 2) le travail de El-Halees *et al.* (2011) qui fait la traduction de Sentistrength en arabe également, 3) le travail de Al-Ayyoub *et al.* (2015) qui traduit des lemmes arabes en anglais.

4.1.2. *Extraction des mots en dialecte algérien et calcul de leurs scores*

Après la réalisation de la première phase (section 4.1.1), nous nous sommes rendu compte qu'un mot en DALG est associé à plusieurs mots en anglais et peut donc avoir plusieurs scores. Prenons par exemple le mot مليح *mliyH* 'bien'. Ce mot peut être associé aux mots anglais : *excellent, best, generous,* etc. Nous observons donc que les mots anglais auxquels est associé le mot مليح *mliyH* peuvent avoir différents scores ('excellent' $(+5)$, 'generous' $(+2)$, etc). Nous extrayons donc, au sein de cette partie, tous les mots en DALG et sans répétition et calculons leurs scores. Pour le calcul du score, nous prenons la moyenne des scores de tous les mots anglais auxquels notre mot en DALG est associé. Nous obtenons ainsi notre lexique en DALG contenant par exemple le mot مليح *mliyH* avec un score égal à 1,90, le mot كره *krah* avec un score égal à $-3,16$, etc.

4.2. *Étape 2 : calcul de la valence d'un message écrit en dialecte algérien*

Cette étape reçoit en entrée un message écrit en DALG et retourne en sortie sa valence et son intensité. Si nous prenons, par exemple, la phrase القررررراية مليحـــــة # أميرة *mliyHaħ AlqrrrrraAyaħ # Âmiyraħ* traduite en 'C'est bien d'étudier Amira'. Si nous nous appuyons sur le lexique construit à l'étape 1, cette phrase est reconnue positive avec une intensité globale égale à 1,69. Nous constatons cependant que pour arriver à un tel résultat, un ensemble de traitements doivent être appliqués à ce message dont 1) différents prétraitements allant de la suppression des lettres répétées à la recherche des émoticônes et expression fortes, 2) recherche des n-gramme du message présents dans le lexique, 3) extraction du lemme de chaque mot du message non identifié comme n-gramme, 4) traitement du passé, et enfin 5) analyse de la négation.

4.2.1. *Prétraitement d'un message en dialecte algérien*

Nous nous inspirons des différents travaux de Cherif *et al.* (2015a) pour proposer plusieurs prétraitements nécessaires au traitement de l'arabe et de ses dialectes. Nous nous inspirons également des différents travaux de Guellil et Azouaou (2017), Harrat *et al.* (2016) et Saâdane et Habash (2015) se focalisant sur les caractéristiques propres au dialecte algérien. Nous proposons donc l'ensemble des prétraitements suivants :
– suppression des blancs, des lettres longues (reconnus par tatweel), par exemple

مليحـــــة devient مليحة *mliyHaħ* ;

– suppression des exagérations, par exemple القرررررراية *AlqrrrrraAyaħ* est transformé en القراية *AlqraAyaħ* ;

– suppression de certaines ponctuations telles que le # et espacements de certains points ('., !, ?') attachés au mots ;

– remplacement des caractères arabes par leurs Unicodes pour traiter le phénomène relié à la présence de différentes lettres selon leurs emplacements (traités au sein de la section 2.1) ;

– recherche d'émoticônes et d'expressions fortes tels ,نموت على *nmuwt ςlaý* 'j'adore' ou encore ينعل الله *yanςal Allah* 'Que Dieu maudisse', etc. Le but étant d'attribuer directement au message la valence de l'émoticône ou de l'expression trouvée. Dans le cas où il y a plusieurs émoticônes ou expressions, nous ne prenons en considération que la première ;

– prétraitements reliés à l'opposition exprimée en dialecte algérien à l'aide du mot بصح *baSaH* 'mais', notre système ne prend en considération que la partie du message qui se trouve après بصح. Nous procédons ainsi, car nous avons constaté que le mot بصح annulait le sentiment de la partie qui le précède. Prenons l'exemple du message : حبيت نخرج نلعب نفرح بصح راني مريضة *Habiyt nuxruj nalςb nafraH baSaH raAniy mriyĎaħ* traduit en 'je voulais sortir jouer être heureuse mais je suis malade'. Ce message à beau contenir plus de mots positifs que négatifs, il reste négatif parce que c'est la partie qui est après l'opposition qui détermine véritablement le sentiment.

4.2.2. *Recherche des expressions du message dans le lexique de sentiments*

Notre lexique de sentiments peut contenir des séquences d'un mot tel que مليح 'bien', de deux mots telles que عالي بصوت *bSawt ςaAliy* 'à haute voix', de trois mots tels que مرا مرا على *ςlaŷ maraA maraA* 'des fois' ou encore de quatre mots tels que برك وحدة مرة يستعمل *yastaςmal maraħ waHdaħ bark* 'il utilise une seule fois seulement'. Pour rechercher la présence d'une séquence de mots du message dans le lexique, nous formons d'abord l'ensemble de ces séquences, par exemple القراية مليحة أميرة *Âmiyraħ mliyHaħ AlqraAyaħ*, ce dernier contient trois mots individuels (القراية ، مليحة ، أميرة), deux séquences de deux mots (القراية مليحة ، مليحة أميرة) et une seule séquence de trois mots (القراية مليحة أميرة). Une fois les séquences de mots du message construites, nous commençons leur recherche dans le lexique. Dans le cas de l'exemple présenté, un seul unigramme مليحة *mliyHaħ* a pu être trouvé dans le lexique. Un traitement de la négation est ensuite indispensable pour déterminer le score des séquences retrouvées (que nous présenterons dans la section 4.2.5).

4.2.3. *Extraction du lemme des mots en dialecte algérien*

Dans cette étape, nous extrayons de chaque mot (non reconnu par l'étape 4.2.2) son lemme (à l'aide du lexique de sentiments). Nous procédons comme suit : 1) nous recherchons tous les mots du lexique inclus dans notre mot, 2) nous sélectionnons les mots ayant le plus grand nombre de lettres et 3) nous vérifions que le mot sélectionné peut être découpé en *préfixe + lemme + suffixe* et nous enlevons ainsi les préfixes et suffixes pour ne garder que le lemme et ce, en se fondant sur le travail de Cherif *et al.* (2015a). Pour cela nous définissons une liste globale des préfixes et suffixes du DALG. Nous illustrons cette étape en réutilisant l'exemple القراية مليحة أميرة. Rappelons juste que l'unigramme مليحة *mliyHaħ* a été identifié au sein de l'étape 4.2.2. Nous aurons donc à traiter les deux mots restants : القراية *AlqraAyaħ* et أميرة *Âmiyraħ*. Aucune partie du mot أميرة n'a été retrouvée dans notre lexique de sentiments. Néanmoins le mot قراي *qraAy* a été détecté comme faisant partie du mot القراية. Ce mot se présente donc sous cette forme : ال+قرايها+ة *Al+qraAy+ħ*. Comme le mot ال (déterminant) est un préfixe reconnu du DALG et la lettre ة est un suffixe reconnu également, le lemme قراي *qraAy* est donc validé. La négation doit ensuite être traitée (se référer à la section 4.2.5).

4.2.4. *Traitement du passé*

Certains verbes conjugués au passé ne peuvent pas être traités comme la plupart des mots cités dans la section 4.2.1. Nous ne pouvons donc pas directement extraire les affixes de ces verbes car nous devons d'abord faire des transformations sur leurs lemmes. Nous citons par exemple, les deux verbes نسى *nsaý* 'oublier' et بكى *bkaý* 'pleurer'. La conjugaison de ces deux verbes au passé est : نسيت *nsiyt* et بكيت *bkiyt* (pour la première personne du singulier). La lettre ى *ý* est donc supprimée afin de rajouter les suffixes nécessaires. Pour pouvoir former le lemme de ces verbes, il faut supprimer les suffixes et ajouter la lettre ى *ý*. Pour réaliser cette étape, nous procédons comme suit : 1) identification de la liste des suffixes passés (par exemple يتو، ينا، يت,etc.) et suppression de ces derniers dans les mots analysés (à cette étape نسيت *nsyt* est transformé en نس *ns*, puisque le suffixe يت est enlevé), 2) ajout de la lettre ى à la fin des mots récoltés (à cette étape نس est transformé en نسى *nsý*) et 3) recherche du lemme obtenu dans le lexique de sentiments, si le lemme obtenu est trouvé, le mot est validé et la négation est ensuite traitée (voir la section 4.2.5).

4.2.5. *Analyse de la négation*

L'analyse de la négation représente un défi de recherche important concernant l'AS et pas seulement pour l'arabe, mais pour toutes les langues. Néanmoins ce

défi est accentué dans le cas de l'arabe et ses dialectes où la négation s'agglutine le plus souvent au mot, au même titre que les différents pronoms. Pour plus de détails sur la négation, nous vous invitons à vous référer à la section 2.2.2.2. Les utilisateurs peuvent faire appel à la négation de différentes manières, par exemple le mot ماتحبكمش *mAnHabkumš* 'je ne vous aime pas' peut s'écrire ما نحبكمش *mA nHabkumš*, ما نحبكم ش *mAnHabkum* š ou encore ما نحبكم ش *mA nHabkum š*. Nous constatons que la négation peut être agglutinée aux termes comme elle peut être séparée de ces derniers. Nous traitons dans ce travail deux sortes de négations : 1) la négation agglutinée au mot, et 2) la négation séparée du mot. Pour les deux cas, nous définissons une liste de préfixes et de suffixes reliés à la négation. Nous avons cependant constaté que, dans la plupart des cas, la négation n'influe pas seulement sur le mot qu'elle précède, mais sur le reste de la phrase également. Une fois qu'un préfixe ou un suffixe de négation est détecté, nous inversons le score des mots succédant à cette négation (multipliant le score par (-1)).

5. Étude expérimentale

Pour développer notre solution nous nous sommes inspirés du programme élaboré par Taboada *et al.* (2011). Le programme et les différents lexiques de ses auteurs sont librement téléchargeables[10]. La différence entre cette solution et la nôtre réside au niveau de l'identification des parties du discours faite par ces auteurs et non réalisée par notre système. Ces auteurs utilisent un outil très répandu pour l'identification des parties du discours[11], qui ne peut pas être utilisé pour le DALG. Nous définissons donc un seul lexique regroupant les quatre parties grammaticales : adjectifs, verbes, noms et adverbes. Pour illustrer les résultats de l'AS du DALG, nous présentons les quatre parties : 1) l'environnement expérimental, 2) les résultats expérimentaux et leurs analyse, 3) l'analyse des cas d'erreurs, et 4) perspective d'extension de notre approche à l'ASM.

5.1. *Environnement expérimental*

Nous présentons dans cette section l'ensemble des données et des paramètres utilisés dans nos expérimentations : 1) les lexiques, 2) les corpus de test, et 3) les différents types d'expérimentations effectuées.

5.1.1. *Lexiques utilisés*

Pour la construction des lexiques, nous avons fait appel à deux lexiques anglais. D'une part, le SOCAL, qui est utilisé dans (Taboada *et al.*, 2011), et d'autre part, le SentiWordNet qui est utilisé dans (Baccianella *et al.*, 2010). Concernant SOCAL, nous

10. https ://github.com/sfu-discourse-lab/SO-CAL
11. Stanford CoreNLP : https ://stanfordnlp.github.io/CoreNLP/

avons d'abord fusionné les lexiques d'adjectifs, de verbes, de noms et d'adverbes. Nous avons ainsi obtenu 6 769 termes dont le sentiment est étiqueté entre -1 et -5 pour les termes négatifs et entre $+1$ et $+5$ pour les termes positifs. Après que l'intégralité des termes a été envoyée à l'API de traduction, 3 952 termes en anglais ont été reconnus et traduits. Notre lexique final, que nous nommons SOCALALG, contient 2 375 termes en DALG dont 1 363 termes négatifs, 948 termes positifs et 64 termes neutres (avec un sentiment égal à 0). Pour SentiWordNet, nous avons commencé par construire un lexique contenant l'ensemble des termes de SentiWordNet avec la moyenne du sentiment de chaque terme. Pour ce faire, nous avons fait appel à l'API JAVA de Petter Tönberg fourni dans le site officiel de SentiWordNet[12]. Comme les termes de SentiWordNet ont un sentiment étiqueté entre -1 et $+1$, nous avons multiplié tous les sentiments par 5 (pour l'aligner au lexique SOCAL). Le lexique obtenu contient 39 885 termes étiquetés entre $+0,05$ et $+5$ pour les termes positifs et entre $-0,05$ et -5 pour les termes négatifs. Une fois envoyés à l'API de traduction, 12 780 termes en anglais ont été reconnus et traduits. Notre lexique final que nous nommons SentiALG, contient 3 408 termes en DALG dont 1 856 négatifs, 1 539 positifs et 13 neutres.

5.1.2. *Corpus de test utilisés*

Nous avons utilisé dans ce travail deux corpus de test : 1) le premier corpus contient 323 messages (phrases) pris du corpus PADIC, dont 157 sont positifs et 166 négatifs, PADIC étant le seul corpus parallèle multidialectal, contenant également le DALG (Meftouh *et al.*, 2015), 2) le deuxième corpus contient 426 messages extraits du média social Facebook[13], dont 220 sont positifs et 206 sont négatifs. Les statistiques relatives à ces corpus sont présentées dans le tableau 5.1.2. Ces corpus ont été annotés par deux annotateurs natifs du dialecte algérien (un des auteurs de ce travail étant l'un des annotateurs. L'accord entre annotateurs (Kappa) est de 0,954 (0,956 pour le corpus PADIC et 0,952 pour le corpus Facebook). Les annotateurs ont reçu les instructions suivantes :

– les messages doivent être annotés en deux classes (positive ou négative). Les messages objectives ou neutres ne doivent pas être pris en considération ;

– les annotateurs ne doivent en aucun cas se référer à leur opinion personnelle mais plutôt au sentiment général de la phrase ;

– prendre en considération les exagérations et les émoticônes qui pourraient accentuer le sentiment ;

– dans le cas où un message contiendrait du texte et des émoticônes de valences différentes, il faudrait privilégier le sentiment porté par le texte pour l'annotation.

5.1.3. *Types d'expérimentations effectuées*

Nos expérimentations ont été effectuées sur 749 messages répartis en deux corpus et sur les deux lexiques SOCALALG et SentiALG. En plus de cela, nous avons mené

12. http ://sentiwordnet.isti.cnr.it/

13. https ://fr-fr.facebook.com/policy.php

Corpus	PADIC			Facebook		
	Pos.	Nég.	Tout	Pos.	Nég.	Tout
Nbre messages	157	166	323	220	206	426
Nbre mots	849	952	1 802	1 711	1 735	3 446
Nbre mots/message	5,41	5,73	5,57	7,78	8,42	8,1
Nbre caractères/message	21,9	24,0	23,0	33,3	35,9	34,6
Nbre messages avec émoticônse	0	0	0	38	19	57

Tableau 4. *Statistiques relatives aux corpus de test*

	Lexique utilisé	PADIC			Facebook		
		P	R	F1	P	R	F1
n-gramme (1)	SOCALALG	0,71	0,45	0,55	0,68	0,36	0,47
	SentiALG	0,73	0,45	0,56	0,67	0,38	0,48
n-gramme + prétraitement (2)	SOCALALG	0,72	0,46	0,56	0,74	0 ,42	0,53
	SentiALG	0,74	0,47	0,57	0,71	0,42	0,53
N-gramme + prétraitement + lemme (3)	SOCALALG	0,70	0,69	0,70	0,70	0,63	0,67
	SentiALG	0,75	0,74	0,74	0,69	0,63	0,66
n-gramme + prétraitement + lemme + passé (4)	SOCALALG	0,70	0,70	0,70	0,72	0,64	0,67
	SentiALG	0,75	0,74	0,74	0,69	0,64	0,66
n-gramme + prétraitement + lemme + passé + négation (5)	SOCALALG	0,75	0,74	0,75	0,68	0,61	0,64
	SentiALG	0,78	0,78	0,78	0,67	0,61	0,64

Tableau 5. *Résultats expérimentaux avec les deux lexiques utilisés*

pour chaque corpus avec chaque lexique cinq expérimentations : 1) n-gramme, 2) n-gramme + prétraitement, 3) n-gramme + prétraitement + lemme, 4) n-gramme + prétraitement + lemme + passé, et 5) n-gramme + prétraitement + lemme + passé + négation. Nous souhaitons montrer à travers ces expérimentations l'impact de chaque étape de notre approche sur les résultats obtenus.

5.2. *Résultats expérimentaux*

Nous illustrons nos résultats à l'aide de trois métriques : la précision (P), le rappel (R) et la F-mesure (F1). Nous présentons dans le tableau 5, les résultats obtenus (P, R et F1) sur les deux corpus de test utilisés (PADIC et Facebook) avec les deux lexiques SOCALALG et SentiALG.

D'après le tableau 5, nous constatons que les résultats évoluent positivement après l'exécution de chaque étape de notre approche et ce pour nos deux lexiques (sauf pour le traitement de la négation où il y a une légère régression des résultats). L'évolution la plus importante en termes de F1 a été observée au sein du corpus PADIC où les résultats initiaux étaient de 0,56 pour finir avec un score de 0,78.

5.3. *Analyse des cas d'erreurs*

Après une analyse approfondie des messages qui sont mal classés par notre système, nous présentons dans le tableau 6 les principales erreurs de classification du sentiment du DALG. D'après le tableau 6, nous constatons qu'il existe sept principales erreurs 1) les pluriels irréguliers, 2) les mots non existants dans nos lexiques, 3) les mots ayant une valence différente dans nos lexiques, 4) les mots en ASM, non utilisés en DALG, 5) les mots ayant une intensité trop élevée, 6) le non-traitement des intensificateurs, 7) les lemmes retrouvés dans le lexique par erreur. Les pluriels irréguliers signifient que la formation du pluriel de certains mots ne suit pas les règles que nous avons présentées dans la section 2.2.2.4. Prenons par exemple le mot مليح *mliyH* signifiant 'bien', son pluriel n'est pas مليحين *mliyHiyn* selon la règle, mais plutôt le mot ملاح *mlaAH*. Nous avons également constaté que plusieurs mots importants ne sont pas présents dans l'un de nos lexiques ou parfois même dans les deux. Parmi ces derniers, nous citons : كادو *kaAduw* 'un cadeau', زعف *zʕaf* 'Il est en colère', etc. D'autres mots tels que عجيب *ʕjiyb* 'bizarre' ont une valence différente d'un lexique à un autre. Certains mots tels que مطبوعة 'imprimé' *maTbuwʕaħ* ou encore احسن *AHsan* 'le meilleur' sont des mots en ASM et donc non reconnus en DALG. Certains mots tels que واش *waAš* 'Comment ?' ont une intensité trop élevée dans les lexiques, ce qui peut fausser le calcul du score final. D'autres mots comme بزاف *bzaAf* 'très', représentant des intensificateurs, ne sont pas reconnus comme tels. Leurs traitements amélioraient le score final. Enfin, les mots comme بكينا *bkiynaA* 'Nous avons pleuré' dont un lemme a été reconnu par erreur, avant même de faire appel au traitement du passé.

Afin de résoudre ces problématiques, nous proposons les améliorations suivantes de notre système :
– proposition d'un lemmatiseur propre au dialecte algérien dédié à la segmentation et à l'analyse des pluriels irréguliers ;
– extension de notre lexique pour qu'il ait un champ plus étendu. Nous prévoyons de faire cet extension à l'aide des techniques du « word Embedding » ;
– intégration du contexte dans l'annotation du lexique ;
– enrichissement de notre lexique en DALG avec un autre lexique en ASM.

Messages mal classés	Traduction	Cause principale de l'erreur	Annotation _cible	Annotation _système
الحمد لله انا نعرف غير ***الملاح*** AlHmd llh AnA nçrf ɣyr AlmlAH	Dieu soit loué je ne connais que des gens bien	Le mot 'الملاح' qui est le pluriel de 'مليح' n'a pas été reconnu	Positif	Négatif
هاديك لي جابتلي ***كادو*** في اللخر تاع العام HAdyk ly jAbtly kAdw fy Allxr tAç AlçAm	Celle qui m'a ramené un cadeau en fin d'année	Le mot 'كادو' n'existe pas dans nos lexiques	Positif	Négatif
قولك نقرا تلقى ***يواش***	Qu'est-ce qu'on dit tu études tu trouveras	Le mot 'واش' ayant une intensité négative élevée	Positif	Négatif
القراية في وقتنا عادت ***بزاف*** صعيبة AlqrAyħ fy wqtnA çAdt bzAf Sçybħ	Les études en notre temps sont devenues très difficiles	Le non-traitement des intensificateurs tels que 'بزاف'	Négatif	Positif
بكينا! bkiynaA. !	On a pleuré !	Le mot 'كي' est retrouvé ne donnant pas la chance au mot 'بكى' d'être recherché	Négatif	Positif
المطبوعة احسن الف مرة AlmTbwçħ AHsn Alf mrħ	L'imprimerie est mieux mille fois.	Contient des mots en ASM qui n'ont pas été reconnus	Positif	Négatif
عجيب! çjiyb	Merveilleux	Le mot 'عجيب' ayant une valence positive dans un lexique et négative dans un autre	Positif	Négatif

Tableau 6. *Les principales erreurs de classification de notre système*

5.4. *Perspective d'extension de notre approche à l'ASM*

Afin de comparer notre approche avec les travaux existants, nous proposons l'extension de cette dernière à l'ASM. Pour ce faire, nous construisons en premier lieu nos lexiques : SOCAL_ASM et Senti_ASM en suivant la même méthode que celle présentée dans la section 4.1. Nous obtenons ainsi 5 190 termes pour SOCAL_ASM et 15 838 pour Senti_ASM. Afin de pouvoir appliquer notre approche sur les deux lexiques obtenus, nous utilisons le corpus ASM utilisé dans (Altowayan et Tao, 2016). Ce corpus contient 4 294 messages, dont 2 147 positifs et 2 147 négatifs. Il a été construit en combinant plusieurs autres corpus présentés au sein de la littérature. En ajoutant certains affixes à notre approche, dédiés à l'ASM, nous obtenons un F1-score égal à 0,58 pour Senti_ASM et 0,62 pour SOCAL_ASM. Nous constatons que la mise à l'échelle de notre approche sur l'ASM donne des résultats compétitifs dans le cadre d'une approche fondée sur les lexiques. Les résultats fournis par SOCAL_ASM (F1 égale à 0,62) ne s'éloignent pas trop des résultats que nous avons présentés dans le tableau 5 pour le corpus Facebook. Ceci est tout à fait justifié car le corpus ASM que nous avons utilisé est essentiellement alimenté avec des données provenant des médias sociaux.

6. Conclusion et perspectives

Dans cet article, nous avons proposé et implémenté une approche d'AS de messages écrits en DALG. Cette approche s'appuie sur la construction et l'utilisation de lexiques de sentiments en DALG. Elle est fondée également sur l'agglutination qui est une problématique très importante dans le traitement de l'ASM et de ses dialectes. Nous avons évalué cette approche à l'aide des deux lexiques de sentiments construits, SOCALALG et SentiALG, ainsi qu'un corpus de test annoté manuellement contenant 747 messages. Les résultats expérimentaux indiquent une amélioration continue après l'exécution de chaque étape de notre approche atteignant une précision de 0,78, un rappel de 0,78 et une F-mesure de 0,78. Ces résultats pourraient cependant être améliorés en prenant en considération plusieurs facteurs étudiés dans la section 5.3. Nos travaux futurs s'orientent vers une proposition d'intégrer l'ensemble des critères suivants :

– l'analyse des pluriels irréguliers et proposition d'une liste de changements et d'affixes qui pourraient traiter ces pluriels ;
– la fusion des deux lexiques SOCAL_ALG et Senti_ALG ainsi que la fusion (ASM et DALG), car plusieurs utilisateurs utilisent l'alternance codique entre l'ASM et le DALG. Il faudrait également procéder à l'intégration d'autres lexiques de sentiments anglais tels que MPQA ou SentiStrenght ;
– la définition d'une méthode combinant le traitement des lemmes et du passé en même temps ;
– le traitement de l'intensification ;
– la revue manuelle des lexiques utilisés pour pouvoir y intégrer la notion de contexte ainsi que l'enrichissement du lexique obtenu à l'aide des techniques de l'apprentissage profond (*deep learning*).

Enfin, nous comptons également étendre cette approche aux corpus annotés pour pouvoir exploiter les techniques de classification usuelles. Néanmoins ces approches requièrent des corpus annotés. La construction de ces corpus est très coûteuse en termes de temps et d'efforts. Nous prévoyons donc de proposer une approche de construction automatique ou semi-supervisée de ces corpus.

7. Remerciements

Les premiers auteurs sont soutenus par l'École nationale supérieure d'informatique (ESI) à Alger ainsi que l'École supérieure des sciences appliquées d'Alger (ESSAA). Le troisième auteur est soutenu par la DGE (ministère de l'Industrie de France) et par la DGE (ministère de l'Économie de France) : projet « DRIRS », référencé par le numéro 172906108. Nous tenons à remercier Billel Gueni pour sa collaboration et ses précieux retours.

8. Bibliographie

Abdul-Mageed M., Diab M., Kübler S., « SAMAR : Subjectivity and sentiment analysis for Arabic social media », *Computer Speech & Language*, vol. 28, n° 1, p. 20-37, 2014.

Abdulla N. A., Ahmed N. A., Shehab M. A., Al-Ayyoub M., Al-Kabi M. N., Al-rifai S., « Towards improving the lexicon-based approach for arabic sentiment analysis », *International Journal of Information Technology and Web Engineering (IJITWE)*, vol. 9, n° 3, p. 55-71, 2014a.

Abdulla N., Mohammed S., Al-Ayyoub M., Al-Kabi M. *et al.*, « Automatic lexicon construction for arabic sentiment analysis », *Future Internet of Things and Cloud (FiCloud), 2014 International Conference on*, IEEE, p. 547-552, 2014b.

Al-Ayyoub M., Essa S. B., Alsmadi I., « Lexicon-based sentiment analysis of arabic tweets », *International Journal of Social Network Mining*, vol. 2, n° 2, p. 101-114, 2015.

AL-Khawaldeh F. T., « A Study of the Effect of Resolving Negation and Sentiment Analysis in Recognizing Text Entailment for Arabic. », *World of Computer Science & Information Technology Journal*, 2015.

Alhumoud S. O., Altuwaijri M. I., Albuhairi T. M., Alohaideb W. M., « Survey on arabic sentiment analysis in twitter », *International Science Index*, vol. 9, n° 1, p. 364-368, 2015.

Altowayan A. A., Tao L., « Word embeddings for Arabic sentiment analysis », *Big Data (Big Data), 2016 IEEE International Conference on*, IEEE, p. 3820-3825, 2016.

Assiri A., Emam A., Aldossari H., « Arabic sentiment analysis : a survey », *International Journal of Advanced Computer Science and Applications*, vol. 6, n° 12, p. 75-85, 2015.

Baccianella S., Esuli A., Sebastiani F., « Sentiwordnet 3.0 : an enhanced lexical resource for sentiment analysis and opinion mining. », *LREC*, vol. 10, p. 2200-2204, 2010.

Badaro G., Baly R., Hajj H., Habash N., El-Hajj W., « A large scale Arabic sentiment lexicon for Arabic opinion mining », *ANLP 2014*, 2014.

Biltawi M., Etaiwi W., Tedmori S., Hudaib A., Awajan A., « Sentiment classification techniques for Arabic language : A survey », *Information and Communication Systems (ICICS), 2016 7th International Conference on*, IEEE, p. 339-346, 2016a.

Biltawi M., Etaiwi W., Tedmori S., Hudaib A., Awajan A., « Sentiment classification techniques for Arabic language : A survey », *7th International Conference on Information and Communication Systems (ICICS),*, IEEE, p. 339-346, 2016b.

Cherif W., Madani A., Kissi M., « Towards an efficient opinion measurement in Arabic comments », *Procedia Computer Science*, vol. 73, p. 122-129, 2015a.

Cherif W., Madani A., Kissi M., « Towards an efficient opinion measurement in Arabic comments », *Procedia Computer Science*, vol. 73, p. 122-129, 2015b.

Diab M., Habash N., Rambow O., Altantawy M., Benajiba Y., « COLABA : Arabic dialect annotation and processing », 2010.

El-Halees A. *et al.*, « Arabic opinion mining using combined classification approach », 2011.

Elarnaoty M., AbdelRahman S., Fahmy A., « A machine learning approach for opinion holder extraction in Arabic language », *arXiv preprint arXiv :1206.1011*, 2012.

Fishman J. A., « Bilingualism with and without diglossia ; diglossia with and without bilingualism », *Journal of social issues*, vol. 23, n° 2, p. 29-38, 1967.

Guellil I., Azouaou F., « Arabic dialect identification with an unsupervised learning (based on a lexicon). application case : Algerian dialect », *Computational Science and Engineering (CSE) and IEEE Intl Conference on Embedded and Ubiquitous Computing (EUC) and 15th Intl Symposium on Distributed Computing and Applications for Business Engineering (DCABES), 2016 IEEE Intl Conference on*, IEEE, p. 724-731, 2016.

Guellil I., Azouaou F., « ASDA : Analyseur Syntaxique du Dialecte Alg {\'e} rien dans un but d'analyse s {\'e} mantique », *arXiv preprint arXiv :1707.08998*, 2017.

Guellil I., Azouaou F., Abbas M., « Comparison between Neural and Statistical translation after transliteration of Algerian Arabic Dialect », *WiNLP : Women & Underrepresented Minorities in Natural Language Processing (co-located withACL 2017)*, p. 1-5, 2017a.

Guellil I., Azouaou F., Abbas M., Fatiha S., « Arabizi transliteration of Algerian Arabic dialect into Modern Standard Arabic », *Social MT 2017/First workshop on Social Media and User Generated Content Machine Translation*, p. 1-8, 2017b.

Guellil I., Boukhalfa K., « Social big data mining : A survey focused on opinion mining and sentiments analysis », *Programming and Systems (ISPS), 2015 12th International Symposium on*, IEEE, p. 1-10, 2015.

Habash N., Soudi A., Buckwalter T., « On arabic transliteration », *Arabic computational morphology*, Springer, p. 15-22, 2007.

Hadi W., « Classification of Arabic Social Media Data », *Advances in Computational Sciences and Technology*, vol. 8, n° 1, p. 29-34, 2015.

Harrag F., « Estimating the sentiment of arabic social media contents : A survey », *5th International Conference on Arabic Language Processing*, 2014.

Harrat S., Meftouh K., Abbas M., Hidouci W.-K., Smaili K., « An algerian dialect : Study and resources », *International journal of advanced computer science and applications (IJACSA)*, vol. 7, n° 3, p. 384-396, 2016.

Harrat S., Meftouh K., Abbas M., Smaili K., « Building resources for algerian arabic dialects », *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

Harrat S., Meftouh K., Smaïli K., « Machine translation for Arabic dialects (survey) », *Information Processing & Management*, 2017.

Hedar A. R., Doss M., « Mining social networks arabic slang comments », *IEEE Symposium on Computational Intelligence and Data Mining (CIDM)*, 2013.

Itani M. M., Zantout R. N., Hamandi L., Elkabani I., « Classifying sentiment in arabic social networks : Naive search versus naive bayes », *Advances in Computational Tools for Engineering Applications (ACTEA), 2012 2nd International Conference on*, IEEE, p. 192-197, 2012.

Joachims T., *Learning to classify text using support vector machines : Methods, theory and algorithms*, vol. 186, Kluwer Academic Publishers Norwell, 2002.

Kaseb G. S., Ahmed M. F., « Arabic Sentiment Analysis approaches : An analytical survey », 2016.

Khalifa K., Omar N., « A hybrid method using lexicon-based approach and naive Bayes classifier for Arabic opinion question answering », *Journal of Computer Science*, vol. 10, n° 10, p. 1961, 2014.

Khoja S., Garside R., « Stemming arabic text », *Lancaster, UK, Computing Department, Lancaster University*, 1999.

Korayem M., Crandall D., Abdul-Mageed M., « Subjectivity and sentiment analysis of arabic : A survey », *International Conference on Advanced Machine Learning Technologies and Applications*, Springer, p. 128-139, 2012.

Mataoui M., Zelmati O., Boumechache M., « A proposed lexicon-based sentiment analysis approach for the vernacular Algerian Arabic », *Research in Computing Science*, vol. 110, p. 55-70, 2016.

Mdhaffar S., Bougares F., Esteve Y., Hadrich-Belguith L., « Sentiment Analysis of Tunisian Dialect : Linguistic Resources and Experiments », *WANLP 2017 (co-located with EACL 2017)*, 2017.

Medhat W., Hassan A., Korashy H., « Sentiment analysis algorithms and applications : A survey », *Ain Shams Engineering Journal*, vol. 5, n° 4, p. 1093-1113, 2014.

Meftouh K., Bouchemal N., Smaïli K., « A Study of a Non-Resourced Language : The Case of one of the Algerian Dialects », *The third International Workshop on Spoken Languages Technologies for Under-resourced Languages-SLTU'12*, 2012.

Meftouh K., Harrat S., Jamoussi S., Abbas M., Smaili K., « Machine translation experiments on padic : A parallel arabic dialect corpus », *The 29th Pacific Asia conference on language, information and computation*, 2015.

Mohammad S. M., Turney P. D., « Crowdsourcing a word–emotion association lexicon », *Computational Intelligence*, vol. 29, n° 3, p. 436-465, 2013.

Saâdane H., Le traitement automatique de l'arabe dialectalisé : aspects méthodologiques et algorithmiques, PhD thesis, Grenoble Alpes, 2015.

Saâdane H., Guidere M., Fluhr C., « La reconnaissance automatique des dialectes arabes à l'écrit », *colloque international «Quelle place pour la langue arabe aujourd'hui*, p. 18-20, 2013.

Saâdane H., Habash N., « A conventional orthography for Algerian Arabic », *Proceedings of the Second Workshop on Arabic Natural Language Processing*, p. 69-79, 2015.

Sadat F., Kazemi F., Farzindar A., « Automatic identification of arabic dialects in social media », *Proceedings of the first international workshop on Social media retrieval and analysis*, ACM, p. 35-40, 2014.

Shoufan A., Alameri S., « Natural language processing for dialectical Arabic : A Survey », *Proceedings of the Second Workshop on Arabic Natural Language Processing*, p. 36-48, 2015.

Siddiqui S., Monem A. A., Shaalan K., « Sentiment analysis in Arabic », *International Conference on Applications of Natural Language to Information Systems*, Springer, p. 409-414, 2016.

Taboada M., Brooke J., Tofiloski M., Voll K., Stede M., « Lexicon-based methods for sentiment analysis », *Computational linguistics*, vol. 37, n° 2, p. 267-307, 2011.

L'analyse et l'annotation à base de FrameNet : contribution à l'étude contrastive des événements de mouvement en arabe et en anglais

Abdelaziz Lakhfif[*], Mohamed Tayeb Laskri[**]

[*] *LRSD, département d'informatique, Université Ferhat Abbas Sétif 1, Algérie.*
[**] *LRI, département d'informatique, Université Badji Mokhtar – Annaba, Algérie.*

abdelaziz.lakhfif@univ-setif.dz, laskri@univ-annaba.org

RÉSUMÉ*. Dans cet article, nous présentons une nouvelle approche computationnelle d'analyse et d'annotation de la langue arabe. L'approche proposée est fondée sur la théorie des frames sémantiques de* Fillmore *(Frame Semantics). Nous abordons la question de l'applicabilité de cette théorie à la langue arabe qui diffère typologiquement de l'anglais, langue sur laquelle la théorie a été fondée à l'origine. Nous allons aussi explorer l'utilisation de cette approche pour l'analyse sémantique de l'arabe, notamment, l'annotation en rôles sémantiques et son application à l'analyse contrastive des événements de mouvement et de déplacement en arabe et en anglais, en utilisant un outil développé spécialement afin de répondre aux spécificités de l'arabe et de satisfaire les principes du projet FrameNet. En dépit des différences entre l'arabe et l'anglais sur plusieurs aspects, allant du type d'écriture à la typologie de langue, les résultats de notre projet confirment, une fois de plus, la nature cross-langues de la théorie des frames sémantiques. Les résultats de ce travail sont fondés sur l'analyse d'un corpus constitué dans sa majorité des expressions du domaine des mouvements et déplacements.*

ABSTRACT*. In this paper, we describe a Frame Semantics based computational approach for Arabic language processing. We explore the representational of Frame Semantics approach to Arabic text semantics, the adaptability of Berkeley FrameNet database and the transferability of FrameNet tools for Arabic, a language that differ typologically from English. We describe our attempts to build an equivalent Arabic FrameNet and the use of such a semantic resource for Arabic text semantic analysis, representation and annotation. Here we present a frame based contrastive study of motion events expression in bilingual text (English-Arabic) using our FrameNet based tool for semantic annotation. Our study results confirm the cross-linguistic nature of Frame Semantics approach and the suitability of the theory for Arabic processing. The current work is based on an analysis of a representative corpus of motion events expressions.*

MOTS-CLÉS *: frames sémantiques, langue arabe, TAL, rôles sémantiques.*

KEYWORDS*: Frame Semantics, Arabic language, NLP, Semantic roles.*

1. Introduction

La théorie des *frames* ou des *cadres* (Baker, 2009) sémantiques (*frame semantics*) (Fillmore, 1982 ; Fillmore et Baker, 2010), a connu, depuis son apparition, un essor considérable et elle est devenue rapidement une source d'inspiration pour des centaines de travaux de recherche dans le domaine du TAL. La théorie de Fillmore, dont l'idée était largement inspirée des travaux de recherche sur les formalismes de représentation des connaissances dans les domaines de l'intelligence artificielle et de la psychologie cognitive, durant les années soixante-dix, représente une évolution naturelle de son idée originale sur la grammaire des cas (Fillmore, 1968). La théorie affirme que « *les gens comprennent, en grande partie, la signification des mots en se référant aux frames (sémantiques) qu'ils évoquent* » (Ruppenhofer *et al.*, 2016a). Par exemple, pour comprendre le sens du mot anglais '*Arrive'* (arriver), il faut en premier lieu avoir des connaissances sur le processus du mouvement, qui évoque principalement la notion de déplacement d'une entité (*Theme*) occupant une location initiale (*Source*), tout au long d'un chemin (*Path*), vers une location finale (*Goal*) et ainsi de suite (section 2). La théorie s'inscrit dans le domaine de la sémantique empirique qui accorde une grande importance à la relation entre la langue et l'expérience (Petruck, 1996), où un mot prend sens dans un contexte conventionnel interprété par le biais d'une structure cognitive (frame) évoquée dans l'esprit des interlocuteurs, plutôt qu'explicite (Baker, 2009). Du point de vue lexicographique, la théorie, cherche à associer des formes linguistiques à leurs structures cognitives (frames) sous-jacentes décrivant leurs usages validés par un usage langagier.

L'application pratique de la théorie est la base sémantique lexicale bien connue FrameNet (Baker *et al.*, 1998 ; Fillmore *et al.*, 2003), une ressource lexicale disponible en ligne conçue et élaborée sur la base des principes des frames sémantiques. La base FrameNet est construite à partir d'un ensemble de frames organisées hiérarchiquement et considérées comme étant des classes sémantiques regroupant des unités lexicales, ainsi que sur un réseau de relations frame à frame, fournissant la possibilité de représenter des scénarios larges (Burchardt, 2008) permettant ainsi de capturer le sens d'un discours composé de plusieurs phrases. Malgré le fait que la base Berkeley FrameNet (BFN) ait été élaborée à l'origine pour l'anglais, la nature « cross-langue » des frames (Boas, 2009 ; Fillmore et Baker, 2010) et la libre disponibilité de la base ont joué pour motiver d'autres chercheurs à réutiliser la base BFN et ses outils d'annotation pour l'analyse d'autres langues (Pitel, 2009 ; Boas, 2009), rendant ainsi FrameNet la ressource lexicale et sémantique la plus intéressante en termes d'utilisation cette dernière décennie.

L'approche des frames sémantiques a été expérimentée avec succès dans plusieurs axes de recherche dans le domaine du TAL, principalement dans la construction des

bases lexicales multilingues (Boas, 2002 ; Torrent *et al.*, 2014), l'alignement de différentes ressources lexicales sémantiques (Shi et Mihalcea, 2005 ; Ferrandez *et al.* 2010 ; Baker *et al.*, 2017), l'extraction de connaissances à partir de textes (Nuzzolese *et al.*, 2011 ; Søgaard *et al.*, 2015), la construction d'ontologies (Reed et Pease, 2015), l'analyse sémantique de surface (étiquetage en rôles sémantiques) (Gildea et Jurafsky, 2002 ; Erk et Padó, 2006 ; Johansson et Nugues, 2008 ; Das *et al.*, 2014 ; Lakhfif et Laskri., 2015 ; Hartmann *et al.*, 2017), la catégorisation automatique de textes (Moschitti, 2008), les systèmes questions-réponses (Narayanan et Harabagiu, 2004 ; Shen et Lapata, 2007), la reconnaissance des paraphrases (Padó et Erk, 2005) et l'implication textuelle (Burchardt *et al.*, 2007 ; Burchardt et Pennacchiotti,2017), la traduction automatique (Gimenez et Marquez, 2007 ; Wu et Fung, 2009, Lakhfif et Laskri, 2015 ; 2016), l'analyse des sentiments (Ruppenhofer et Michaelis, 2016b) et l'extraction d'événements (Agarwal *et al.*, 2014) dans les réseaux sociaux. L'approche a été expérimentée aussi dans des domaines spécialisés à l'image du projet « Kicktionary » (Schmidt, 2007) pour l'analyse des commentaires footballistiques, la génération de scènes visuelles à partir du texte (VigNet) (Coyne, *et al.*, 2010), le projet « BioFrameNet » (Dolbey *et al.*, 2006), une extension de BFN vers une ontologie biomédicale et le projet de représentation et d'annotation de textes dans le domaine juridique (Venturi *et al.*, 2009). Cependant, l'application de la théorie des frames sémantiques sur d'autres langues se heurte à des difficultés liées à l'applicabilité de l'outil d'annotation et d'affichage, conçu initialement pour l'anglais, sur d'autres langues (le chinois, l'arabe, etc.) et à la capacité des frames à représenter les divergences des langues dans la représentation linguistique des conceptualisations. Parmi les contraintes que l'approche de Fillmore doit résoudre dans un contexte multilingue, on trouve le phénomène de la divergence des langues (Dorr, 1993). La divergence dans l'expression d'un événement dans les langues se produit lorsque le sens sous-jacent est réparti sur différentes formes linguistiques, provoquant ainsi des transformations touchant un des aspects de l'expression tels que la structure, le lexique, la catégorisation des mots, etc. (Habash et Dorr, 2002). Dans cet article nous allons présenter les premiers résultats de l'adaptation de la théorie des frames sémantiques pour le traitement de la langue arabe, en nous intéressant, en particulier, à la conceptualisation des événements de mouvement et de déplacement en termes de frames (Ellsworth *et al.*, 2006 ; Ohara, 2007 ; Petruck, 2008) dans différentes langues (anglais, français et arabe). Nous allons montrer que la théorie des frames sémantiques dispose d'un mécanisme, fondé sur les relations *frame à frame*, pour la résolution de certaines divergences liées à la typologie des langues (Talmy, 1991).

Après une présentation concise de la notion des frames sémantiques (section 2), nous présentons dans la section 3, notre tentative de génération d'une base FrameNet pour l'arabe. Nous commençons par une brève introduction sur les caractéristiques de l'arabe. Ensuite, nous abordons la question de l'utilisation des ressources lexicales sémantiques

dans notre projet et l'application des principes de la théorie dans l'analyse et l'annotation de l'arabe moderne standard (MSA). La section 4 présente une étude contrastive de la conceptualisation des expressions des événements de mouvement selon le point de vue des frames sémantiques. Nous terminons par une conclusion dans la section 5.

2. Les frames sémantiques et le projet FrameNet

Une frame sémantique (figure 1) est une représentation schématique d'une situation, d'un objet, d'une relation ou d'un événement impliquant des participants sémantiques et regroupant tous les mots (FEE) qui peuvent évoquer cette situation dans une phrase. Chaque frame sémantique est associée avec un ensemble de participants sémantiques, appelés « éléments de la frame » (*frame elements*) (FE) (Pado et Pitel, 2007), qui représentent le sens véhiculé par les arguments syntaxiques du prédicat de la phrase. Fillmore et Baker (2010) attestent que la notion de frames est largement indépendante des langues, et qu'elle est plutôt proche de la notion de scripts (Schank et Abelson, 1977) et de la notion des frames proposée par Minsky (1975) qui représentent des situations stéréotypes et qui jouent un rôle important dans le processus de perception, de mémorisation et de raisonnement. Cependant, à la différence des ces anciens formalismes, qui ne tiennent pas compte du processus de la production langagière dans la mise en place de ses structures cognitives (Fillmore *et al*., 2003 ; Fillmore et Baker, 2010), les frames de Fillmore sont considérées selon une perspective linguistique et, par conséquent, visent à fournir un cadre descriptif de l'interface entre le sens et les formes linguistiques (support du sens véhiculé).

Arriving	
Définition	An object Theme moves in the direction of a Goal. The Goal may be expressed or it may be understood from context, but its is always implied by the verb itself. Our visitors ARRIVED yesterday. Amy ARRIVED home from school early one afternoon. The senator ARRIVED to a standing ovation.
Core FE	1. Goal [Goal] **Semantic Type :** Goal 2. Theme[Thm] **Semantic Type :** Sentient
UL anglaises	*approach.n, approach.v, arrival.n, arrive.v, come.v, crest.v, descend_(on).v, enter.v, entrance.n, entry.n, get.v, influx.n, make it.v, make.v, reach.v, return.n, return.v, visit.n, visit.v*

Arriving	
UL arabes	إقْتَرَبَ.**v** ، دَنا.**v** ، بَلَغَ.**v** ، وَصَلَ.**v** ، إنْتَهى إلى.**v** ، دَخَلَ.**v** ، رَجَعَ.**v** ، أقْبَلَ.**v** ، أتى.**v** ، آبَ.**v** ، إقْتَحَمَ.**v** ، تَوجَّه إلى.**v** ، زارَ.**v** ، عادَ.**v** ، نَزَلَ بـ.**v** ، وَفَدَ.**v** ، قَدِمَ.**v** ، جاءَ.**v** ، إنْتَقَلَ إلى.**v**

Figure 1. *Une illustration sommaire d’une entrée de la base* FrameNet *: **‘Arriving’***

La notion de la frame est à la base de notre compréhension des mots de la langue où un mot est associé à un cadre descriptif de son usage (frame sémantique). L’objectif principal de l’approche des frames sémantiques est la description de différentes significations véhiculées par un lexème, en s’appuyant sur des structures cognitives (frames), où chaque frame encapsule un de ses sens véhiculés, ainsi que les constructions grammaticales possibles en énumérant toutes les réalisations syntaxiques supportées par un usage langagier des rôles sémantiques joués par les différents arguments syntaxiques liés au lexème donné (le prédicat). Ces rôles sémantiques (FE) représentent les participants sémantiques spécifiques à la frame. Les FE nécessaires pour la compréhension de la situation sont des FE noyaux ‘Cores’. Les frames sont reliées entre elles par des relations frame à frame (*inheritance*, *using*, *sub*frame, *precedes*, *causative of*, *etc.*) (figure 2).

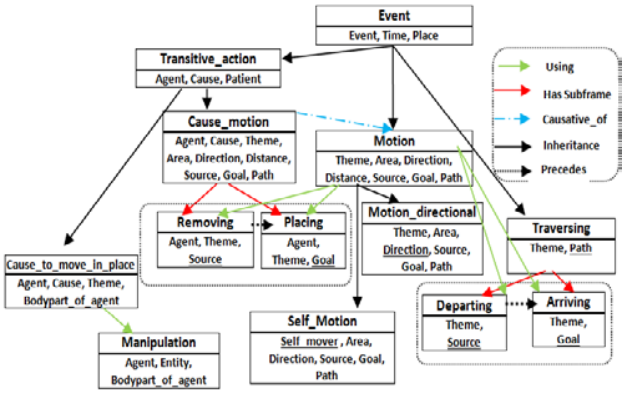


Figure 2. *Une représentation partielle des relations frame à frame*

Ces relations (BFN contient 1 687 relations) sont d’une importance capitale pour les applications à base de frames qui nécessitent des mécanismes d’inférence. Revenons sur notre exemple initial (section 1) concernant le processus mental de compréhension du

mot anglais *'Arrive'* (arriver), l'image cognitive sous-jacente est organisée linguistiquement autour de la notion des frames. Dans notre cas, la frame caractérisant partiellement cette situation est appelée '*Arriving*' et les participants sémantiques jouant les rôles thème (*Theme*) et but (*Goal*) sont indispensables à l'instanciation de la frame, parmi d'autres participants secondaires. Cette frame '*Arriving*' avec la frame *'Departing'* sont des sous-frames de la frame complexe *'Traversing'* qui utilise (*using*) la frame *'Motion'*, qui de son côté caractérise un mouvement plus général que la frame *'Arriving'*. La frame *'Arriving'* représente le mouvement d'un point de vue de la destination (*Goal-profiling*), alors que la frame *'Departing'*, qui est reliée à *'Arriving'* avec la relation *'Precedes'*, représente le côté source (*Source-profiling*).

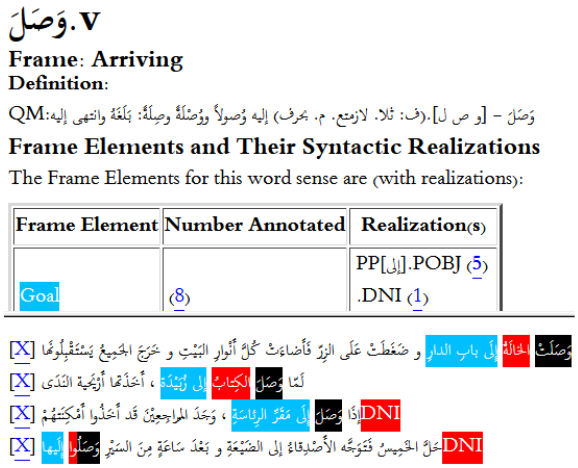
La frame *'Arriving'* peut être évoquée par un groupe de mots partageant les mêmes caractéristiques sémantiques essentielles par rapport à l'image cognitive (frame), tels les verbes : *arrive* (arriver), *reach* (atteindre), *enter* (entrer), *visit* (visiter), etc. En général, un mot peut avoir plusieurs sens et, par conséquent, peut évoquer différentes frames. Le verbe '*reach'* (atteindre) est un mot avec plusieurs sens et, par conséquent, il appartient à plusieurs frames. Parmi les sens du verbe '*reach'* (atteindre) un qui correspond à la frame *'Contacting'*, dans laquelle un communicateur '*Communicator*' adresse une '*Communication*' à un destinateur '*Addressee*' *via* une adresse particulière '*Address*' (courriel, téléphone, etc.). Le reste des sens sont liés aux frames *'Path_shape' et 'Body_mouvement'* respectivement. Dans ce cas de figure, la paire (lemme, frame évoquée) correspond à une unité lexicale (UL).
Le développement continu de la base BFN permet d'accroître le niveau de sa couverture linguistique, elle contient actuellement (juin, 2018) plus de 1 087 frames lexicales caractérisant plus de 13 640 UL. La base comporte aussi plus de 200 000 ensembles d'annotations de phrases extraites à partir du corpus « *British National Corpus* »(BNC), validant l'usage langagier des mots.

2.1. *Procédure et outils d'annotation*

La procédure d'annotation en frames sémantiques au sein du projet BFN consiste à associer les formes linguistiques avec leurs composants sémantiques (FE) de la structure cognitive (Fillmore et Baker, 2010).

Le projet BFN dispose d'un système client-serveur développé au centre ICSI Berkeley pour la création des frames, l'ajout des UL, et l'annotation en rôles sémantiques. Les données du système sont stockées dans une base de données relationnelle (MySql). L'affichage des informations sur les frames et sur les annotations est assuré par un outil visuel à base de technologies Web. BFN a suivi une stratégie d'annotation qui consiste à procéder frame par frame dans la création des UL et l'annotation des phrases, tout en limitant le nombre des types de phrases (PT) et des

fonctions grammaticales (GF), afin de faciliter la tâche d'annotation. L'annotation de textes consiste à sélectionner pour chaque UL, des exemples de son utilisation sous forme de phrases à partir du corpus BNC qui attestent son usage et son adéquation avec le sens caractérisé par la frame. Une fois la phrase sélectionnée et importée dans l'outil d'annotation, l'opération d'annotation en rôles sémantiques va consister à regrouper en plusieurs couches et pour chaque phrase annotée l'étiquette représentant *l'élément de la frame* (FE) avec sa *fonction grammaticale* (GF), et son *type de phrase* (PT) sous la forme d'un triplet. Le processus d'annotation offre la possibilité d'ajouter d'autres couches pour les informations non couvertes par les couches précédentes (prédicats, relatifs, etc.).


وَصَلَ.V

Frame: **Arriving**

Definition:

QM:وَصَلَ – [و ص ل].(ف: ثلا. لازمتع. م. بحرف) إليه وُصولاً ووُصْلَةً وصِلَةً: بَلَغَهُ وانتهى إليه

Frame Elements and Their Syntactic Realizations

The Frame Elements for this word sense are (with realizations):

Frame Element	Number Annotated	Realization(s)
Goal	(8)	PP[إلى].POBJ (5) .DNI (1)

[X] وَصَلَتْ الحَالَةُ إلى بابِ الدارِ و ضَغَطَتْ عَلَى الزِرّ فَأَضاءَتْ كُلَّ أَنْوارِ البَيْتِ و خَرَجَ الجَمِيعُ يَسْتَقْبِلُونَها

[X] لَمّا وَصَلَ الكِتابُ إلى زَيْنَةَ ، أَخَذَتْها أَرْيَحِيَة النَدَى

[X] إذَا DNI وَصَلَ إلى مَقَرِّ الرِئاسَةِ ، وَجَدَ المَراجِعِيْنَ قَد أَخَذُوا أَمْكِنَتَهُمْ

[X] حَلَّ الخَمِيسُ فَتَوَجَّه الأَصْدِقاءُ إلى الضَيْعَةِ و بَعْدَ سَاعَةٍ مِنَ السَيْرِ وَصَلُوا إليها DNI

Figure 3. *Une partie de la table de réalisation pour " وَصَلَ (arriver)" « arrive »*

La figure 3 montre une partie de la table récapitulative des réalisations syntaxiques des rôles sémantiques (FE) *pour l'unité lexicale arabe وَصَلَ (arriver) « arrive »* avec des exemples de son usage langagier, ainsi que sa définition extraite à partir d'un dictionnaire spécialisé (*AL- Qamoos AL-Muheet* ou *« Le Dictionnaire complet »*, Al-Firoz-Abadi (1329-1415 AD)). Ce rapport offre ainsi aux utilisateurs un accès rapide aux informations sur la correspondance entre les formes syntaxiques et les composants sémantiques de la phrase.

Les données du projet BFN sont réparties entre deux bases différentes. La base lexicale qui regroupe toutes les informations sur les frames et les différents types de relations entre ces frames ainsi que les informations sur les UL et la base d'annotations

qui regroupe toutes les phrases annotées avec les sous-corpus à partir desquels ces mots sont extraits. Ces annotations riches en informations lexicographiques et en valences de mots sont stockées avec l'ensemble des valences de l'UL dans un fichier XML. Ce fichier XML contient aussi des données sur les modèles de valences des prédicats (UL), générées automatiquement durant l'opération d'annotation et qui peuvent être exploitées par des applications TAL.

2.2. *La dimension cross-langue de la théorie des frames sémantiques*

Suite aux succès de la théorie des frames sémantiques et de sa base lexicale FrameNet, l'équipe de Berkeley essaye actuellement de répondre aux questions liées à l'extension du projet BFN vers d'autres langues, notamment vers les langues dont la typologie est différente de celle de l'anglais, telles que l'espagnol, le français et les langues sémitiques (l'arabe, l'hébreu, le maltais, etc.). Ainsi, les chercheurs s'interrogent pour savoir dans quelle mesure les structures cognitives, ou frames sémantiques développées initialement pour l'anglais, sont universelles, et s'il existe des frames spécifiques à chaque langue ou à chaque classe de langues. Fillmore (1982) et d'autres chercheurs (Boas, 2005 ; Burchardt *et al.,* 2009) ont défendu la nature cross-langue de la théorie en avançant le fait que la description conceptuelle des frames, qui sont dans la majorité des cas des structures cognitives, est indépendante de la langue. Cependant, il est prévisible que certaines frames de la base BFN sont liées à l'anglais, principalement dans le cas de la description de certaines situations ou de certains événements liés à la culture et aux traditions américaines. Nous allons montrer (section 3.1) que l'utilisation et l'adaptation de la théorie des frames sémantiques pour l'arabe, langue sémitique qui diffère largement de l'anglais, doivent prendre en considération certaines caractéristiques linguistiques non rencontrées dans les langues indo-européennes.

3. Un FrameNet pour la langue arabe

Le succès du projet FrameNet pour l'anglais a encouragé le lancement des projets similaires pour d'autres langues, à l'image des projets pour l'espagnol (Subirats et Petruck, 2003), le japonais (Ohara *et al.,* 2004), le chinois (You et liu, 2005), l'allemand (Burchardt *et al.,* 2009), le brésilien (Salomão, 2009), l'italien (Johnson et Lenci, 2011), le danois (Bick, 2011), le coréen (Kim *et al.,* 2016), le finnois (Lindén *et al.*, 2017), et enfin le français. En effet, récemment ASFALDA, un FrameNet pilote pour le français (Candito *et al.,* 2014), a été mis en ligne. En utilisant l'outil SALTO (Burchardt *et al.,* 2006), le projet ASFALDA a opté, en première étape, pour une démarche ciblant

l'analyse et l'annotation d'un corpus orienté (Djemaa *et al.*, 2016) couvrant les sept domaines de discours (communications, transactions commerciales, relations spatiales, etc.) les plus discutés au sein de la communauté FrameNet. Cependant, ce genre de tentatives pour les langues sémitiques, spécialement pour l'arabe, n'a pas atteint le niveau attendu par rapport à cette importante ressource sémantique. En effet, la première étude visant l'application de la théorie de Fillmore sur une langue sémitique a été lancée par Petruck (2008), et, récemment, un FrameNet pour l'hébreu a vu le jour grâce à Hayoun et Elhadad (2016). Le projet suit les principes de développement adoptés par BFN, en utilisant la même base des frames avec leurs définitions, FE et toutes les relations frame à frame.

En revanche, très peu de travaux de recherche se sont intéressés à l'analyse de l'arabe sous la lumière des frames sémantiques. En effet, le travail de Sharaf et Atwell (2009) représente la première tentative sérieuse de l'utilisation de la théorie des frames sémantiques pour l'analyse de l'arabe en essayant de développer un prototype FrameNet explorant les valences des verbes extraits du Coran. De son côté, Alshehri (2014) a effectué une analyse contrastive (anglais et arabe) sur un petit corpus de textes contenant cinq verbes (*walk, run, fly, climb* et *crawl* et leurs équivalents en arabe) de la frame *Self_motion* dans le but de déterminer la capacité et la généralité de la structure de la frame à représenter les deux langues. En effet, le travail de Lakhfif et Laskri (2015) est la première tentative, grandeur nature, de la génération d'un *Arabic FrameNet.* Ce projet, qui a débuté en 2011, vise à construire une base lexicale sémantique pour l'arabe à partir de la base originale, en utilisant la ressource *Arabic WordNet* (Black *et al.*, 2006) (AWN) comme passerelle et en suivant la méthodologie de transfert recommandée par BFN. Le processus d'annotation consistait à extraire des phrases attestant l'usage des UL induites en se fondant sur le corpus CCA *Corpus of Contemporary Arabic* (Al-Sulaiti et Atwell, 2006), en plus d'autres sources scolaires et religieuses. La collection des exemples et l'annotation semi-automatique d'informations morphosyntaxiques ont été effectuées par trois personnes qualifiées en linguistique arabe, cependant que l'annotation en rôles sémantiques était faite par un expert. Cette base sémantique a été utilisée pour l'analyse, la représentation sémantique et l'annotation (Lakhfif, *et al.*, 2013) des phrases en arabe (ASM). Ce projet (thèse de doctorat) entre dans le cadre du développement d'un outil de traduction automatique de l'arabe vers la langue des signes algérienne (LSA) (Lakhfif, 2016).

Langue	# Frames lexicales	# Unités lexicales	# Ensembles annotés
Anglais (Berkley FrameNet)	1 087	13 640	174 022
Espagnol	–	1 268	11 000
Suisse	1 200	38 028	9 000
Japonais	565	8 500	60 480
Finnois	938	6 639	40 721
Brésil	179	196	12 000
Hébreu (Hayoun et Elhadad)	167	3 000	500
Arabe (Lakhfif et Laskri)	> 600	> 10 000	> 3 000
DiCoInfo : A Framed Version (Ghazzawi)	57	106	–

Tableau 1. *Comparaison de l'état de certains projets FrameNet nationaux*

Dans sa thèse de doctorat, Ghazzawi (2016), tout en s'appuyant sur la ressource BFN, a ajouté à la base DiCoInfo (L'Homme, 2008) la traduction arabe des mots appartenant au domaine de l'informatique. Cependant, malgré le fait que ce projet adopte l'approche de Fillmore, l'annotation en frames sémantiques est un peu différente de celle du projet BFN, ainsi que la codification des informations annotées. Le tableau 1 donne un résumé de la couverture de certains projets FrameNet nationaux, dont la majorité ont adopté une approche étendue (Vossen, 1999) dans leurs processus de construction, gardant ainsi l'intégralité de la base originale pour l'anglais (BFN), au vu des difficultés de développement de ce genre de ressource qui nécessite des efforts multidisciplinaires et qui consomme du temps.

3.1. *L'arabe et les technologies du TAL*

La langue arabe est classée à la quatrième position parmi les dix premières langues utilisées sur Internet (http ://www.internetworldstats.com/stats7.htm). Avec l'émergence des réseaux sociaux comme facteur d'influence sociale, économique, et politique, le nombre d'utilisateurs de l'arabe (standard, moderne et dialectes) a été multiplié par trois, en moins de dix années (on est passé de 60 millions d'utilisateurs, en 2008, à 197 millions, en décembre 2017). L'importance de l'arabe est accentuée par le nombre de pays (22 pays, et plus de 420 millions d'habitants) qui l'utilisent comme langue officielle. En plus des pays arabes, l'arabe (classique) est la langue du Coran, livre sacré de l'islam, religion regroupant plus de 1,7 milliard de personnes dans le monde. L'arabe est une langue sémitique caractérisée par une riche morphologie inflectionnelle et un système de dérivation à base d'une douzaine de schèmes. L'arabe est une langue à sujet

pronominal vide « *pro-drop* », où la morphologie du verbe intègre un pronom sujet et peut, dans certains cas, intégrer aussi un pronom objet (encliticisé). En général, l'arabe est une langue cliticisante, où les prépositions et les conjonctions peuvent être aussi procliticisées au mot arabe. Par exemple (tableau 2), le mot arabe (*fa-jA-at-hu*) consiste en la conjonction "ف(fa)" "puis- et", un verbe "**جاءَ** (*jAa*)venir" conjugué dans le passé avec une inflexion pour un pronom sujet à la première personne du féminin du singulier et avec une inflexion pour un pronom objet à la troisième personne du singulier masculin.

Arabe : **فَجاءَتْهُ**
Translittération : *fa-jA-at-hu*
POS *: Conj--V.$_{pass}$--Pro.$_{1sf}$(subj)--Pro.$_{3sm}$(obj)*
Glose-fr *: puis--venir--elle--lui* (*puis elle vint à lui*)

Tableau 2. *La complexité de la morphologie arabe*

La question du transfert des outils du projet BFN pour la représentation, l'analyse et l'annotation de l'arabe représente un défi majeur, vu les différences entre les deux langues, allant de la transcription, l'orientation, la flexion et l'agglutination (Ryding, 2005 ; Habash, 2010), en passant par les contraintes sur l'ordre des mots dans la phrase, jusqu'au système de dérivation. Parmi les traits linguistiques décisifs dans le traitement automatique de l'arabe, on note la terminaison et les voyelles de la fin des mots qui contribue à la définition des cas syntaxiques. Cependant, si l'on regarde la flexibilité dans la distribution des arguments syntaxiques au sein de la phrase arabe, on observe que la position des arguments syntaxiques (agent, objet, etc.) n'est pas un critère grammatical déterministe. L'arabe dispose aussi d'un mécanisme dérivationnel lui permettant de générer des catégories nominales ((le participe actif) اسم الفاعل, (forme infinitive)مصدر , et (le participe passif) اسم المفعول, etc.) qui peuvent occuper des fonctions grammaticales importantes (adjectif, adverbe, etc.), et jouer les mêmes rôles que le verbe, comme avoir un sujet, des objets, etc.

3.2. *Génération d'une base FrameNet pour l'arabe*

L'émergence, vers la fin de ce dernier siècle, des ressources lexicales et sémantiques, telles que Princeton WordNet (Fellbaum, 1998) (PWN), FrameNet et VerbNet (Kipper-Schuler, 2005), représente l'une des avancées les plus importantes dans le domaine du TAL. Ces bases lexicales sémantiques sont devenues indispensables dans les applications nécessitant une représentation profonde de la sémantique de la phrase (Fellbaum *et al.,* 2007), à l'image de la compréhension du texte, de la traduction

automatique, etc. Cependant, la seule ressource lexicale sémantique, disponible actuellement, pour l'arabe est la base « Arabic WordNet » (AWN) (Black et *al*., 2006). AWN est une base lexicale pour les mots arabes, inspirée du réseau sémantique PWN et qui suit la méthodologie de développement de la base EuroWordNet (Vossen, 1998). L'avantage de la base AWN est dû à sa liaison avec PWN (3.0) en termes d'équivalence entre synsets, permettant ainsi la projection des classes de l'ontologie SUMO des mots anglais vers les mots arabes. Parmi ces relations (tableau 3), la relation « *Hyponymy* » qui relie deux mots par la relation taxonomique (*has_hyponym*) et la relation d'implication « *sub_event* » liant un verbe à un autre verbe décrivant un sous-événement de l'événement du premier verbe.

Algorithme 1 – L'algorithme d'induction des LU arabes

Entrée : **une base des alignements** (lemme, PoS, des **LU_Synset** [ID, Frame + PWN_Mots])
/*
enSSID :le synset ID (SSID) anglais ; **arSSID** : le synset ID arabe
aI_LU_SS : alignement d'un LU avec des synsets WN3.0
L_AW : la liste des LU arabes induites à partir de AWN : lemme, forme, pos, LU anglais, sumo, relation
L_ESS : liste des enSSIDs ; **L_ASS** : liste des : arSSID, sumo, AW relation
/
L_AW = {} /*initialiser la liste des LU arabes induites à partir de AWN */
Pour chaque **alignement faire**
Pour chaque **aI_LU_SS faire**
L_ESS = Lire_tous_les_**enSSID**_de_(**lemme**) /* tous les SSID anglais*/
Ajouter tous_les_**enSSID**_ similaires à **L_ESS** /* relation WN '***similar'***/
en_SUMO = Lire_le_concept SUMO du **enSSID**
Pour chaque **enSSID dans L_ESS faire** /*synsetid anglais*/
arSSID = Lire le SSID arabe équivalent dans AWN
ajouter tous les mots de **arSSID dans L_AW** /*mots arabes */
ar_SUMO = le concept SUMO du synset arabe
Si(**PoS** = 'v' ou Pos = 'n') **Alors** */*la relation **related_to***/*
| **Ajouter** les SSID arabes reliés à **arSSID** avec la relation '***related_to***' à la liste **L_ASS**
Fin Si
Ajouter les SSID arabes reliés à **arSSID** avec la relation '***has_derived***' à la liste **L_ASS**
Ajouter les SSID arabes reliés à **arSSID** avec la relation '***has_hyponym***' à la liste **L_ASS**
Si(**PoS** = 'v') **Alors** /*la relation 'verb_group', 'sub_event' */
| **Ajouter** les SSID arabes reliés à **arSSID** avec la relation '***verb_group***' à la liste **L_ASS**
| Ajouter les SSID arabes reliés à **arSSID** avec la relation '***sub_event***' à la liste **L_ASS**
Fin Si
Pour chaque **arSSID dans L_ASS faire** /*synsetid arabes, sumo, relation*/
| **Lire** le concept SUMO de **arSSID**
| **Si** (**ar_SUMO** subsume le concept SUMO de **arSSID**) **Alors**

| | **ajouter** tous les mots du sysnset **arSSID dans L_AW**
| | /*lemme, forme, pos, LU anglais, sumo, relation WN*/
| **Fin Si**
Fin Pour
Fin Pour
Fin Pour
Fin Pour
Sauvegarder la liste **L_AW** dans un Fichier XML

Ces relations parmi d'autres relations sont utilisées dans notre algorithme (algorithme 1) d'induction des unités lexicales pour construire un FrameNet pour l'arabe. La création manuelle d'une base similaire à BFN s'avère une tâche fastidieuse pour les langues moins dotées en ressources lexicales et sémantiques. Dans notre projet, en plus de la traduction manuelle d'un ensemble de Frames (*100* Frames) et les unités lexicales qui les appartiennent à partir de la base originale BFN, nous avons opté pour une approche étendue (Vossen, 1999), en utilisant la ressource AWN comme une source dans notre algorithme (algorithme 1) d'induction de la base FrameNet pour l'arabe (Lakhfif et Laskri, 2015 ; Lakhfif, 2016). Comme les ensembles de mots AWN sont traduits et liés directement aux ensembles des mots de la base PWN, nous avons exploité un algorithme d'alignement (Ferrandez *et al.* 2010) qui induit les UL de la base BFN (version 1.3) à partir des synsets de la base PWN (version 3.0), en se fondant sur les relations ontologiques telles que les relations "*equivalent*", "*related_to*", "*verb_group*".etc. (figure 4).

Relations	Synsets	Mots arabes	Équivalents anglais
sub_event	balaEa_v1AR	بَلَع	***swallow***
sub_event	maDaga_v1AR	مَضَغَ	***chew***
related_to	>akol_n1AR	أَكْل	***food***
verb_group	>akol_v2AR	أَكَلَ,تَنَاوَلَ الطَّعَام	***eat***
has_hyponym	qaDama_v1AR	قَضَمَ	***nibble***

Tableau 3. *Les relations associées au synset : >akala_v1AR 'أَكَلَ-eat'*

Wn30synset **ID** = 2005948 **PoS** = verb **corrélation** = **1.09**		
Synsetid = arrive_v1EN **sumo** = Motion + **Members** *: arrive, get, come* **Gloss** : reach a destination ; arrive by movement or progress ; …….		
frame *= Arriving* **ID** *= 940* **lemma** *= arrive* **PoS** *= v* **Sumo** = Motion +		
Projection		
Relation	**Lemmes arabes**	**SynsetID**
equivalent	وَصَلَ	waSala_v4AR
related_to	مَجِيء – قُدُوم , وُصُوْل	quduwm_n1AR, wuSuwol_n1AR

Figure 4. *Une entrée de l'induction des lemmes arabes pour "Arrrive.v"*

Cet algorithme, fondé sur une mesure de ressemblance sémantique entre les mots, rapporte une précision de 77 %, ce qui lui permet d'obtenir le meilleur score parmi les algorithmes d'induction proposés jusqu'à présent. Les UL induites automatiquement sont passées par une phase de vérification manuelle, afin d'éliminer les unités qui n'ont pas d'exemple d'utilisation en arabe, attestant sa cohérence avec la frame d'appartenance. Une partie de l'évaluation de notre projection ainsi que le nombre d'UL dans la base originale BFN sont présentés dans le tableau 4 qui montre aussi le nombre d'UL dans la base BFN utilisées dans la projection et le nombre d'UL générées pour l'arabe à partir des AWN respectivement. Il est à noter que malgré le fait que la taille de la base BFN soit plus large de treize fois par rapport à la base AWN (tableau 5), l'induction valable des UL arabes a atteint un rapport avoisinant la moitié de la taille de la base BFN. Cela ne veut pas dire que notre alignement couvre 50 % de la base BFN (seul le tiers (32 %) des UL dans la base BFN avait des équivalents dans AWN), mais plutôt qu'il donne un indice de la richesse du lexique arabe.

Catégorie de frames	# Frames	# UL (verbes) dans BFN	# (En) UL utilisées	# (Ar) UL	Précision
Mouvement	35	765	202	689	0,687
Causation	10	189	60	188	0,534
Activités	5	30	12	43	0,699
Frames vérifiées	80	1 523	484	1 490	**0,711**

Tableau 4. *La précision de l'induction par catégorie de Frames (Lakhfif, 2016)*

	# lemmes	# verbes	# noms	# adjectifs	# autres
PWN 3.0	155 287	13 600	81 000	19 000	3 600
FrameNet	13 640	5 200	5 558	2 396	486
AWN	11 256	2 525	7 960	500	271
Ar. FrameNet	10 273	4 500	5 530	232	11

Tableau 5. *Comparaison de la couverture de différentes ressources sémantiques*

3.3. *L'analyse et l'annotation de l'arabe en frames sémantiques*

La base FrameNet offre des informations sur les mots (UL) anglais et sur les frames sémantiques qui les caractérisent sous forme de données (XML) exploitables par les applications TAL. Parmi les informations utiles offertes par cette base, un nombre important de phrases extraites du corpus BNC sont annotées avec la triple information (FE, GF et PT). Ces informations illustrent, avec un niveau de détail sans précédent (Fillmore *et al.*, 2003), comment ces rôles sémantiques sont exprimés à travers des compléments et des modifiants du prédicat déclencheur de la frame. À cet effet, le projet offre une base lexicale avec un outil logiciel d'annotation et de visualisation à base des technologies XML. L'outil d'annotation BFN a été réutilisé et porté avec succès par plusieurs projets d'annotation en frames sémantiques pour des langues indo-européennes. Cependant, vu les divergences entre l'arabe et l'anglais, la réutilisation de cet outil pose plusieurs problèmes techniques, liés principalement à l'écriture des caractères et des mots arabes (langue cursive), rendant l'adaptation des outils de BFN pour la visualisation et l'annotation de l'arabe inappropriée. À cet effet, et afin de répondre aux défis soulevés par les spécificités de la langue arabe (transcription, orientation, flexion, agglutination, etc.), nous avons opté pour le développement d'un outil d'analyse et d'annotation de textes arabes (Lakhfif *et al.*, 2013), un système multiniveau permettant l'analyse lexico-morphologique, syntaxique et sémantique de l'arabe (ASM). Le système intègre en cascade l'analyseur morphologique AraMorph (Buckwalter, 2002) et la ressource AWN dans un système d'analyse fournissant plusieurs couches d'informations (PoS, grammaticale, sémantique). L'intégration de ces outils et de ces ressources offre une riche description des phrases arabes à l'image d'AWN qui ajoute des informations ontologiques sur les mots et FrameNet qui fournit des informations sémantiques sur les arguments du prédicat. Ces informations et d'autres informations utiles sont stockées dans un fichier XML, compatible avec le système de codification du projet BFN, et peuvent être exploitées par d'autres applications TAL tierces.

3.3.1. *Traitement de certaines caractéristiques de l'arabe*

Dans notre application de l'approche des frames sémantiques sur l'arabe, et afin de prendre en charge certaines spécificités linguistiques, nous avons opté pour des solutions qui préservent l'intégrité de l'approche adoptée par BFN et qui prennent en charge les caractéristiques propres de l'arabe. L'un des paris concerne l'annotation en rôles sémantiques des pronoms incorporés dans la morphologie du verbe (inflexion). Ces pronoms occupent généralement la position de sujet et dans certaines constructions peuvent aussi occuper la position de complément d'objet. Ce phénomène linguistique, très répandu dans certaines langues à sujet nul (pro-drop) telles que l'espagnol, l'italien et les langues sémitiques, se sert de la richesse morphologique afin de réaliser certaines constructions complexes. En arabe, le pronom sujet caché joue un rôle important dans la résolution de l'anaphore et de la référence dans certaines phrases complexes où le sujet est omis (CNI). Par exemple, dans le verbe '-شرَعنا' (tableau 6, figure 5), le pronom sujet 'nA-نا' incorporé dans le verbe correspond au pronom pronominal non réalisé morphologiquement 'nous – نحن'. Ce pronom contient des informations sur la personne, le nombre et le genre (1PM). Dans notre projet, dans le cas d'absence de sujet lexical dans une phrase, nous avons opté pour l'ajout d'une étiquette grammaticale 'SBJp' pour 'sujet pro-drop' pour marquer ce type de pronom explicitement avec l'étiquette sémantique adéquate (rôle sémantique). Cette solution nous a permis de garder des traces sur les arguments dans la phrase. Le deuxième cas concerne l'annotation en FE des constituants syntaxiques non locaux dans les constructions grammaticales à montée (*raising*) et à contrôle. Dans ces constructions à deux prédicats (un prédicat gouverne un autre), chaque prédicat évoque une frame différente, et l'argument du premier prédicat est aussi considéré comme un argument pour le deuxième prédicat (cible). Dans l'exemple ci-dessous (Ruppenhofer *et al.*, 2016a), '*John*' est aussi un argument du prédicat '*retaliate*' et, par conséquent, doit avoir le rôle sémantique '*Offender*'.

– We **expect** [John **Avenger**] to **retaliate** [against us **Offender**] [INI **Punishment**] [DNI **Injury**].

وَ [Activity_start] شَرَعـنا Target [Agent :pr-drop-1MP] في [Self_motion] السَّيْرِ Target [نَحْوَ ميريبيت هاوس Goal] [Activity
Glose : *et nous commençons à marcher vers 'Merripit House'*
we started to [walk Self_motion] [to Merripit House Goal]

Tableau 6. *Une clause avec une structure à contrôle* (The Hound of the Baskerville)

Dans BFN, la fonction grammaticale 'externe' (GF : Ext) est attribuée aux deux arguments ('*John*' et '*We*') des deux prédicats. Cette démarche pose des problèmes d'ambiguïté pour les applications TAL qui utilisent les données BFN dans des

mécanismes d'inférence. Dans notre projet, et afin d'éviter cette ambiguïté, un argument est considéré uniquement vis-à-vis du prédicat qu'il gouverne, tout en étendant les fonctions grammaticales vers les constituants pronoms, attachés morphologiquement au verbe, et qui occupent la position syntaxique 'sujet' ou 'objet'. Cette solution donne une distinction claire entre les rôles sémantiques dans ce type de construction.

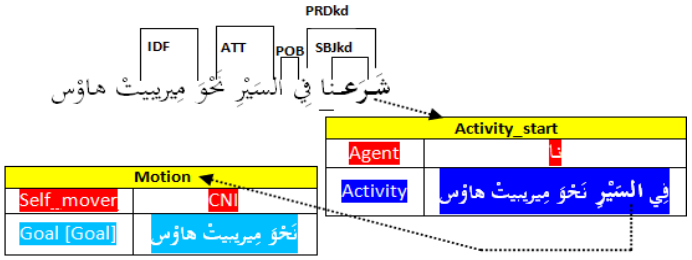


Figure 5. *Conceptualisation en frames d'une clause arabe (sujet nul, éléments non locaux)*

Dans l'exemple de la figure 5, le prédicat (la marche)السَّيْرِ déclenche la frame '*Self_motion'*, et son rôle sémantique '*Self_mover'* doit être réalisé par le sujet du prédicat de la construction à contrôle, du verbe appartenant au groupe des verbes de rapprochement et de commencement (أفعال الشروع و المقاربة). La liaison entre le sujet (étiqueté par SBJkd c'est-à-dire sujet des verbes de rapprochement et de commencement) du premier prédicat '*à commencer*- شَرَع' et le sujet du prédicat cible (*la marche)* السَّيْرِ est facilement établie en se fondant sur les informations d'agrément du pronom sujet caché (SBJp) qui fait référence sémantiquement à un argument non local. Le troisième cas de différence concerne l'annotation des verbes à préposition. Dans BFN, un verbe à préposition (*take off, take up*, etc.) est considéré avec sa particule comme une expression polylexicale '*multiword'* et, par conséquent, cette particule sera marquée comme une partie du verbe dans la couche '*Target'*. Cependant, dans notre projet, et en suivant les principes de la grammaire arabe, la préposition est marquée avec le syntagme prépositionnel qu'elle gouverne. Dans notre cas, l'association de la préposition avec le verbe dans l'annotation, génère des incohérences dans la description syntaxique du syntagme.

3.3.2. *L'annotation de corpus*

Suite aux succès des projets d'annotation de textes pour l'anglais, plusieurs tentatives ont été faites pour l'annotation de l'arabe. Les projets à l'image de Penn Arabic Treebank (PATB) (Maamouri *et al.* 2004), Prague ADTB (Hajic *et al.*, 2004), CATiB (Habash et Roth, 2009) et The Quranic Arabic Dependency Treebank (QADT) (Dukes *et al.* 2010) représentent les projets ayant le plus d'impact sur les travaux de recherche sur l'arabe. Cependant, à l'exception du projet Arabic PropBank (Diab *et al.*, 2008) qui

s'intéresse à l'annotation de la sémantique des arguments du prédicat, on note que l'annotation sémantique des corpus en langue arabe n'a pas encore attiré l'attention méritée. Dans notre projet d'annotation en frames sémantiques, le processus d'annotation de corpus suit les mêmes procédures de collection et d'organisation des données adoptées par le projet BFN qui est décrit en détail dans (Ruppenhofer *et al.*, 2016a). Pour chaque UL, les phrases contenant l'entrée avec le sens adéquat sont sélectionnées à partir de plusieurs sources. Comme notre objectif de l'annotation est de fournir une riche description de textes selon le niveau morpholexical, syntaxique ou sémantique intéressant pour les applications TAL (figure 6), notre outil semi-automatique d'analyse fournit plusieurs types de descriptions selon ces trois niveaux. Au début d'analyse, notre outil établit automatiquement une analyse complète de la phrase en entrée et propose tous les résultats possibles pour chaque lexème (ou groupe de lexèmes), dans chaque niveau d'analyse, selon le degré d'ambiguïté.

```
<text>جَـــرَعنا فِـــي الـسَّـــيْرِ نَحْـــوَ مِيريبِيـــتْ هـــاوْس</text>
<annotationSet cDate="04/10/2017" status="S_MANUAL" ID="480">
  <layer rank="1" name="Target">
    <label cBy="Aziz" end="4" start="0" name="Target"/>
  </layer>
  <layer rank="1" name="FE">
    <label cBy="Aziz" feID="2542" end="6" start="5" name="Agent"/>
    <label cBy="Aziz" feID="2085" end="42" start="8" name="Activity"/>
  </layer>
  <layer rank="1" name="GF">
    <label end="6" start="5" name="SBJp"/>
    <label end="42" start="8" name="POBJ"/>
  </layer>
  <layer rank="1" name="PT">
    <label end="6" start="5" name="NP"/>
    <label end="42" start="8" name="PP"/>
  </layer>
  <layer rank="1" name="AWP">
    <label end="6" start="0" name="V;CAT:VPKd;TEN:PV;VOI:A;TRA:VPKd;GEN:_;I
    <label end="6" start="5" name="SBJp:Drop;pron:ـنا;nA;pgn:1_P"/>
    <label end="12" start="10" name="P;CAT:PREP;TEN:_"/>
    <label end="21" start="14" name="N;DEF:D;CAT:N;NUM:_;GEN:M;PER:_;CAS:G
    <label end="29" start="24" name="P;CAT:ADV1;TEN:_"/>
    <label end="39" start="31" name="N;DEF:_;CAT:NOP;NUM:S;GEN:M;PER:1;CAS
    <label end="45" start="41" name="N;DEF:_;CAT:NOP;NUM:S;GEN:M;PER:1;CAS
  </layer>
```

Figure 6. *Une représentation en XML pour une phrase annotée selon différentes couches*

Par la suite, l'utilisateur choisit la solution adéquate parmi les suggestions, elle sera enregistrée comme un résultat final de l'analyse (figure 6) (Lakhfif, 2016).

Dans le niveau morphologique, en plus de la description des lemmes après segmentation, l'outil offre la possibilité d'annotation et de génération des fichiers annotés en symboles préterminaux (tags) les plus utilisés pour l'arabe tels que Penn TreeBank (Marcus *et al.*, 1994), BAMA (BuckWalter tags) et CATiB, QADT. Le système d'analyse et d'annotation génère automatiquement des couches d'annotations

concernant les classes SUMO pour les mots reconnus durant l'analyse à base d'AWN en plus des mots membres des ensembles « synsets » équivalents et leurs traductions anglaises 'Gloses'. Les données d'analyse et d'annotation sont organisées par unité lexicale, ce qui signifie que toutes les couches d'annotations pour une unité lexicale donnée sont enregistrées dans un seul fichier. Cependant, en plus de l'annotation style BFN qui s'appuie sur une représentation syntagmatique de constituants, l'annotation en *fonctions grammaticales* (GF) dans notre projet, et à l'instar des projets SALSA (Burchardt *et al.*, 2009), FrameNet danois, et ASFALDA, est fondée sur la grammaire de dépendance (Tesnière, 1959 ; Mel'čuk, 1988), une grammaire bien adaptée pour l'arabe, une langue ayant un réseau de relations de dépendance dans chaque phrase ou proposition (Ryding, 2005). L'annotation utilise un ensemble large d'étiquettes syntaxiques (supérieur à cinquante étiquettes) telles que le complément accusatif absolu (المفعول المطلق), les compléments circonstanciels (المفعول فيه), le complément de manière (الحال), etc., permettant ainsi d'améliorer la tâche de classification des arguments syntaxiques (Johansson et Nugues, 2008). Ce genre de relation est nécessaire dans certaines applications telles que la traduction automatique et la compréhension du texte.

4. Étude de cas : la théorie des frames sémantiques et le domaine des événements de mouvement et de déplacement.

Les langues naturelles présentent des divergences dans leurs descriptions des événements de mouvement et de déplacement, notamment dans les moyens linguistiques accordés à l'expression de la trajectoire et de la manière. Talmy (1991, 2000) propose de classer les langues, selon la façon d'exprimer le composant décrivant la trajectoire (*path*) du mouvement, en deux grandes familles : la famille des langues à cadre satellitaire (*satellite-framed langages* ou *s-langues*) (comme l'anglais), qui utilisent des particules (satellites) pour exprimer la trajectoire, comme stratégie de base, tandis que la manière est généralement confondue avec la sémantique du verbe (*run, limp, crawl*, etc.) ; la deuxième famille, concerne les langues à cadre verbal (*verb-framed langages* ou *v-langues*), (comme l'espagnol) qui fusionnent le concept de la trajectoire avec celui du mouvement au sein du verbe (*entrer, arriver, sortir*). Dans cette dernière famille, la manière du mouvement est généralement exprimée à travers des compléments de phrase (*il entre à la maison en courant*). Le français, l'italien, l'hébreu (Talmy, 1991) et l'arabe (Talmy, 1991 ; Bernini, 2010) sont classés parmi les *v-langues*. Par exemple, les traductions en arabe de *go in*, *go out* et *go down* sont دَخَلَ (*daxala)* « entrer », خَرَجَ (*xaraja)* « sortir » et نَزَلَ (*nazala)* « descendre » respectivement. Cependant, l'anglais et l'allemand, parmi d'autres, sont des *s-langues* connues pour leur utilisation intensive des verbes de manière dans leurs descriptions des événements de mouvement. Deux conséquences capitales résultent de cette divergence en typologie. La première

conséquence concerne la divergence dans l'expression des informations de la manière du mouvement, et la deuxième conséquence concerne la divergence dans l'expression des informations de la trajectoire du mouvement.

4.1. *Une analyse contrastive à base de FrameNet*

Depuis des décennies, le domaine des expressions de mouvement et de déplacement est devenu le champ préféré des études contrastives cross-langues. Parmi ces études, les travaux de Slobin (2006) et son équipe essayent de montrer l'importance accordée par la langue à l'expression de la manière et le degré d'élaboration des éléments de la trajectoire dans l'expression des événements de mouvement. Ces études ont porté sur des corpus narratifs multilingues. C'est dans cette perspective que le livre d'images *Frog story* et le chapitre 6 du roman *The Hobbit* (Tolkien, 1937) représentent les supports les plus utilisés dans les études contrastives sur la typologie des langues. De son côté, et afin de montrer l'apport des frames sémantiques dans la compréhension du texte, l'équipe du projet BFN a lancé plusieurs projets d'annotation des corpus constitués à partir de romans célèbres tels que *The Tiger of San Pedro* (Doyle, 1908) et le chapitre 14 de *The Hound of the Baskerville*.(Doyle, 1901). Ce dernier a été choisi comme objet d'étude par Ellsworth *et al.*, (2006), Ohara (2007) et Petruck (2008), dans leurs analyses contrastives sous la lumière de la théorie des frames sémantique. Ellsworth *et al.*, (2006) ont effectué une étude comparative en termes de prédicats déclencheurs de frames dans la version anglaise et de ses traductions (espagnole et japonaise). En plus de la validation des hypothèses établies par Talmy et Slobin sur la typologie des langues, l'étude a révélé l'existence de différentes stratégies de conceptualisation d'un même événement. Cette divergence apparaît également au sein d'une même famille de langues. Par exemple, en anglais, une scène décrivant une action avec des participants peut se voir traduite, en japonais, en une expression décrivant la scène entière ou l'état résultant de l'action. De son côté, Ohara (2007) a mis l'accent sur l'importance de l'interaction entre la sémantique du prédicat déclencheur de la frame et ses constructions grammaticales dans la conceptualisation des événements. Ces études contrastives à base de FrameNet ont révélé que la structure conceptuelle des frames et les relations frame à frame permettent de comparer les langues avec plus de détails que dans les études contrastives fondées uniquement sur la typologie des langues.

Figure 7. *Annotation multilingue de texte en rôles sémantiques*

Dans cette direction de recherche, et afin de savoir si l'anglais et l'arabe conceptualisent les événements de mouvement avec les mêmes structures conceptuelles (frames), nous avons procédé à l'annotation multilingue (EN-FR-AR) en rôles sémantiques du chapitre 6 du célèbre roman *The Hobbit*, en utilisant notre outil d'annotation (Lakhfif *et al.*, 2013) (figure 7). Le tableau 7, qui résume les divergences en termes de conceptualisation (frames), montre les types et le nombre des frames évoqués dans les expressions de mouvement des créatures ressemblant aux humains (Bilbo, nains, etc.) dans la version originale et sa traduction en arabe. Notre étude révèle que parmi les soixante-douze frames utilisées dans la version originale, la version traduite en arabe a employé les mêmes frames dans soixante et une expressions. En effet, parmi les soixante expressions évoquant une '*Self_motion*' dans la version anglaise, uniquement quatre expressions dans la version arabe ont montré différentes frames, avec trois expressions qui étaient caractérisées par les frames '*Motion_directional*' *et* '*Arriving*' évoquées par les prédicats نَزَلَ (descendre), عادَ (retourner), إقترب (s'approcher de), respectivement. Cette divergence est due à la différence en typologie des deux langues, car les verbes de manière '*climb*' et '*creep*' n'ont pas d'équivalent dans les *v-langues* et ils se traduisent généralement par des verbes directionnels. Cette remarque confirme l'hypothèse de Slobin (2004) qui affirme que les *v-langues* n'ont pas d'équivalent pour l'expression '*climb down*'.

Frames évoquées (**EN**)	Frames évoquées (**AR**)	# d'expressions	
Self_motion	*Self_motion*	**56**	-
Self_motion	*Motion_directional*	**01**	(**نَزَلَ**) - descendre
Self_motion	*Arriving*	**02**	(**عادَ** – **إقترب**) – retourner, s'approcher de
Self_motion	*Manipulation*	**01**	(**تَعلَّقَ**) - s'accrocher
Motion	*Motion_directional*	**02**	(**تَدَحْرَجَ**) - dégringoler
Motion	*Self_motion*	**02**	(**إِنْزَلَقَ**) – se glisser
Motion_directional	*Motion_directional*	**04**	(**سَقَط**, **وَقَع**) - tomber
Motion_directional	*Cause_motion*	**01**	(**أَوْقَعَ**) – faire tomber
Cause_to_move_in_place	*Manipulation*	**01**	(**تَعلَّقَ**) – s'accrocher
Fleeing	*Fleeing*	**01**	(**هَرَب**) – s'échapper
Dispersal	*Self_motion*	**01**	(**تَفَرَّقَ**) – se disperser

Tableau 7 *Type de frames évoqués dans les expressions de mouvement des créatures ressemblant aux humains dans* The Hobbit *chap. 6* (*EN-AR*)

Un autre point important, que nous avons constaté, est que, lorsque la version originale utilise deux verbes de manière (*run*, *creep*) qui évoquent la frame '*Self_motion'* pour caractériser un mouvement à manière en direction d'un but, exprimée *via* '*back – en arrière*' et '*still nearer – plus près*' respectivement, la traduction arabe se focalise sur le but à atteindre, en utilisant la frame '*Arriving'* (qui caractérise le mouvement du côté but (*Goal-profiling*)). Cette frame '*Arriving'* est évoquée par les deux prédicats عاد (*retourner*) et إقْتَرَبَ (*s'approcher*), respectivement. Cette façon de conceptualisation est prévisible dans la version arabe où la manière du mouvement est généralement omise, comme l'atteste la première expression (*run*). Cependant dans la deuxième expression (*creep*), la traduction était fidèle même dans la description de la manière, et ce, grâce à l'adverbe (participe actif) arabe زاحفا (*creeping*) « en rampant ». Il est à noter que cette stratégie représente une exception dans les *v-Langues* (la version française a omis la manière : '*Il se glissa encore plus près'*). En dépit des différences en termes de caractérisation en frames (*Self_motion vs Arriving*, qui partagent certaines caractéristiques de la frame générale '*Motion'*), les deux représentations ont capturé l'essentiel du sens (manière et location) véhiculé par chaque expression avec des points de vue différents. En résumé, 85 % des expressions ont été

caractérisées par les mêmes frames, présentant ainsi un parallélisme conceptuel remarquable vu la divergence dans la typologie.

	Goal	Source	Path	Area	Place	Direction
Anglais	21	14	37	10	5	1
Arabe	35	09	15	6	5	3

Tableau 8. *Distribution des éléments des frames (FE) (EN/AR)*

Cependant, le tableau 8 montre les conséquences de la divergence de typologie (v-langue *vs* s-langue) sur la réalisation sémantique des composants du mouvement dans les deux langues. Cette divergence est due aux différentes stratégies de lexicalisation des éléments de la trajectoire. L'anglais utilise généralement des verbes de manière associés à des particules (satellites : *in*, *out*, *up*, *down*, etc.) pour l'expression des éléments du trajet, alors que l'arabe utilise des verbes qui, en plus de la notion de mouvement, expriment aussi certains éléments de la trajectoire.

FR	Vers le milieu du pont, Cosette, ayant les pieds engourdis, désira [**marcher** **Self_motion**]. [Il *Agent*] [la *Theme*] [**posa** **Placing**] [à terre *Goal*] et la reprit par la main.
EN	Towards the middle of the Bridge, Cosette, whose feet were benumbed, wanted to [**walk** **Self_motion**]. [He *Agent*] [**set** **Placing**] [her *Theme*] [on the ground *Goal*] and took her hand again.
AR	و حوالي منتصف الجسر رَغِبت كوزيت، و قد خدرت رجلاها، في أن [تسير **Self_motion**] [فَأَنْزَلَهَا [**Cause_motion**]إلى الأَرْضِ *Goal*] و أَمْسَكَ بِيَدِهَا. Trad. *: Et environ la moitié du pont, Cosette, ces deux pieds s'engourdirent, désira de marcher et donc il la fit descendre à la terre.*
Glose	w HawAly munotaSafi Aljisori ragibato kuwziyto wa qado xadirato rijolAhA fiy >no tasiyra fa>anozalahaA <lY Al>aroDi w >amosaka biyadihaA

Tableau 9. *Conceptualisation en Frames (*Les Misérables*) (EN/AR/FR)*

Une autre divergence dans les points de vue en termes de conceptualisation est présentée dans le tableau 9 représentant un exemple à partir du célèbre roman français *Les Misérables* (Hugo, 1845-1862**)**. Le texte original (français) et sa traduction en anglais utilisent la frame '*Placing'* (un '*Agent*' place un '*Theme*' dans une location '*Goal*') pour caractériser la scène où *Jean Valjean* essaye de poser par terre *Cosette* qui est dans ses bras. Cependant, dans la version arabe, le traducteur a préféré utiliser le verbe أنزل (faire descendre) évoquant, ainsi la frame '*Cause_motion'* (un *'Agent'* cause à

un *'Theme'* de subir un mouvement direct). Selon la figure 2 (section 2), la conceptualisation arabe de l'événement n'est qu'un point de vue général de la scène caractérisée par *'Placing'* qui est une sous-frame de *'Cause_motion'*. Un point fort de FrameNet et que, contrairement aux autres bases lexicales, elle offre des mécanismes (relations frame à frame) permettant de relier toutes les UL qui peuvent concevoir la même scène. En dépit de ces divergences de conceptualisation en termes de frames, la théorie de Fillmore dispose de mécanismes d'interprétation et de moyens de prise en charge des effets qui résultent des différences dans les stratégies de lexicalisation des événements de mouvement et de déplacement.

5. Conclusion

Dans cet article nous avons présenté les premiers résultats de notre tentative d'application de la théorie des frames sémantiques pour l'analyse de l'arabe, langue qui diffère de l'anglais typologiquement. Nous avons montré qu'une intégration de plusieurs ressources linguistique peut améliorer la qualité des systèmes TAL (Baker et Felbaum, 2009) (section 3.3). Nous avons aussi montré qu'il était possible d'appliquer les procédures d'analyse et d'annotation du projet BFN sur l'arabe, avec quelques adaptations nécessaires afin de respecter les caractéristiques linguistiques propres à l'arabe. Nous avons effectué une étude contrastive à base des frames sémantiques sur la conceptualisation des événements de mouvement et de déplacement. Nous avons aussi montré qu'une approche par l'analyse sémantique peut apporter des réponses aux questions relatives à la divergence des langues dans l'expression des événements de mouvement.

Nous croyons que notre analyse contrastive offre un apport aux études sur la typologie des langues initiées par Talmy et Slobin et représente les premiers résultats pour l'arabe fondés sur un objet d'étude de référence (*The Hobbit*) dans les analyses cross-langues. Sur cette base, nous avons confirmé les conclusions des chercheurs (Fillmore, Ohara, Ellsworth *et al.*, Petruck) sur le rôle important de la structure des frames et des relations frame à frame dans l'interprétation des divergences en conceptualisation entre les différentes langues. Cependant, afin d'arriver à une généralisation de ces conclusions, une étude approfondie doit être établie sur un large corpus représentatif. Notre travail peut être étendu sur d'autres phénomènes langagiers afin de répondre aux questions liées à l'applicabilité de la théorie des frames sémantiques sur l'arabe.

6. Bibliographie

Al-Sulaiti L., Atwell E., « The design of a corpus of contemporary Arabic », *International Journal of Corpus Linguistics*, vol. 11, p. 135-171, 2006.

Alshehri A. M., « The frame semantics of 'selfmotion'frame in Arabic and English », Thèse de doctorat, San Francisco : San Francisco State University, 2014.

Baker, C.F., Fillmore, C.J., Lowe, J.B., « The Berkeley FrameNet Project », *In COLING-ACL '98 : Proceedings of the Conference*, p. 86-90, 1998.

Baker C., « La sémantique des cadres et le projet FRAMENET : une approche différente de la notion de « valence »«, *Langages,* vol. 4, n° 176, p. 32-49,2009.

Baker C., Fellbaum C., « WordNet and FrameNet as complementary resources for annotation », *In : Association for Computational Linguistics* (ed.), In *Proceedings of the Third Linguistic Annotation Workshop*, Singapore, p.125-129, 2009.

Baker C., Fellbaum C., Passonneau R.J., « Semantic Annotation of MASC », *Handbook of Linguistic Annotation*, Springer, Dordrecht, p. 699-717, 2017.

Bernini, G., « Word classes and the coding of spatial relations in motion events : A contrastive typological approach », In *Space in Language*, p. 29-52, Edizioni ETS, 2010.

Bick E., « A FrameNet for Danish », *In : Proceedings of Nodalida*, Riga, Latvia, NEALT Proceedings Series, *Riga, Latvia,* vol. 11, p.34-41, 2011.

Black W., Elkateb S., Rodriguez H., Alkhalifa M., Vossen P., Pease A., Fellbaum C.. « Introducing the Arabic wordnet project », *In Proceedings of the third international WordNet conference*, p. 295-300, Seogwipo, Korea, 2006.

Boas H. C., « Bilingual FrameNet Dictionaries for Machine Translation », *In LREC,* 2002, p. 1364-1371, Las Palmas, Iles Canaries, 2002.

Boas H. C., « Semantic frames as interlingual representations for multilingual databases », *International Journal of Lexicography*, vol. 18, n° 4, p. 445-478, 2005.

Boas H. C., « Recent trends in multilingual computational lexicography », *Multilingual* FrameNet*s in Computational Lexicography : Methods and Applications. Berlin and New York : Mouton de Gruyter*, p. 1-36, 2009.

Buckwalter T., « Buckwalter Arabic morphological analyzer version 1.0 », linguistic data consortium, *University of Pennsylvania, LDC Catalog n° : LDC2002L49,* 2002.

Burchardt A., Erk K., Frank A., Kowalski A. Pado S., « SALTO : A versatile multi-level annotation tool », *In Proceedings of LREC-*, Genoa, Italy, 2006.

Burchardt A., Reiter N., Thater S., Frank A., « A semantic approach to textual entailment : system evaluation and task analysis ». *In : Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing,* Prague, République Tchèque, 2007.

Burchardt A., Erk K., Frank A., Kowalski A., Padó S., Pinkal M., « Using FrameNet for the semantic analysis of German : annotation, representation, and automation », *Multilingual* FrameNet*s in Computational Lexicography : methods and applications*, p. 209-244, 2009.

Burchardt A., Pennacchiotti M., « FATE : Annotating a Textual Entailment Corpus with FrameNet », *In Handbook of Linguistic Annotation*, p. 1101-1118, Springer, 2017.

Candito M., Amsili P., Barque L., Benamara F., De Chalendar G., Djemaa M., Haas P., Huy-ghe R., Mathieu Y. Y., Muller P., Sagot B., Vieu L., « Developing a French FrameNet : Methodology and First results », *Proceedings of the International Conference on Language Resources and Evaluation Conference* (*LREC*), Reykjavik, Iceland, 2014.

Coyne O. Rambow Hirschberg J., Sproat R., « Frame Semantics in Text-to-Scene Generation », *In Proceedings of the KES'10 workshop on 3D Visualisation of Natural Language*, Cardiff, Wales, 2010.

Diab M., Mansouri A., Palmer M., Babko-Malaya O., Zaghouani W., Bies A., Maamouri M., « A Pilot Arabic Propbank », In *Proceedings of the 7th International Conference on Language Resources and Evaluation*, 2008.

Djemaa M., Candito M., Muller P. Vieu L., May., « Corpus annotation within the french framenet : a domain-by-domain methodology », *In Tenth International Conference on Language Resources and Evaluation* (*LREC*), Portorož, Slovenia, 2016.

Dolbey, A., Ellsworth, M. and Scheffczyk, J., « BioFrameNet : A Domain-Specific FrameNet Extension with Links to Biomedical Ontologies », In *KR-MED*, Vol. 222, 2006.

Doyle, A. C. *The hound of the Baskervilles.* London : Strand Magazine, 1901. [Trad.arabe. M.H. Mahmoud. (1997). "شبح باسكرفيل".] [Trad. Franç. A. de Jassard. *Le Chien des Baskerville*, 1905. La Bibliothèque électronique du Québec]

Dukes K., Buckwalter T., « A Dependency Treebank of the Quran using Traditional Arabic Grammar », *In Proceedings of the 7th international conference on Informatics and Systems*, Cairo, Égypte, 2010.

Ellsworth M., Ohara K., Subirats K., Schmidt T., « Frame-semantic analysis of motion scenarios in English, German, Spanish, and Japanese », Paper presented at *The Fourth International Conference on Construction Grammar*, Tokyo, 2006.

Erk K., Pado S., « Shalmaneser a toolchain for shallow semantic parsing », In *Proceedings of LREC,* vol. 6, 2006.

Ferrández Ó., Ellsworth M., Muñoz R., Baker C.F., « Aligning FrameNet and WordNet based on semantic neighborhoods », In *Proceedings of the 7th international language resources and evaluation conference* (*LREC*), p. 310-314, 2010.

Fellbaum C., « WordNet : An Electronic Database ». MIT Press, Cambridge, MA. 1998.

Fellbaum C., Osherson A., Clark P.E., « Putting semantics into WordNet's "morphosemantic" links », *In Proceedings of the 3th Language and Technology Conference, Poland*, 2007.

Fillmore C.J., « The case for case », *In : Bach, E., Harms, R.* (*eds.*)*, Universals in Linguistic Theory*, Holt, Rinehart & Winston, New York, 1968.

Fillmore C.J., « Frame Semantics », *Linguistics in the Morning Calm*, p. 111-38, Seoul, 1982.

Fillmore C.J., Johnson C.R., Petruck M.R.L., « Background to Framenet », *International Journal of Lexicography*, vol. 16, n° 3, p. 235-250, 2003.

Fillmore C. J., Baker C., « A frames approach to semantic analysis », *In Heine, Bernd and Heiko Narrog* (*Eds.*), The Oxford Handbook of Linguistic Analysis, p. 313-341, Oxford University Press, 2010.

Ghazzawi N., « Du terme prédicatif au cadre sémantique : méthodologie de compilation d'une ressource terminologique pour les termes arabes de l'informatique », *Thèse de doctorat*, mai 2016.

Gildea D., Jurafsky D., « Automatic labeling of semantic roles », *Computational linguistics*, vol. 28, n° 3, p. 245-288, 2002.

Habash N., Dorr B., « Handling translation divergences : Combining statistical and symbolic techniques in generation-heavy machine translation », *Machine Translation : From Research to Real Users*, p. 84-93, 2002.

Habash N., Roth, R. M., « CATiB : The columbia arabic treebank », In *Proceedings of the ACL-IJCNLP, Conference Short Papers*, p. 221-224, 2009.

Habash N. Y., « Introduction to Arabic natural language processing », *Synthesis Lectures on Human Language Technologies*, vol. 3 n° 1, p. 1-187, 2010.

Hajic J., Smrz O., Zemánek P., Šnaidauf J., Beška E., « Prague Arabic dependency treebank : Development in data and tools ». *In Proceeding of the NEMLAR Intern. Conf. on Arabic Language Resources and Tools,*Cairo, Egypt, p. 110-117, 2004.

Hartmann S., Kuznetsov I., Martin T., Gurevych I., « Out-of-domain FrameNet Semantic Role Labeling », In *Proceedings of the 15th Conference of the European Chapter of the ACL : vol. 1, Long Papers*, p. 471-482, 2017.

Hayoun A., Elhadad M., « The Hebrew FrameNet Project », *In Proceedings of the International Conference on Language Resources and Evaluation Conference* (*LREC*), Portorož Slovenia, 2016.

Hugo, V., *Les Misérables*.1845-1862 [1- Trad. arabe. M. Albalabki. (1979), « البؤساء ». Dar El Ilm Lilmalayin, Liban] [2- Trad. anglaise. I. F. Hapgood. *Les Misérables*,1887, Thomas Y. Cowell :New York, USA]

Johansson R., Nugues P., « The effect of syntactic representation on semantic role labeling », *In Proceedings of International Conference on Computational Linguistics* (COLING), Manchester, UK, 2008.

Jeonguk k., Hahm y., Choi K.S., « Korean FrameNet Expansion Based on Projection of Japanese FrameNet », *In Proceeding COLING (Demos)*, Osaka, Japan, 2016.

Kipper-Schuler K., « VerbNet : A broad-coverage, comprehensive verb lexicon », PhD thesis, University of Pennsylvania, 2005.

Lakhfif A., Laskri M. T., Atwell E., « Multi-Level Analysis and Annotation of Arabic Corpora for Text-to-Sign Language MT », In Second Workshop on Arabic Corpus Linguistics (WACL-2), Lancaster University, UK, 2013.

Lakhfif A., Laskri M. T., « A frame-based approach for capturing semantics from Arabic text for text-to-sign language », *International Journal of Speech Technology*, doi : 10.1007/s10772-015-9290-8, 2015.

Lakhfif A., « Un signeur virtuel 3D pour la traduction automatique de textes arabes vers la langue des signes algérienne », thèse de doctorat en sciences, avril 2016.

L'Homme M.C., *Le DiCoInfo. Méthodologie pour une nouvelle génération de dictionnaires spécialisés*, Traduire, vol. 217, p. 78-103, 2008.

Johnson M., Lenci A., « Verbs of visual perception in Italian FrameNet », Constructions and Frames, vol. 3, n° 1, p. 9-45, 2011.

Lindén K., Haltia H., Luukkonen J., Laine A. O. Roivainen H., Väisänen N., « FinnFN 1.0 :The Finnish frame semantic database », *Nordic Journal of Linguistics*, vol. 40, n° 3, p. 287-311, 2017.

Maamouri M., Bies A., Buckwalter T., Mekki W., « The penn arabic treebank : Building a large-scale annotated arabic corpus », In *NEMLAR conference on Arabic language resources and tools*, p. 102-109, 2004.

Marcus M. P., Santorini B., Marcinkiewicz M. A., « Building a large annotated corpus of English : The Penn TreeBank », Computational Linguistics, vol. 19, n° 2, p. 313, 1994.

Minsky M., « A framework for representing knowledge », *In Proceeding Winston* (*Ed.*), *The Psychology of Computer Vision*, McGraw-Hill, 1975.

Mel'čuk I., « *Dependency Syntax : Theory and Practice* », State University of NY Press, Albany, 1988.

Moschitti, A., « Kernel methods, syntax and semantics for relational text categorization », In *Proceedings of the 17th ACM conference on Information and knowledge management,* p. 253-262, ACM, 2008.

Narayanan S., Harabagiu S., « Question answering based on semantic structures », In *Proceedings of the 20th international conference on Computational Linguistics* (*ACL*), p. 693-, 2004.

Nuzzolese A. G., Gangemi A., Presutti V., « Gathering lexical linked data and knowledge patterns from FrameNet », In *Proceedings of the sixth international conference on Knowledge capture*, p. 41-48, ACM, 2011.

Ohara K. H., Fujii S., Ohori T., Suzuki R., Saito H., Ishizaki S., « The japanese framenet project : An introduction », *In Proceedings of LREC-04, Satellite Workshop* « *Building Lexical Resources from Semantically Annotated Corpora* », p. 9-11, 2004.

Ohara, K.H., « Frame Semantics in Action : A Frame-based Contrastive Text Analysis Using FrameNet », In *the 10th International Cognitive Linguistics Conference*, Poland, 2007.

Pado S., Pitel G., « Annotation précise du français en sémantique de rôles par projection cross-linguistique », *Proceedings of TALN-07*, Toulouse, France, 2007.

Petruck M.R.L., « Frame semantics *», In Verschueren, J., Östman, J-O.* 2007*, Blommaert, J. & Bulcaen, C., eds, Handbook of pragmatics*, Amsterdam : John Benjamins, p. 1-13, 1996.

Reed S. K., Pease A., « A framework for constructing cognition ontologies using WordNet, FrameNet, and SUMO », *Cognitive Systems Research*, vol. 33, p.122-144, 2015.

Ruppenhofer J., Ellsworth M., Petruck M. R., Johnson C. R., Scheffczyk J., « FrameNet *II : Extended theory and practice* », Institut für Deutsche Sprache, Bibliothek, 2016a.

Ruppenhofer J., Michaelis L. A., « Frames, polarity and causation », Corpora, vol. 11, n° 2, p. 259-290, 2016b.

Ryding KC., *A Reference Grammar of Modern Standard Arabic*, Cambridge University Press, 2005.

Salomão, M.M.M., « FrameNet Brasil : um trabalho em progresso », Calidoscópio, vol. 7, n° 2, p. 171-182, 2009.

Schank R. C., Abelson R. P., « Scripts, Plans, Goals, and Understanding », Hillsdale, N. J. : Erlbaum Associates, 1977.

Sharaf A., Atwell E., « Knowledge Representation of the Quran Through Frame Semantics : A Corpus-Based Approach », *In Proceedings of the Fifth Corpus Linguistics Conference*. University of Liverpool, UK, 2009.

Shen D., Lapata M., « Using Semantic Roles to Improve Question Answering », *In EMNLP-CoNLL,* p. 12-21, 2007.

Shi L., Mihalcea R., « Putting pieces together : Combining FrameNet, VerbNet and WordNet for robust semantic parsing », In *Computational linguistics and intelligent text processing,* p. 100-111, Springer Berlin Heidelberg, 2005.

Schmidt T., « The Kicktionary – a Multilingual Lexical Recource of Football Language », [w :] Boas, HC (red.).*Multilingual FrameNets*, 2007.

Slobin Dan I., « The many ways to search for a frog : Linguistic typology and the expression of motion events », *In Stromqvist & Verhoeven* (*eds.*), p.219-257, 2004a.

Slobin D. I., « Relating events in translation », *In Perspectives on language and language development : Essays in honor of Ruth A. Berman*. Dordrecht : Kluwer, p. 115-130, 2004b.

Søgaard A., Plank B., Alonso H. M., « Using Frame Semantics for Knowledge Extraction from Twitter », *In AAAI*, p. 2447-2452, 2015.

Subirats C., Petruck M.R.L., « Surprise : Spanish FrameNet », *In E. Hajicova, A. Kotesovcova & J. Mirovsky* (*eds.*), *Proceedings of CIL* 17. CD-ROM. Prague : Matfyzpress, 2003.

Talmy, L., « Toward a cognitive semantics », Cambridge, MA : MIT Press, 2000.

Talmy L., « Path to realization : A typology of event conflation », *Proceedings of the Seventeenth Annual Meeting of the Berkeley Linguistics Society*, p. 480-519, 1991.

Tolkien J. R. R., *The Hobbit or there and back again*, London : George Allen & Unwin, 1937. [1-Trad.arabe. H. Fahmy, M. Ghanim. (2008), « الهوبيت (أو ذهابا و عودة) ». Dar Lila-Boeken.] [Trad.française. F. Ledoux. *Bilbo le Hobbit*, 1980, Paris : Hachette.]

Tesnière L., « Éléments de syntaxe structurale », Librairie C. Klincksieck, 1959.

Torrent T., Salomão M. M., Campos F. A., Braga R. M., Matos E. E., Gamonal M., Gonçalves J., Souza B. C., Gomes D., Peron S., « Copa FrameNet Brasil : a frame -based trilingual electronic dictionary for the Football World Cup », In COLING (Demos), 10-14, 2014.

Venturi G., Lenci A., Montemagni S., Vecchi E. M., Sagri M. T., Tiscornia D., Agnoloni T., « Towards a FrameNet resource for the legal domain », *LOAIT*, p. 67-76, 2009.

Vossen P. « Introduction to eurowordnet », In *EuroWordNet : A multilingual database with lexical semantic networks,* Springer Netherlands, p. 1-17, 1998.

Vossen P. (ed) « EuroWordNet », (LE2-4003, LE4-8328), Part A, Final Document Deliverable D032D033/2D014, 1999.

Morphology-based Entity and Relational Entity Extraction Framework for Arabic

Amin Jaber* — **Fadi A. Zaraket****

*Purdue University, West Lafayette, IN
**American University of Beirut, Beirut 1107 2020, Lebanon

ABSTRACT. *Rule-based techniques to extract relational entities from documents allow users to specify desired entities with natural language questions, finite state automata, regular expressions and structured query language. They require linguistic and programming expertise and lack support for Arabic morphological analysis. We present a morphology-based entity and relational entity extraction framework for Arabic (*MERF*). *MERF *requires basic knowledge of linguistic features and regular expressions, and provides the ability to interactively specify Arabic morphological and synonymity features, tag types associated with regular expressions, and relations and code actions defined over matches of subexpressions.* MERF *constructs entities and relational entities from matches of the specifications. We evaluated* MERF *with several case studies. The results show that* MERF *requires shorter development time and effort compared to existing application specific techniques and produces reasonably accurate results within a reasonable overhead in run time.*

RÉSUMÉ. *Les techniques à base de règles pour extraire des entités permettent de spécifier les entités souhaitées en utilisant des questions de langage naturel, des automates à états finis, des expressions régulières et des instructions d'extraction de données. Ils nécessitent des expertises en linguistique et en programmation, et ne soutiennent pas l'analyse morphologique de l'arabe. On présente pour l'arabe un cadre d'extraction d'entité renforcé par l'analyse morphologique (*MERF*). Il exige des connaissances de base des caractéristiques linguistiques et des expressions régulières, et fournit la possibilité de spécifier de façon interactive des fonctionnalités de morphologie et synonymie arabes, des types de tag associés avec des expressions régulières, et des relations et actions de code définies sur les correspondances de sous-expressions.* MERF *construit des entités relationnelles à partir des correspondances des spécifications. On évalue* MERF *avec des études de cas. Les résultats montrent que* MERF *nécessite un effort de développement plus court par rapport aux techniques existantes et produit des résultats raisonnablement précis avec une surcharge raisonnable en temps d'exécution.*

KEYWORDS: *Arabic, information extraction, natural language processing, tagging.*

MOTS-CLÉS : *Arabe, extraction d'information, traitement du langage naturel, marquage.*

TAL. Volume 58 – n° 3/2017, pages 97 à 121

1. Introduction

Computational Linguistics (CL) is concerned with building accurate linguistic computational models. *Natural Language Processing* (NLP) is concerned with automating the understanding of natural language. CL and NLP tasks range from simple ones such as spell checking and typing error correction to more complex tasks including *named entity recognition* (NER), *cross-document analysis*, *machine translation*, and *relational entity extraction* (Linckels and Meinel, 2011; Ferilli, 2011). Entities are elements of text that are of interest to an NLP task. Relational entities are elements that connect entities. *Annotations* relate chunks of text to *labels* denoting semantic values such as entities or relational entities. We refer to annotations and labels as *tags* and *tag types*, respectively, in the sequel.

Supervised and unsupervised empirical learning techniques tackle NLP and CL tasks. They employ machine learning without the need to manually encode the requisite knowledge (Soudi *et al.*, 2007). Supervised learning techniques require training corpora annotated with *correct* tags to learn a computational model. Supervised and unsupervised techniques require annotated reference corpora to evaluate the accuracy of the technique using metrics such as precision and recall (Marcus *et al.*, 1993; Maamouri *et al.*, 2004; Xue *et al.*, 2005).

Researchers build training and reference corpora either manually, incrementally using learning techniques, or using knowledge-based annotation techniques that recognize and extract entities and relational entities from text. Knowledge-based techniques use linguistic and rhetorical domain specific knowledge encoded into sets of rules to extract entities and relational entities (Soudi *et al.*, 2007). While existing annotation, entity, and relational entity extraction tools exist (Chiticariu *et al.*, 2010; Atzmueller *et al.*, 2008; Urbain, 2012; Settles, 2011; Müller and Strube, 2006; Stenetorp *et al.*, 2012), most of them lack Arabic language support, and almost all of them lack Arabic morphological analysis support (Habash and Sadat, 2006). Fassieh (Attia *et al.*, 2009) is a *commercial* Arabic annotation tool with morphological analysis support and text factorization. However, this tool lacks support for entity and relational entity extraction.

Figure 1 illustrates the target of MERF using the directions to Dubai Mall example [1]. The figure also presents a transliteration and an English translation of the Arabic text. The framed words in the text are entities referring to names of people (n_1, n_2, n_3), names of places ($p_1, \ldots, p_7$), relative positions ($r_1, \ldots, r_4$), and numerical terms (u_1, u_2). We would like to extract those entities, and then extract the relational entities forming the graph in Figure 1 where vertices express entities, and edges represent the relational entities.

In this paper, we present MERF, a morphology-based entity and relational entity extraction framework for Arabic text. MERF provides a user-friendly interface where the user defines tag types and associates them with regular expressions over Boolean

1. Text taken from the Dubai Mall website `http://www.thedubaimall.com/ar/`.

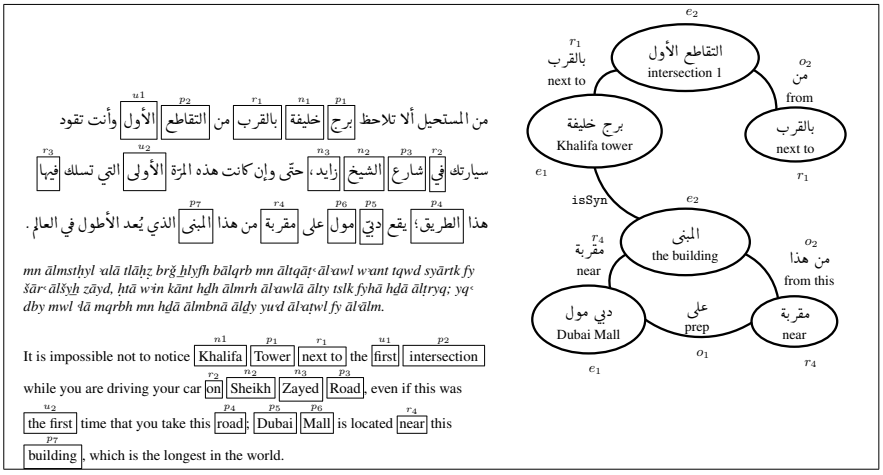


Figure 1. *Direction example with Arabic text, annotated with entities, transliteration, translation, and extracted relational entities in a graph.*

formulae. A Boolean formula is defined by a term, negation of a term, or disjunction of terms. Terms are matches to Arabic morphological features including prefix, stem, suffix, part of speech (POS) tags, gloss tags, extended synonym tags, and semantic categories. For example, entity p_1 in Figure 1 has a "place" semantic category. MERF regular expressions support operators such as concatenation, zero or one, zero or more, one or more, up to M repetitions where M is a non-zero positive integer, and logical conjunction and disjunction. For example, the sequence between p_1 and p_2 matches a regular expression re that requires two semantic place categories with a place-preposition POS tag (r_1) in between.

An editor allows the user to associate an action with each subexpression. The user specifies the action with C++ code and uses an API to access information related to the matches such as text, position, length, morphological features, and numerical value. Each regular expression is associated with a named identifier to form a *local grammar* like structure (Traboulsi, 2009). A relation definition GUI allows the user to provide relational tuples where each tuple has a source, a destination and an edge label. The user uses the regular expression identifiers to define the relational tuple elements. For example, the relation between e_1, e_2 and r shown in Figure 1 is a match of a relational tuple over the components of re. We refer to regular expressions and Boolean formulae as expressions and formulae, respectively. We also refer to expressions as rules when used in a grammar context; e.g. when used with an identifier.

MERF takes an Arabic text and the local grammar defined by the Boolean formulae and the regular expressions. MERF computes the morphological solutions of the input text then computes matches to the Boolean formulae therein. MERF then generates a *non-deterministic finite state automata* (NDFSA) for each expression and simulates it with the sequence of Boolean formulae matches to compute the regular

expression matches. MERF generates executable code for the actions associated with the regular expressions, compiles, links, and executes the generated code as shared object libraries. Finally, MERF constructs the semantic relations and cross-reference between entities. MERF also provides visualization tools to present the matches, and estimate their accuracy with respect to reference tags.

This work significantly extends Jaber and Zaraket (2013) that allows for manual, and morphology annotation. MERF enables a user to incrementally create complex annotations for Arabic based on automatic extraction of morphological tags through a user-friendly interactive interface. MERF has the following advantages.

– MERF provides a novel and intuitive visual interface to build formulae over morphological features, build regular expressions over the resulting formulae, and thereafter compute automatic tags.

– To our knowledge, this morphology-based framework is the first for Arabic entity and relational entity extraction.

– MERF provides the user with the ability to rapidly create annotated Arabic text corpora with sophisticated morphology-based tags.

In MERF, we make the following contributions.

– MERF enables the user to define relations in a simple manner and automatically detects relational entities matching the user defined relations.

– MERF enables the user to associate subexpressions with code actions, and executes the code action when a corresponding match is found. It also provides an API to enable access to match features such as text, position, length, numerical value, and morphological features.

– MERF enables the user to tag words based on a novel light Arabic WordNet relation that leverages the synonym Syn^k feature.

– MERF is open source and available online for the research community under `https://github.com/codelogicanalysis/atmine`.

The rest of the paper is structured as follows. Section 2 introduces Arabic morphological analysis and its important role in Arabic NLP. Section 3 explains the methodology of MERF. Section 4 presents MERF components. Section 5 presents MERF GUI. Section 6 presents and discusses related work. Section 7 presents the evaluation results. Finally, we conclude and discuss future work in Section 8.

2. Background: Morphological Analyzer

Morphological analysis is key to Arabic NLP due to the exceptional degree of ambiguity in writing, the rich morphology, and the complex word derivation system (Al-Sughaiyer and Al-Kharashi, 2003; Shahrour *et al.*, 2016; Pasha *et al.*, 2014). Short

vowels, also known as diacritics, are typically omitted in Arabic text and inferred by readers (Habash and Sadat, 2006). For example, the word بنbn can be interpreted as بُنbon (“coffee”) with a *damma* diacritic on the letter بـb or بِنbin (“son of”) with a *kasra* diacritic on the letter بـb .

Morphological analysis is required even for tokenization of Arabic text. The position of an Arabic letter in a word (beginning, middle, end, and standalone) changes its visual form. Some letters have non-connecting end forms which allows visual word separation without the need of a white space separator. For example, the word ياسمينyāsmyn can be interpreted as the “Jasmine” flower, as well as يا (the calling word) followed by the word سمين (obese). Consider the sentence ذهب المدرسة الوَلدالَى ḏhb alwald-ilā ’lmdrsh (“the kid went to school”). The letters د and ى have non-connecting end of word forms and the words الولد, الى, and المدرسة are visually separable, yet there is no space character in between. Newspaper articles with text justification requirements, SMS messages, and automatically digitized documents are examples where such problems occur.

MERF is integrated with *Sarf*, an in-house open source Arabic morphological analyzer based on finite state transducers (Zaraket and Makhlouta, 2012b). Given an Arabic word, Sarf returns a set of morphological solutions. A word might have more than one solution due to multiple possible segmentations and multiple tags associated with each word. A morphological solution is the internal structure of the word composed of several morphemes including *affixes* (*prefixes* and *suffixes*), and a *stem*, where each morpheme is associated with tags such as POS, gloss, and category tags (Al-Sughaiyer and Al-Kharashi, 2003; Habash, 2010).

Prefixes attach before the stem and a word can have multiple prefixes. Suffixes attach after the stem and a word can have multiple suffixes. Infixes are inserted inside the stem to form a new stem. In this work we consider a set of stems that includes infix morphological changes. The part-of-speech tag, referred to as POS, assigns a morpho-syntactic tag for a morpheme. The gloss is a brief semantic notation of morpheme in English. A morpheme might have multiple glosses as it could stand for multiple meanings. The category is a custom tag that we assign to multiple morphemes. For example, we define the `Name of Person` category to include proper names.

	Prefixes			Stem	Suffix
Data	فَ*fa*	سَ*sa*	يَ*ya*	أْكُلُ*akul*	ها*hā*
POS	CONJ+	FUT+	IV3MS+	VERB_IMPERFECT	IVSUFF_DO:3FS
Gloss	and/so	will	he/it	eat/consume	it/them/her
index		10		13	16
length		3		3	2

Table 1. *Sample solution vector for* فَسَيَأْكُلُها*fasaya'akulhā* .

We denote by $\mathcal{S}$, $\mathcal{P}$, $\mathcal{X}$, POS, $GLOSS$, and CAT, the set of all stems, prefixes, suffixes, POS, gloss, and user defined category tags, respectively. Let $T = \langle t_1, t_2, \ldots, t_M \rangle$ be a set of Arabic words denoting the text documents. MERF uses Sarf to compute a set of morphological solutions $M(t) = \{m_1, m_2, \ldots, m_N\}$ for each word $t \in T$. Each morphological solution $m \in M(t)$ is a tuple of the form $\langle p, s, x, P, G, C \rangle \in \mathcal{P} \times \mathcal{S} \times \mathcal{X} \times POS \times GLOSS \times CAT$ where $p = p_1 \ldots p_{|p|}$, $x = x_1 \ldots x_{|x|}$, $P = P_{p_1} \ldots P_{p_{|p|}} P_s P_{x_1} \ldots P_{x_{|x|}}$, $G = G_{p_1} \ldots G_{p_{|p|}} G_s G_{x_1} \ldots G_{x_{|x|}}$, and $C = C_{p_1} \ldots C_{p_{|p|}} C_s C_{x_1} \ldots C_{x_{|x|}}$. P_{p_i}, G_{p_i}, and $C_{p_i}, 1 \leqslant i \leqslant |p|$ are the POS, gloss and category tags of prefix p_i. P_{x_j}, G_{x_j}, and $C_{x_j}, 1 \leqslant j \leqslant |x|$ are the POS, gloss and category tags of suffix x_i. P_s, G_s, and C_s are the POS, gloss and category tags of stem s. Intuitively, p, x, P, G and C are concatenations of prefix, suffix, POS, gloss and category values, respectively.

Table 1 shows the morphological analysis of the word فَسَيَأْكُلُها. The word is composed of the prefix morphemes فَ*fa* , سَ*sa* , and يَ*ya* , followed by the stem أْكُلُ*akul* , and then followed by the suffix morpheme هـا*hā* . Each morpheme is associated with a number of morphological features. The `CONJ`, `FUT`, `IV3MS` `VERB_IMPERFECT`, and `IVSUFF_DO:3FS` POS tags indicate conjunction, future, third person masculine singular subject pronoun, an imperfect verb, and a third person feminine singular object pronoun, respectively. The POS and gloss notations follow the Buckwalter notation (Buckwalter, 2002).

3. MERF Methodology

Figure 2 illustrates the four processes involved in MERF methodology. The first process takes Arabic text and provides the user with a morphology-based Boolean (MB) formulae GUI. The user interactively composes MB-formulae using the GUI and the output of the simulator and the Syn^k detector. The simulator and the detector apply the formulae over the morphological solutions of the Arabic text and produce the MB-formulae tags.

The second process takes the MB-formulae tags and the Arabic text and provides the user with a morphology-based grammar rule GUI. The user interactively composes MB-grammar rules using the GUI and the output of the MB-grammar rule simulator. The grammar rule simulator applies the rules over the MB-formulae tags and produces the MB-grammar rule tags.

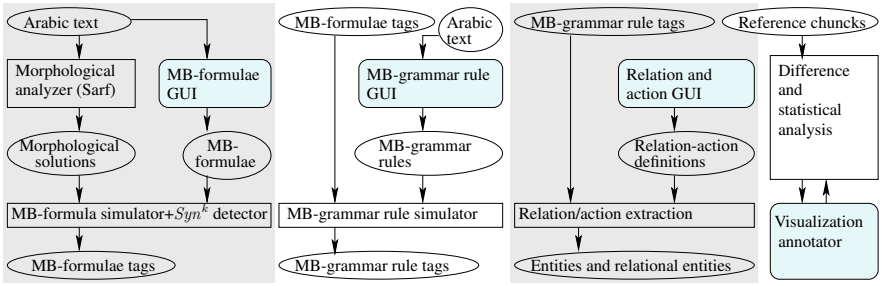


Figure 2. MERF *four process methodology with rounded corner blocks for GUI.*

The third process takes the MB-grammar rule tags and provides the user with a relation and action GUI. The user interactively provides (1) the relation definitions and (2) the actions in terms of identifiers from the MB-grammar rules. The relation extraction produces the target entities and relational entities. The action execution enriches the entities and the relational entities with powerful semantics. For example, users can utilize actions to compute statistical features, store intermediate results, or apply intelligent entity inference techniques as we show later in the numerical extraction example of Subsection 7.4. Finally, in the fourth process the user compares the results with golden reference chunks and visualizes the difference. This allows the user to refine the formulae, rules, relations and actions.

After relation extraction, we are interested to relate entities that express the same concept. MERF provides the extended synonym feature of second order as a default cross-reference relation (Syn^2). In Figure 1, triggering this feature creates the edge labeled with `isSyn` between the nodes `Khalifa Tower` and `The building`.

The user may refine the defined formulae, rules and relations and the corresponding formulae tags, rule tags, entities and relational entities either using the GUI or directly through readable output files. The files are in the javascript object notation (JSON) (Nolan and Lang, 2014) format that is intuitive to read and modify. MERF separates the user defined formulae, rules, actions and relations in a MERF tag type file and the matching tags in a tags files. The separation serves the user to apply the tag types to multiple case studies and to obtain a separate file of resulting tags for each.

4. MERF Components

4.1. *The extended synonymy feature* Syn^k

Up to our knowledge, Syn^k provides the first light Arabic WordNet based on the lexicon of Sarf. The sets E, A, and L denote all English words, Arabic words, and Arabic lexicon words, respectively. Recall that $GLOSS$ and $\mathcal{S}$ denote the set of glosses and stems in the morphological analyzer, respectively. We have $GLOSS \subset E$ and $\mathcal{S} \subset L \subset A$. Function $\alpha : \mathcal{S} \rightarrow 2^{GLOSS}$ maps Arabic stems to subsets of related

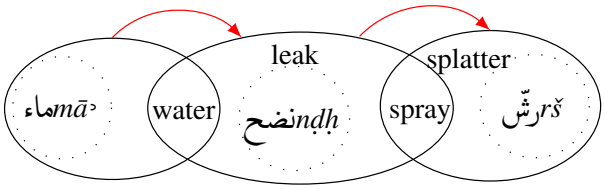


Figure 3. Syn^2(ماء *mā'*).

English glosses, where 2^{GLOSS} denotes the power set of $GLOSS$ which is the set of all subsets of $GLOSS$. Function $\gamma : L \rightarrow 2^{\mathcal{S}}$ maps Arabic lexicon words to subsets of relevant Arabic stems.

Given a word $w \in L$, $Sy(w) = \{u \mid u \in \mathcal{S} \wedge \exists s \in \gamma(w) \wedge \alpha(u) \cap \alpha(s) \neq \varnothing\}$ is the set of Arabic stems directly related to w through the gloss map. Let $Sy^i(w)$ denote stems related to w using the gloss map of order i recursively such that $Sy^1(w) = Sy(w)$ and $Sy^{i+1}_k(w) = \{u \mid u \in S \wedge \exists s \in Sy^i(w) \wedge \alpha(u) \cap \alpha(s) \neq \varnothing\}$. Formally, $Syn^k(w) = \bigcup_{i=1} Sy^i(w)$ for $i \in [1 \ldots k]$. The example in Figure 3 illustrates the computation. Let w denote an input Arabic word ماء *mā'*, which has the gloss `water`, i.e. `water` $\in \alpha(w)$. w shares this gloss with the stem نضح *ndḥ* , denoted s_1, i.e. $s_1 \in Sy^1(w)$. Next, the stem رشّ *ršš* , denoted s_2, shares the gloss `spray` with s_1, i.e. $s_2 \in Sy^1(s1) \subset Sy^2(w)$. Therefore, $Syn^2(w)$ relates the words ماء *mā'* and رشّ *ršš* .

4.2. *MRE: Morphology-based regular expressions*

Let $\mathcal{O} = \{isA, contains\}$ be the set of atomic term predicates, where isA and $contains$ denote exact match and containment, respectively. Also, let $\mathcal{F} = \{\mathcal{P}, \mathcal{S}, \mathcal{X}, POS, GLOSS, CAT\}$ be the set of morphological features where each morphological feature $A \in \mathcal{F}$ is in turn a set of morphological feature values. Given a word w, a user defined constant feature value $CF \in A$, and an integer $k, 1 \leqslant k \leqslant 7$, the following are morphology-based atomic terms (MAT), *terms* for short.

– $a(w) := \exists m \in M(w).\ m = \langle p, s, x, P, G, C \rangle.r \circ CF$ where $\circ \in \mathcal{O}$, $r \in \{p, s, x, P, G, C\}$, and $r \in A$. Informally, a solution vector of w exists with a feature containing or exactly matching the user-chosen feature value CF.

MBF	description	formula	matches
N	name of person	$category = Name\_of\_Person$	n_1, n_2, n_3
P	name of place	$category = Name\_of\_Place$	$p_1, p_2, ..., p_7$
R	relative position	$stem \in \{$قرب,في,...$\}$	r_1, r_2, r_3, r_4
U	numerical term	$stem \in \{$أول,ثاني,...$\}$	u_1, u_2

Table 2. *Boolean formulae corresponding to task in Figure 1.*

– $a(w) := w \in Syn^k(CF), CF \in \mathcal{S}$. Informally, this checks if w is an extended synonym of a stem CF. We limit k to a maximum of 7 since we practically noticed that (1) values above 7 introduce significant semantic noise and (2) the computation is expensive without a bound.

A morphology-based Boolean formula (MBF) is of the following form.

– a and $\neg a$ are MBF formulae where a is a MAT and $\neg$ is the negation operator.

– $(f \vee g)$ is an MBF where f and g are MBF formulae, and $\vee$ is the disjunction (union) operator.

Moreover, MERF provides O to be a default Boolean formula that tags all *other* words in the text that do not match a user defined formula. We also refer to those words as *null* words.

Consider the task we discussed in the introduction (Figure 1) and recall that we are interested in identifying names of people, names of places, relative positions, and numerical terms. Table 2 presents the defined formulae. The user denotes the "name of person" entities with formula N which requires the *category* feature in the morphological solution of a word to be `Name_of_Person`. The entities n_1, n_2, and n_3 are matches of the formula N in the text. Similarly, the user specifies formula P to denote "name of place" entities. The user specifies formula R to denote "relative position" entities, and defines it as a disjunction of terms that check for solutions matching stems such as قرب*qrb* ("near") and في*fy* ("in"). Similarly, U denotes numerical terms and is a disjunction of constraints requiring the stem feature to belong to a set of stems such as أول*wl* ("first"), ثَاني*ṯāny* ("second"), ... عَاشر*āšr* ("tenth").

Next, we define a morphology-based regular expression (MRE) as follows.

– m is an MRE where m is an MBF.

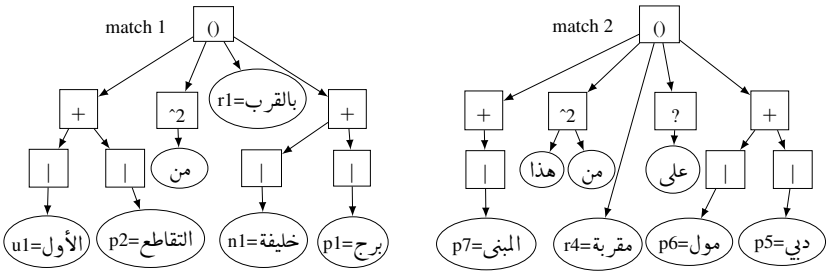


Figure 4. *Matches of regular expression* $(P|N)+\ O?\ R\ O\hat{}2\ (P|N|U)+$.

– fg is an MRE where f and g are both MRE expressions. A match of f followed by a match of g satisfies this concatenation operation.

– $f*$, $f+$, $f\hat{}x$, and $f?$ are MRE where f is an MRE, and are satisfied by zero or more, one or more, up to x matches, and an optional single match of f, respectively.

– $f\&g$, (conjunction) and $f|g$ (disjunction) are MRE where f and g are MRE, and are satisfied by the intersection of f and g matches, and the union of the f and g matches, respectively.

We denote by $[\![f]\!]$ the set of matches of an MRE f.

Back to the example in Figure 1. We use the formulae defined in Table 2 to construct an MRE such as $(P|N)+\ O?\ R\ O\hat{}2\ (P|N|U)+$ where $|, +, ?$, and $\hat{}\,k$ denote disjunction, one or more, zero or one, and up to k matches, respectively. The expression specifies a sequence of places or names of persons, optionally followed by a null word, followed by one relative position, followed by up to two possible null words, followed by one or more match of name of place, name of person, or numerical term. $O?$ and $O\hat{}2$ are used in the expression to allow for flexible matches.

The matching parse trees in Figure 4 illustrate two matches of the expression computed by MERF. The first tree refers to the text برج خليفة بَالقرب من التقَاطع الأول*brǧ ẖlyfh bālqrb mn āltqāṭ͑ āl-ʷwl* (“Khalifa Tower next to the first intersection”). The second tree refers to the text دبي مول علَى مقربة من هذَا المبنَى*dby mwl ɬā mqrbh mn hḏā ālmbnā* (“Dubai Mall is

located near this building"). The leaf nodes of the trees are matches to formulae and the internal nodes represent roots to subexpression matches. For instance, برج خليفة*brǧ ẖlyfh* in match 1 tree corresponds to the subexpression $(P|N)+$.

4.3. *User-defined relations and actions*

A relation is defined by the user as a tuple $\langle e_1, e_2, r\rangle$ where e_1, e_2, and r are identifiers associated with subexpressions of an MRE f. Matches of the relation are a set of labeled binary edges where matches of e_1 and e_2 are the source and destination nodes and matches of r are the edge labels. We denote $[\![\langle e_1, e_2, r\rangle]\!]$ to be the set of matches of the corresponding relation, and we refer to them as relational entities.

We are interested in constructing the relational entity graph in Figure 1. Let e_1, o_1, r, o_2,, and e_2 be identifiers to the subexpressions $(P|N)+$, $O?$, R, $O \wedge 2$, and $(P|N|U)+$, respectively. The matches to e_1, r, o_2, and e_2 in match 1 (Fig. 4) are برج خليفة*brǧ ẖlyfh* ("Khalifa Tower"), بالقرب*bālqrb* ("next"), من*mn* ("to"), and التقاطع الأول*āltqāṭ ̣ āl-ʾwl* ("first intersection"). Note that there is no match to the optional O formula in match 1. Similarly, the matches to e_1, o_1, r, o_2, and e_2 in the second matching tree are دبي مول*dby mwl* ("Dubai Mall"), علىٰ*lā* ("is located"), مقربة*mqrbh* ("near"), من هذا*mn hḏā* ("this"), and المبنى*ālmbnā* ("building"), respectively.

We define the semantic relations $\langle e_1, e_2, r\rangle$, $\langle r, e_1, o_1\rangle$, and $\langle r, e_2, o_2\rangle$. Relation $\langle e_1, e_2, r\rangle$ creates the edge labeled `next to` between `Khalifa tower` and `intersection 1` nodes from match 1, and the edge labeled `near` between `Dubai Mall` and `the building` nodes from match 2. Relation $\langle r, e_1, o_1\rangle$ creates the edge labeled `prep` between `Dubai Mall` and `near` nodes from match 2. Relation $\langle r, e_2, o_2\rangle$ creates the edge labeled `from` between `intersection 1` and `next to` nodes in match 1, and the edge labeled `from this` between `near` and `the building` nodes in match 2.

Moreover, MERF allows advanced users to write C++ code snippets to process matches of subexpressions. Each subexpression can be associated with two computational actions: `pre-match` and `on-match`. MERF provides an API that enriches the actions with detailed access to all solution features of an expression or a formula match including text, position, length, equivalent numerical value when applicable, and morphological features. The API follows a decorator pattern in that it incrementally adds the action results to the matching entities. Once MERF computes all matching parse

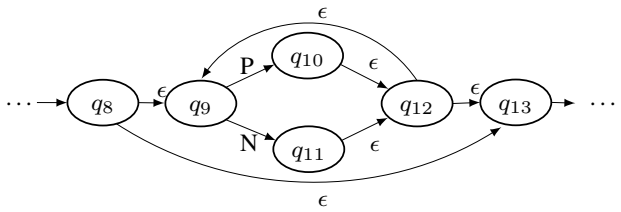


Figure 5. *Equivalent NFA of direction expression.*

trees, it traverses each tree to execute the user defined `pre-match` actions in pre-order manner and the `on-match` actions in post-order manner. This follows an observer pattern that notifies listeners with each produced match.

4.4. MERF *simulators*

The set of tag types $\mathcal{T}$ contains tuples of the form $\langle l, f, d\rangle$ where l is a text label with a descriptive name, f is an MRE, and d is a visualization legend with font and color information. For the example of Figure 1, l is "direction", f is $(P|N)+\ O?\ R\ O\hat{}2\ (P|N|U)+$, and d is italic.

For each word $t_i \in T, 0 \leqslant i < |T|$. MERF computes a Boolean value for all MBFs. For example, *brğ* برج matches MBF P. Then, it computes the set of MBF tags $R_i = \{(t_i, tt) | tt = \langle l, f, d\rangle \wedge f\ is\ an\ MBF \wedge f(t_i)\} \subseteq T \times \mathcal{T}$ which tags a word t_i with tt iff the MBF f associated with tag type tt is true for t_i. The MBF evaluation results in a sequence of tag sets $\langle R_0, R_1, \ldots, R_{n-1}\rangle$. If a word t_i has no tag type match, its tag set R_i is by default the singleton $O = \{NONE\}$. For example, the tag sets for the text in Figure 2 follows $\{\{NONE\}, \{NONE\}, \{NONE\}, \{NONE\}, \{($ *brğ* برج$, P)\}, \{($ *hlyfh* خليفة$, N)\}, \ldots\}$.

For each MRE, MERF generates its equivalent non-deterministic finite automaton (NFA) in the typical manner (Sipser, 2012). We support the upto operation ($f\hat{}x$), which is not directly supported in Sipser (2012), by expanding it into a regular expression form; for example $f\hat{}3$ is equivalent to $f?|ff|fff$. Consider the example of Figure 1 and the corresponding expression $(P|N)+\ O?\ R\ O\hat{}2\ (P|N|U)+$. Figure 5 shows part of the corresponding NFA where $q_8, q_9, \ldots, q_{13}$ represent NFA states, and edges are transitions based on MBF tags such as P, and N. Edges labeled with the empty string ϵ are non-deterministic.

MERF simulates the generated NFA over the sequence of tag sets matching the MBF formulae. A simulation match m of an expression f is a parse tree where the root spans the expression, the internal nodes are roots to subexpressions of f, and the leaves are matches of the MBF formulae of f, e.g. Figure 4. The sequence of leaf matches forms a vector of tags $\langle r_k, r_{k+1}, \ldots, r_j\rangle$ corresponding to the text sequence $\langle t_k, t_{k+1}, \ldots, t_j\rangle$ where $r_\ell \in R_\ell, 0 \leqslant k \leqslant \ell \leqslant j < n$. If we have more than one match for an expression, MERF returns the longest.

Finally, MERF computes the relational entities corresponding to each user defined relation $[\![\langle e_1, e_2, r\rangle]\!] \subseteq [\![e_1]\!] \times [\![e_2]\!] \times [\![r]\!]$.

5. MERF GUI

MERF provides a user friendly interface to specify the atomic terms, the MBFs, the MREs, the tag types, and the legends. The GUI also allows the user to modify and correct the tag set R. The GUI allows the user also to compute accuracy results that compare different tag sets and that can serve well as inter annotation agreement results when the tag sets come from two human annotators, or as evaluation results when comparing with reference tag sets.

5.1. *Tag type Boolean formula editor*

The user writes MBF tag types with the tag type editor introduced in Jaber and Zaraket (2013). First the user specifies atomic terms by selecting a feature from $\mathcal{F}$. The user can also choose whether to require an exact match using the `isA` predicate, or a substring match using the `contains` predicate option.

The user can add and remove feature values to the atomic terms using push buttons. A check box in the "Feature" column allows negating the term, and the "Relation" column switches the predicate between `isA` and `contains`. The list of feature and value pairs is interpreted as a disjunction to form the MBF. A right pane shows a description of the tag type and a set of legend descriptors. When the stem or gloss features are selected, the user has the option to use the Syn^k feature.

In the direction extraction task example, the user specifies four MBF-based tag types with labels N, P, R, and U with "name of person", "name of place", "relative position", and "numerical term" descriptions, respectively. For each MBF, the user selects the morphological features, specifies the constant value CF, and adds it to the Boolean formula editor.

5.2. *MBF match visualization*

The MBF match visualizer shows color sensitive text view, the tag list view, and the tag description view. The tag description view presents the details of the selected tag along with the relevant tag type information. The user can edit the tags using a context sensitive menus. MERF GUI also allows manual tag types and corresponding tags that are not based on morphological features. This enables building reference corpora without help from the morphological analyzer.

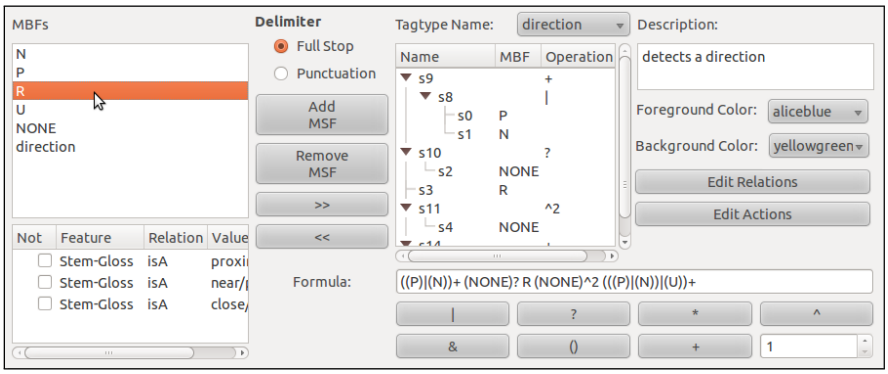


Figure 6. MERF *tag type regular expression editor.*

5.3. *Tag type regular expression editor*

After interacting with the MBF editor, the user moves to specify the regular expressions. The MRE editor of Figure 6 allows the definition of an MRE tag type in a user-friendly manner. The user first adds the required MBF formulae by selecting a label from $\mathcal{T}$ under MBFs. The Boolean formula of a highlighted tag type is shown in the table on the lower left pane. Each selected MBF is associated with an automatic name. The user can nest the MRE expression using a tree view of the MRE operations. The tree features the name, MBF, and operation for each subexpression.

To specify a binary operation the user selects two subexpressions and clicks the corresponding operation button. The operations include disjunction, conjunction, zero or one, sequence, zero or more, one or more, and up to a user defined constant. The right pane shows a description of the tag type and a set of legend descriptors.

5.4. *MRE match visualization*

While specifying an MRE, the user can interact with the visualization and editor views to make sure the MRE expresses the intent. The color-sensitive text view in Figure 7 shows the highlighted tag matches after the user called the MRE simulator using the `Tagtypes` menu.

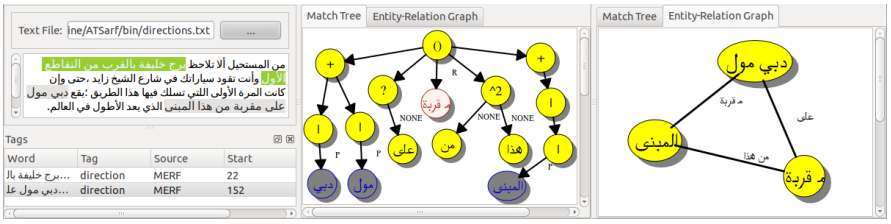


Figure 7. *MRE annotated Text, MRE matching parse tree, and entity-relation graph.*

The matching parse tree view shows the selected match in a graph view. Figure 7 shows the matching parse tree of the direction task دبي مول علَى مقربة من هذَا المبنَى *dby mwl ʿlā mqrbh mn hḏā ālmbnā* (“Dubai Mall is located near this building”).

5.5. *User defined relation editor*

After the user is satisfied with the MRE matches, the user moves to define relations and code actions. The relation editor allows the user to define relations by specifying $\langle e_1, e_2, r \rangle$ tuples, where e_1 and e_2 denote source and destination entities, and r denotes the label. The editor shows the MRE tree and allows the user to select the subexpressions and select features of the matches of the subexpressions to define the three components of the relation.

A snapshot of the GUI in Figure 7 shows in an interactive graph view the entity-relation graph of the match of the user defined relation extracted from the matching parse tree of the MRE. In the computational action editor, an advanced user can enter C++ code and use the MERF API to program and process subexpression matches.

5.6. *Analysis*

In the analysis view, the user provides two tag sets R_1 and R_2 and two tag type sets $\mathcal{T}_1$ and $\mathcal{T}_2$ as input. The tag type difference view shows the text annotated in three panes: (i) the common tag types $\mathcal{T}_1 \cap \mathcal{T}_2$, (ii) the tag types in $\mathcal{T}_1$ but not in $\mathcal{T}_2$, and (iii) the tag types in $\mathcal{T}_2$ and not in $\mathcal{T}_1$. Similarly, the tag difference view shows $R_1 \cap R_2$, R_1/R_2 and R_2/R_1 in addition to precision, recall and F-measure values. The user selects a predicate to compute the metrics from the following predicates: (1) “Intersection”: a tag from R_1 intersects in text with a tag in R_2, (2) “Exact”: a tag from R_1 exactly matches a tag in R_2, (3) “A includes B”: a tag from R_1 contains a tag from R_2, and (4) “B includes A”: a tag from R_2 contains a tag from R_1.

Features	MERF	SystemT	TEXTMARKER	Urbain	QARAB
Query type	MRE	AQL	matching rules	natural language	natural language
Morphology support	✓	-	-	OpenNLP	Parser
Relations	✓	-	-	✓	-
Actions	✓	-	-	-	-
Editor	✓	-	✓	-	-
Tag visualization	✓	-	✓	-	-
Graph visualization	✓	-	-	-	-

Table 3. *Comparison of* MERF *with SystemT, TEXTMARKER, Urbain, QARAB.*

6. Related Work

In this section we review the literature on entity and relation IE and on automatic and manual annotation techniques and compare to MERF.

Information Extraction. The common pattern specification language (CPSL) targets system independent IE specifications (Appelt and Onyshkevych, 1998). MERF extends CPSL with Arabic morphological features, code actions, and user defined relations. SystemT (Chiticariu *et al.*, 2010) aims to overcome the limitations of CPSL. It is based on an algebraic approach to declarative information extraction, uses the declarative annotation query language (AQL), and uses an optimizer to generate high performance execution plans for the AQL rules. MERF supports multiple tags per word, and supports the MRE conjunction operator which overcomes the overlapping annotation problem discussed in SystemT.

TEXTMARKER is a semi-automatic rule-based IE system for structured data acquisition (Atzmueller *et al.*, 2008). Both TEXTMARKER and MERF provide the user with GUI editor and result visualizer.

The work in Urbain (2012) presents a user-driven relational model and targets entity and relation extraction. The user enters a natural language query, and uses the OpenNLP toolkit to extract tags and relations from the query. Similar to MERF, the system constructs entities and relations.

QARAB is an Arabic question answering system that takes an Arabic natural language query and provides short answers for it (Hammo *et al.*, 2002). QARAB uses traditional information retrieval techniques and an outdated Arabic NLP analyzer with limited features of Arabic words compared to the morphological analysis of MERF.

Table 3 summarizes the comparison between MERF and other systems. MERF differs in that it provides code actions, user defined relations, and an interactive graph visualization of the relational entities. It also differs in that it fully supports Arabic morphological analysis while only QARAB supports Arabic linguistic features using a parser, and the work in Urbain (2012) uses OpenNLP that currently lacks full support for Arabic morphological features. Similar to TEXTMARKER, MERF has the advantage of providing a user-friendly interactive interface to edit the entity and relational specifications and visualize the results.

DUALIST is an annotation system for building classifiers for text processing tasks using machine learning techniques (Settles, 2011). MERF doesn't support classification tasks. However, MERF provides an interactive GUI where the user can edit MBF and MRE tags. This interactive environment contributes to the regular expression extraction and semantic relation construction which increases the overall accuracy.

Another track in the literature targets specific tasks such as NER using statistical and machine-learning techniques such as maximum entropy, optimized feature sets and conditional random fields (Benajiba *et al.*, 2007; Benajiba *et al.*, 2008; Ekbal and Bandyopadhyay, 2008; AbdelRahman *et al.*, 2010). Knowledge-based techniques such as Zaghouani *et al.* (2010) and Traboulsi (2009) propose local grammars with morphological stemming. Makhlouta *et al.* (2012) extract entities and events, and relations among them, from Arabic text using a hierarchy of manually built finite state machines driven by morphological features, and graph transformation algorithms. Such techniques require advanced linguistic and programming expertise.

WordNet is a lexical reference system that mimics human lexical memory and relates words based on their semantic values and their functional categories: nouns, verbs, adjectives, adverbs, and function words (Miller *et al.*, 1990). The Syn^k feature in MERF is inspired by WordNet.

Annotation tools. MMAX2 is a manual multi-level linguistic annotation tool with an XML based data model (Müller and Strube, 2006). BRAT (Stenetorp *et al.*, 2012) and WordFreak (Morton and LaCivita, 2003) are manual multi-lingual user-friendly web-based annotators that allow the construction of entity and relation annotation corpora. Knowtator (Ogren, 2006) is a general purpose incremental text annotation tool implemented as a Protégé (Gennari *et al.*, 2003) plug-in. Protégé is an open-source platform with a suite of tools to construct domain models and knowledge-based applications with ontology. However, it doesn't support the Arabic language.

MERF differs from MMAX2, BRAT, WordFreak, and Knowtator in that it is an automatic annotator that allows manual corrections and sophisticated tag type and relation specifications over Arabic morphological features.

Kholidy and Chatterjee (2010) present an overview of annotation tools and concludes with a set of rules and guidelines needed in an Arabic annotation alignment tool. The work in Dukes *et al.* (2013) presents a collaborative effort towards morphological and syntactic annotation of the Quran. Dorr *et al.* (2010) present a framework for interlingual annotation of parallel text corpora with multi-level representations. Kulick (2010) presents the integration of the Standard Arabic Morphological Analyzer (SAMA) into the workflow of the Arabic Treebank.

The work in Smrz and Pajas (2004) presents a customizable general purpose tree editor, with the Arabic MorphoTrees annotations. The MorphoTrees present the morphological analyses in a hierarchical organization based on common features.

Task specific annotation tools such as Alrahabi *et al.* (2006) use enunciation semantic maps to automatically annotate directly reported Arabic and French speech.

AraTation is another task specific tool for semantic annotation of Arabic news using web ontology based semantic maps (Saleh and Al-Khalifa, 2009). We differ in that MERF is general, and not task specific, and it uses morphology-based features as atomic terms. Fassieh is a commercial Arabic text annotation tool that enables the production of large Arabic text corpora (Attia *et al.*, 2009). The tool supports Arabic text factorization including morphological analysis, POS tagging, full phonetic transcription, and lexical semantics analysis in an automatic mode. Fassieh is not directly accessible to the research community and requires commercial licensing. MERF is open source and differs in that it allows the user to build tag types and extract entities and relations from text.

7. Results

In this section we evaluate MERF with four case studies. We perform a survey-like evaluation where developers manually built task specific information extraction tools for the case studies and other developers built equivalent MERF tools. The aim of the comparison is to showcase that MERF enables fast development of linguistic applications with similar accuracy and a reasonable affordable overhead in computational time. We report development time, size of developed code versus size of grammar, running time, and precision-recall as metrics of cost, complexity, overhead, and accuracy, respectively.

We survey three case studies from the literature: (1) narrator chain, (2) temporal entity, and (3) genealogy entity extraction tasks, and we use the reported development time for the task specific techniques proposed in ANGE (Zaraket and Makhlouta, 2012a), ATEEMA (Zaraket and Makhlouta, 2012c), and GENTREE (Makhlouta *et al.*, 2012), respectively. We also compare a MERF number normalization task to a task specific implementation.

We evaluated ANGE with *Musnad Ahmad*, a hadith book, where we constructed an annotated golden reference containing 1,865 words. We evaluated ATEEMA with articles from issues of the Lebanese *Al-Akhbar* newspaper where we constructed an annotated golden reference containing 1,677 words. For the genealogical tree extraction we used an extract from the Genesis biblical text with 1,227 words. Finally, we used an annotated article from the Lebanese *Assafir* newspaper with 1,399 words to evaluate the NUMNORM case study [2]. In the online appendix [3], we report on eight additional MERF case studies. Manual annotators inspected the outcome and provided corrections where tools made mistakes. The corrections form the manual gold annotation that we compared against.

Table 4 reports the development time, extraction runtime, recall and precision of the output MRE tags, the size of the task in lines of code or in number of MERF rules, for both the standalone task specific and the MERF implementations. The develop-

2. Available at `http://www.assafir.com` and `http://www.al-akhbar.com`.

3. Available at `http://research-fadi.aub.edu.lb/pdfs/merfappendix.pdf`.

Task	Size (words)	Development time	Run time(s)	Accuracy Recall	Accuracy Precision	Ease of Composition
ANGE	1,865	2 months	1.79	0.99	0.99	3,000+ lines of code
MERF		3 hours	7.24	0.99	0.93	8 MBFs and 4 MREs
ATEEMA	1,677	1.5 months	2.53	0.88	0.89	1,000+ lines of code
MERF		3 hours	3.14	0.91	0.81	3 MBFs and 2 MREs
Genealogy tree	1,227	3 weeks	0.74	0.96	0.98	3,000+ lines of code
MERF		4 hours	2.28	0.84	0.93	3 MBFs and 3 MREs
NUMNORM	1,399	1 week	0.32	0.91	0.93	500 lines of code
MERF		1 hour	1.53	0.91	0.90	3 MBFs/1 MRE/57 lines

Table 4. MERF *compared to task specific applications.*

ment time measures the time required for developing the case study. For instance, ANGE (Zaraket and Makhlouta, 2012a) required two months of development by a research assistant with 6 and 14 hours of course work and teaching duties, respectively. Recall refers to the fraction of the entities correctly detected against the total number of entities. Precision refers to the fraction of correctly detected entities against the total number of extracted entities.

Table 4 provides runtime results of MERF compared to the task specific implementations while running MBF and MRE simulations jointly. This is a rough estimate of the complexity of the MERF simulator. The complexity of the MBF simulation is the total number of morphological solutions for all the words multiplied by the number of user-defined MBFs. We do not provide a limit on the number of user defined formulae. In practice, we did not encounter more than ten formulae per case study. As for the complexity of MRE simulation, converting the rules into non-deterministic finite state machines (NDFSM) is done once. Simulating an NDFSM over the MBF tags is potentially exponential. In practice, all our case studies terminated within a predetermined time bound of less than 30 minutes. MERF required reasonably more runtime than the task specific implementations and reported acceptable and slightly less precision metrics with around the same recall.

Table 4 shows that MERF has a clear advantage over task specific techniques in the effort required to develop the application at a reasonable cost in terms of accuracy and run time. Developers needed three hours, three hours, four hours, and one hour to develop the narrator chain, temporal entity, genealogy, and number normalization case studies using MERF, respectively. However, the developers of ANGE, ATEEMA, GENTREE, and NUMNORM needed two months, one and a half months, three weeks, and one week, respectively. MERF needed eight MBFs and four MREs for narrator chain, three MBFs and two MREs for temporal entity, three MBFs and three MREs for genealogy, and three MBFs, one MRE, and 57 lines of code actions for the number normalization tasks. However, ANGE, ATEEMA, GENTREE, and NUMNORM required 3,000+, 1,000+, 3,000+, and 500 lines of code, respectively.

```
name:   PN ((MEAN)? PN)*;
nar:    name ((NONE)^3 FAM (NONE)^3 name)*;
pbuh:   BLESS GOD UPONHIM GREET;
nchain: (s1 =TOLD s2 =nar)+ ((PN|FAM|NONE)^8 pbuh)?
```

القعقاع	بن	عمارة	عن	جرير	حدثنا	سعيد	بن	قتيبة	حدثنا
ālqʿqāʿ	*bn*	*ʿmārh*	*ʿn*	*ǧryr*	*ḥdṯnā*	*sʿyd*	*bn*	*qtybh*	*ḥdṯnā*
PN	FAM	PN	TOLD	PN	TOLD	PN	FAM	PN	TOLD
name		name		name		name		name	
nar				nar		nar			
nchain									

Table 5. *Narrator chain example.*

7.1. *Narrator chain case study*

A narrator chain is a sequence of narrators referencing each other. The chain includes proper nouns, paternal entities, and referencing entities. ANGE uses Arabic morphological analysis, finite state machines, and graph transformations to extract entities and relations including narrator chains (Zaraket and Makhlouta, 2012a).

Table 5 presents the MREs for the narrator chain case study. MBF `PN` checks the abstract category `Name of Person`. MBF `FAM` denotes "family connector" and checks the stem gloss "son". MBF `TOLD` denotes referencing between narrators and checks the disjunction of the stems حدث("spoke to"), عن("about"), سمع("heard"), أخبر("told"), and أنبأ("inform"). MBF `MEAN` checks the stem عني("mean"). MBFs `BLESS`, `GOD`, `UPONHIM`, and `GREET` check the stems صلّى, الله, علي, and سلّم, respectively.

MRE *name* is one or more `PN` tags optionally followed with a `MEAN` tag. MRE `nar` denotes narrator which is a complex Arabic name composed as a sequence of Arabic names (`name`) connected with family indicators (`FAM`). The `NONE` tags in `nar` allow for unexpected words that can occur between names. MRE `pbuh` denotes a praise phrase often associated with the end of a hadith ("peace be upon him"), and is the satisfied by the sequence of `BLESS`, `GOD`, `UPONHIM`, and `GREET` tags. MRE `nchain` denotes narrator chain, and is a sequence of narrators (`nar`) separated with `TOLD` tags, and optionally followed by a `pbuh` tag.

Task	MBF accuracy		relation accuracy	
	Recall	Precision	Recall	Precision
Narrator chain	0.99	0.85	0.99	0.98
Number normalization	0.99	0.99	0.97	0.95
Temporal entity	0.99	0.52	0.98	0.89
Genealogy tree	0.99	0.75	0.81	0.96

Table 6. MERF *MBF and user-defined relation accuracy.*

The first row in Table 5 is an example narrator chain, the second is the transliteration, the third shows the MBF tags. Rows 4, 5, and 6 show the matches for `name`, `nar`, and `nchain`, respectively. MERF assigns the symbols s_1 and s_2 for the MRE subexpressions `TOLD` and `nar`, respectively. We define the relation $\langle s_2, s_2', s_1 \rangle$ to relate sequences of narrators with edges labeled by the tags of `TOLD` where s_2' denotes the next match of `nar` in the one or more MRE subexpression. Table 6 shows that MERF detected almost all the MBF matches with 99% recall and 85% precision and extracted user-defined relations with 98% recall and 99% precision.

7.2. *Temporal entity extraction*

Temporal entities are text chunks that express temporal information. Some represent absolute time such as الخَامس من آب ٢٠١٠ *ālẖāms mn ʾāb 2010*. Others represent relative time such as بعد خمسة أيّام *bʿd ẖmsh ʾayām*, and quantities such as يومَا ١٤ *14 ywmā*. *ATEEMA* presents a temporal entity detection technique for the Arabic language using morphological analysis and finite state transducers (Zaraket and Makhlouta, 2012c). Table 6 shows that MERF detected almost all the MBF matches with 99% recall, however it shows low precision (52%). As for the semantic relation construction, MERF presents a 98% recall and 89% precision.

7.3. *Genealogy tree*

Biblical genealogical lists trace key biblical figures such as Israelite kings and prophets with family relations. The family relations include wife and parenthood. A sample genealogical chunk of text is وولد هَارَان لوطَا *w wld hārān lwṭā* meaning "and Haran became the father of Lot". GENTREE (Makhlouta *et al.*, 2012) automatically extracts the genealogical family trees using morphology, finite state machines, and graph transformations. Table 6 shows that MERF detected MBF matches with 99% recall, and 75% precision, and extracted relations with 81% recall and 96% precision.

TMB algorithm

```
cout << $s1.text;
if(isHundred) {
  if(current != 0) {
    previous += current;
  }
  current = currentH * $s1.number;
  currentH = 0;
  isHundred = false;
  isKey = true;
} else if(current == 0) {
  current = $s1.number;
  isKey = true;
} else if(!isKey) {
  isKey = true;
  current = current * $s1.number;
} else {
  previous += current;
  current = $s1.number;}
```

DT algorithm

```
if(isHundred) {currentH += $s0.number;
} else if(current == 0) {
  current = $s0.number;
} else if(isKey) {
  previous += current;
  current = $s0.number;
} else {current += $s0.number; }
isKey = false;
```

H algorithm

```
isHundred = true;
if(current == 0)  {
  currentH = $s2.number;
} else if(!isKey) {
  currentH = current * $s2.number;
  current = 0;
} else {currentH = $s2.number;}
isKey = false;
```

Figure 8. *Actions for TMB, DT, and H MRE expressions.*

7.4. *Number normalization*

We implemented a number normalization extractor using MERF and compared it with *NUMNORM*, a C++ implementation for number normalization. First, we defined the MBFs `DT`, `H`, and `TMB` to denote (1) digits and tens, (2) hundreds, and (3) thousands, millions, and billions, respectively. The `num` MRE `(DT|TMB|H)+` is one or more `DT`, `TMB`, or `H` tags. MERF assigns the symbols s_1, s_2, and s_3 for the subexpressions `DT`, `TMB`, and `H`, respectively. Figure 8 shows the actions associated with the `DT`, `TMB`, and `H` subexpressions that cumulatively compute the numeric value of the numeric expression match. The actions use MERF API to access features of the matches such as text (`$s1.text`) and numeric value (`$s1.number`) of literal numbers such as numbers from one to ten. Table 6 shows high accuracy in MBF tagging and relation extraction with 99% and 97% recall and 99% and 95% precision, respectively.

7.5. *Discussion*

The results show that MERF provides a friendly environment to develop entity and relational entity extraction tasks with acceptable accuracy and runtime overheads compared to task specific applications. MERF requires the user to understand and interact with basic linguistic concepts such as readable values of morphological features, sequences, repetitions, and bounded repetitions. The user interacts with the MBF editor to specify basic concepts and visualize their matches over highlighted text. Then, the user interacts with the MRE editor to specify sequences of the concepts and visualize the matches in a graph, in conjunction with the highlighted text.

The two levels of interaction allow the user to separate between concepts that relate to word features, and more sophisticated entities that relate to sequences and context. The MBF, MRE, and user defined relations can be used to generate large annotated corpora in a fast manner. MERF visualization can be used later to refine the annotation. The case studies showed that MERF requires some linguistic expertise to successfully execute the tasks. In contrast, the case specific implementations require more sophisticated linguistic and programming expertise to attain similar results.

We notice that ANGE, ATEEMA, and Genealogy tree report higher precision than MERF. This is mainly due to their capacity to learn words and relations that may not have a match in the morphological analyzer based on co-occurrence relations. For example, the sequence $p_1 t_1 p_2$ where p_1 and p_2 are persons and t_1 is a tell relationship helps indicate that x is a tell relationship in $p_1 x p_2$ even if the morphological analyzer did not return the required feature for x to match a tell relationship. MERF does not have that capacity yet unless it is encoded in the C++ actions.

8. Conclusion

In this work, we present a morphology-based entity and relational entity extraction framework for Arabic text. MERF provides a friendly interface where the user defines tag types and associates them with regular expressions defined over Boolean formulae. The Boolean formulae are in turn defined over matches of Arabic morphological features and a novel extended synonymy feature (Syn^k). MERF allows the user to associate code actions with each regular subexpression and to define semantic relations between subexpressions. We evaluate MERF with several case studies and compare with existing application-specific techniques. The results show that MERF requires shorter development time and effort compared to existing techniques and produces reasonably accurate results within a reasonable overhead in run time. In the future, MERF will support user-defined cross-reference predicates, and will infer morphological features from relevant example words to express a concept.

9. Acknowledgment

The authors would like to thank the Lebanese National Council for Scientific Research (CNRS) for their support.

10. References

AbdelRahman S., Elarnaoty M., Magdy M., Fahmy A., "Integrated Machine Learning Techniques for Arabic Named Entity Recognition", *International Journal of Computer Science Issues*, vol. 7, n^{o} 4, p. 27-36, 2010.

Al-Sughaiyer I., Al-Kharashi I., "Arabic morphological analysis techniques: A comprehensive survey", *JASIST*, 2003.

Alrahabi M., Ibrahim A. H., Desclés J.-P., "Semantic Annotation of Reported Information in Arabic", *FLAIRS Conference*, vol. 6, p. 263-268, 2006.

Appelt D., Onyshkevych B., "The common pattern specification language", *TIPSTER workshop*, ACL, 1998.

Attia M., Rashwan M., Al-Badrashiny M., "Fassieh, a semi-automatic visual interactive tool for morphological, PoS-Tags, phonetic, and semantic annotation of Arabic text corpora", *IEEE transactions on audio, speech, and language processing*, vol. 17, n° 5, p. 916-925, 2009.

Atzmueller M., Kluegl P., Puppe F., "Rule-Based Information Extraction for Structured Data Acquisition using TextMarker", *Proceedings of LWA*, Citeseer, p. 1-7, 2008.

Benajiba Y., Diab M., Rosso P., "Arabic named entity recognition using optimized feature sets", *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, p. 284-293, 2008.

Benajiba Y., Rosso P., Benedíruiz J. M., "Anersys: An Arabic named entity recognition system based on maximum entropy", *International Conference on Intelligent Text Processing and Computational Linguistics*, Springer, p. 143-153, 2007.

Buckwalter T., Buckwalter Arabic Morphological Analyzer Version 1.0, Technical report, University of Pennsylvania, 2002.

Chiticariu L., Krishnamurthy R., Li Y., Raghavan S., Reiss F. R., Vaithyanathan S., "SystemT: an algebraic approach to declarative information extraction", *Proceedings of the Association for Computational Linguistics*, p. 128-137, 2010.

Dorr B. J., Passonneau R. J., Farwell D., Green R., Habash N., Helmreich S., Hovy E., Levin L., Miller K. J., Mitamura T. *et al.*, "Interlingual annotation of parallel text corpora: a new framework for annotation and evaluation", *NLE*, vol. 16, n° 3, p. 197-243, 2010.

Dukes K., Atwell E., Habash N., "Supervised collaboration for syntactic annotation of Quranic Arabic", *Language resources and evaluation*, vol. 47, n° 1, p. 33-62, 2013.

Ekbal A., Bandyopadhyay S., "Named Entity Recognition using Support Vector Machine: A Language Independent Approach", *IJCSSE*, 2008.

Ferilli S., "Natural Language Processing", *Automatic Digital Document Processing and Management*, Springer, 2011.

Gennari J., Musen M., Fergerson R., Grosso W., Crubézy M., Eriksson H., Noy N., Tu S., "The evolution of Protégé: an environment for knowledge-based systems development", *International Journal of Human-Computer Studies*, vol. 58, n° 1, p. 89-123, 2003.

Habash N., "Introduction to Arabic natural language processing", *Synthesis Lectures on Human Language Technologies*, 2010.

Habash N., Sadat F., "Arabic Preprocessing Schemes for Statistical Machine Translation", *NAACL*, p. 49-52, 2006.

Hammo B., Abu-Salem H., Lytinen S., "QARAB: A question answering system to support the Arabic language", *Computational approaches to semitic languages*, ACL, p. 1-11, 2002.

Jaber A., Zaraket F., "MATAr: Morphology-based Tagger for Arabic", *AICCSA*, May, 2013.

Kholidy H., Chatterjee N., "Towards developing an Arabic word alignment annotation tool with some Arabic alignment guidelines", *ISDA*, IEEE, p. 778-783, 2010.

Kulick S., "Consistent and flexible integration of morphological annotation in the Arabic Treebank", *LREC*, 2010.

Linckels S., Meinel C., "Natural Language Processing", *E-Librarian Service*, Springer, p. 61-79, 2011.

Maamouri M., Bies A., Buckwalter T., Mekki W., "The penn Arabic treebank: Building a large-scale annotated Arabic corpus", *NEMLAR Conference on Arabic Language Resources and Tools*, p. 102-109, 2004.

Makhlouta J., Zaraket F., Harkous H., "Arabic entity graph extraction using morphology, finite state machines, and graph transformations", *CICLing*, Springer, p. 297-310, 2012.

Marcus M. P., Marcinkiewicz M. A., Santorini B., "Building a large annotated corpus of English: The Penn Treebank", *Computational linguistics*, vol. 19, nº 2, p. 313-330, 1993.

Miller G. A., Beckwith R., Fellbaum C., Gross D., Miller K. J., "Introduction to WordNet: An on-line lexical database", *International journal of lexicography*, vol. 3, nº 4, p. 235-244, 1990.

Morton T., LaCivita J., "WordFreak: an open tool for linguistic annotation", *HLT/NAACL*, 2003.

Müller C., Strube M., "Multi-level annotation of linguistic data with MMAX2", *Corpus technology and language pedagogy: New resources, new tools, new methods*, 2006.

Nolan D., Lang D., "JavaScript Object Notation", *XML and Web Technologies for Data Sciences with R*, Springer, 2014.

Ogren P., "Knowtator: a protégé plug-in for annotated corpus construction", *NAACL-Demonstrations*, ACL, 2006.

Pasha A., Al-Badrashiny M., Diab M. T., El Kholy A., Eskander R., Habash N., Pooleery M., Rambow O., Roth R., "MADAMIRA: A Fast, Comprehensive Tool for Morphological Analysis and Disambiguation of Arabic", *LREC*, vol. 14, p. 1094-1101, 2014.

Saleh L., Al-Khalifa H., "AraTation: an Arabic semantic annotation tool", *IIWAS*, 2009.

Settles B., "Closing the loop: Fast, interactive semi-supervised annotation with queries on features and instances", *Proceedings of EMNLP*, ACL, p. 1467-1478, 2011.

Shahrour A., Khalifa S., Taji D., Habash N., "Camelparser: A system for Arabic syntactic analysis and morphological disambiguation", *COLING Demonstrations*, p. 228-232, 2016.

Sipser M., *Introduction to the Theory of Computation*, Cengage Learning, 2012.

Smrz O., Pajas P., "Morphotrees of Arabic and their annotation in the TrEd environment", *NEMLAR International Conference on Arabic Language Resources and Tools*, 2004.

Soudi A., Neumann G., Van den Bosch A., *Arabic computational morphology: knowledge-based and empirical methods*, Springer, 2007.

Stenetorp P., Pyysalo S., Topić G., Ohta T., Ananiadou S., Tsujii J., "BRAT: a web-based tool for NLP-assisted text annotation", *EACL Demonstrations*, ACL, p. 102-107, 2012.

Traboulsi H., "Arabic named entity extraction: A local grammar-based approach", *IMCSIT*, IEEE, 2009.

Urbain J., "User-driven relational models for entity-relation search and extraction", *Proceedings of JIWES*, ACM, 2012.

Xue N., Xia F., Chiou F.-D., Palmer M., "The Penn Chinese TreeBank: Phrase structure annotation of a large corpus", *Natural language engineering*, vol. 11, nº 2, p. 207-238, 2005.

Zaghouani W., Pouliquen B., Ebrahim M., Steinberger R., "Adapting a resource-light highly multilingual Named Entity Recognition system to Arabic", *LREC*, p. 563-567, 2010.

Zaraket F. A., Makhlouta J., "Arabic Cross-Document NLP for the Hadith and Biography Literature", *FLAIRS*, May, 2012a.

Zaraket F., Makhlouta J., "Arabic Morphological Analyzer with Agglutinative Affix Morphemes and Fusional Concatenation Rules", *COLING*, Mumbai, India, December, 2012b.

Zaraket F., Makhlouta J., "Arabic Temporal Entity Extraction using Morphological Analysis", *IJCLA*, vol. 3, p. 121-136, 2012c.

Notes de lecture

Rubrique préparée par Denis Maurel

Université de Tours, LIFAT (Laboratoire d'informatique fondamentale et appliquée)

Thierry POIBEAU. Machine Translation. *The MIT Press*. 2017. 285 pages. ISBN 978-0-262-53421-5.

Lu par **Claire LEMAIRE**

Université de Grenoble Alpes – LIDILEM/GETALP

Machine Translation *se propose, en quinze chapitres de six à vingt-six pages, de donner une vue d'ensemble sur les progrès réalisés en traduction automatique depuis la Seconde Guerre mondiale. L'auteur aborde le sujet sous quatre angles : par l'histoire de la traduction automatique, les méthodes utilisées pour développer des outils, les types d'évaluation de ces outils, et enfin par l'industrie de la traduction automatique.*

L'ouvrage s'ouvre sur la difficulté de la conception d'outils de traduction automatique et sur le rappel que ces outils sont prévus pour des textes techniques et informatifs et non pour des textes littéraires. L'auteur s'interroge ensuite sur l'acte de traduire, et sur ce que représentent de bonnes traductions. Il propose les cinq critères que sont le respect de la structure du texte source, de son ton, de son style et de ses idées, et la fluidité de la lecture. Par ailleurs, la qualité dépend également du destinataire et de l'usage du texte, or deux tendances coexistent en traductologie : soit le traducteur cherche à rester proche de la forme de la langue source, soit au contraire proche de la forme de la langue cible. Enfin, la difficulté d'une modélisation informatique réside dans la subjectivité des critères de qualité, dans l'ignorance des processus cognitifs humains à reproduire, ainsi que dans le manque de précision des langages naturels.

1. Histoire de la traduction automatique (TA)

L'historique présente tout d'abord une figure du triangle de Vauquois, qui décrit les différents niveaux de transfert d'une langue à une autre lors du processus de TA, et se poursuit par une présentation des différents types de systèmes. Reprenant les publications, entre autres de John Hutchins, l'auteur s'intéresse aux réflexions qui ont précédé la question de la TA, retrace l'évolution en Occident des concepts de langue universelle et de code numérique, ainsi que des premières machines qui ont suivi. Sont ensuite décrits les premiers essais de TA, par transfert direct, puis par transfert syntaxique avec les systèmes à base de règles, aux États-Unis et en Europe, sur une période qui part de 1949 avec le premier document précisant le problème de la TA, le fameux mémorandum de Warren Weaver, et qui se termine en 1959 avec

le rapport de Bar-Hillel et son questionnement sur le bien-fondé de travaux sur le sujet.

La dernière partie de cet historique relate tout d'abord la suppression d'une grande partie des budgets, dans le monde anglophone, due au rapport *Automatic Language Processing Advisory Committee* (ALPAC), en 1966, annonçant la quasi-impossibilité de traduire automatiquement, puis la période plus calme qui a suivi, au cours de laquelle se sont tout de même poursuivis des travaux dans différents centres de recherche et dans différentes sociétés commerciales. Cette période difficile pour le domaine se termine au début des années 1980, avec une recrudescence de l'intérêt pour la TA et une diversification des travaux et des méthodes.

2. Méthodes utilisées aujourd'hui pour développer des outils

Cette partie commence par aborder les méthodes *empiriques*, par opposition à la méthode *experte*. Alors que la méthode experte est fondée sur la modélisation des langues concernées, une méthode empirique repose sur un gigantesque corpus de phrases bilingues. Dans la première méthode empirique, la TA *basée sur l'exemple*, on entre un segment (un ou plusieurs mots) à traduire et, si un segment semblable est stocké dans le corpus bilingue, le système le retrouve et le propose. Si le système ne retrouve pas de segment semblable, il produit une traduction par analogie.

Dans une deuxième méthode empirique appelée TA *probabiliste* (ou *statistique*), on commence par une longue phase d'entraînement sur le corpus bilingue. On entre une phrase à traduire, le programme calcule, grâce à des modèles d'alignement, la phrase cible la plus probable. L'auteur décrit également les différentes approches de création et de traitement des corpus bilingues, puis les modèles d'alignement d'IBM, utilisés pour entraîner les systèmes, qui commencent par des probabilités de traduction au niveau lexical et qui se terminent par un réordonnancement des mots dans la langue cible.

La méthode experte, également appelée *par règles* (pour des raisons historiques), est aujourd'hui intéressante pour traduire des langues peu dotées, puisqu'il n'existe pas assez de données pour entraîner un système probabiliste. L'auteur propose Apertium qui est un logiciel libre adapté pour développer rapidement son propre système de TA, mais qui ne fonctionne bien que pour des langues proches syntaxiquement.

La troisième méthode empirique est le *deep learning*, également appelée *traduction neuronale*, car elle fait appel aux réseaux neuronaux profonds déjà connus en intelligence artificielle. À partir du corpus bilingue, on indique une correspondance exacte entre un segment source et un segment cible à obtenir en sortie. Ensuite, on prépare le segment source en définissant des caractéristiques (genre, lemme, etc.) pondérées et en attribuant un vecteur. Dans le moteur neuronal, on applique une fonction mathématique, puis on récupère le résultat que l'on décode. L'auteur précise que la prochaine étape en traduction neuronale sera donc de maîtriser ces caractéristiques et leur pondération, pour l'instant très obscures.

3. Types d'évaluation de ces outils

Les premières méthodes d'évaluation faisaient appel à des humains, soit des monolingues répondant par QCM à des questions de compréhension d'un texte automatiquement traduit, soit des traducteurs post-éditant ou jugeant directement les traductions. Puis ont suivi les méthodes d'évaluation automatique, qui comparent une traduction du système de TA à une traduction de référence produite par un humain. Si les deux textes sont strictement identiques, le score maximal est atteint. BLEU est la mesure la plus utilisée, NIST reprend l'idée précédente et accorde plus de poids aux groupes de mots plus rares. METEOR ajoute une recherche de mots porteurs de sens (et de leurs lemmes) ; meilleure, mais moins facile à utiliser, cette dernière méthode a moins de succès.

L'auteur reconnaît qu'aucune des méthodes automatiques n'évalue réellement la qualité des traductions, mais constate que les scores fournis donnent tout de même des indications sur le système de TA testé, soit par rapport à ses résultats précédents, soit par rapport aux autres systèmes de TA entraînés avec les mêmes données.

Depuis 2005, la conférence *Workshop on Machine Translation* (WMT) propose des campagnes d'évaluation. Les résultats des tests sont certes fonction des systèmes, mais également des langues : les plus riches morphologiquement, comme l'allemand et ses mots composés, sont pénalisées et les couples de langues proches sont favorisés.

4. Industrie de la TA

Les acteurs principaux sont Systran, Google, PROMT, IBM et Microsoft. Depuis sa création en 1968 (sous le nom de Latsec Inc.), Systran a été en partenariat avec les services du département de la Défense des États Unis, puis avec la Commission européenne. La firme, qui propose une solution gratuite en ligne *Systranet*, a été rachetée en 2014 par le Coréen CSLi, développeur des systèmes d'analyse vocale et de TA utilisés par les téléphones et tablettes Samsung. ProMT, le concurrent russe de Systran, vend des services de traduction dans des sites Web, payables au nombre de mots par mois, tout comme IBM avec sa plate-forme *IBM WebSphere*. Google a développé *Google Translate* initialement basé sur le système *Systran*, l'auteur met cependant en garde les entreprises sur l'absence totale de confidentialité de ce système qui conserve tous les textes sources des utilisateurs. Enfin, Microsoft propose *Bing Translator* dans ses logiciels. Pour conclure, l'auteur note que Google, mais aussi Apple et Facebook, rachètent des start-up du domaine et montent des centres de recherche en traduction neuronale recrutant actuellement des chercheurs au niveau mondial, puis cite les applications consommatrices de TA.

L'ouvrage se referme sur une TA pleine d'avenir et sur les tout derniers systèmes neuronaux capables de prendre en compte, au cours du processus, des phrases entières.

Un glossaire d'une quarantaine de mots du TAL et un index complètent le livre, ainsi qu'une bibliographie des sources d'inspiration de l'auteur, particulièrement détaillée et bienvenue pour quiconque souhaite se forger une petite culture dans le domaine.

Caroline BARRIÈRE. Natural Language Understanding in a Semantic Web Context. *Springer*. 2016. 318 pages. ISBN 978-3-319-41335-8.

Lu par **Thierry POIBEAU**

Lattice-CNRS

Le livre de Caroline Barrière se présente comme un point d'entrée pour les spécialistes du Web sémantique qui souhaiteraient s'initier au traitement automatique des langues (TAL), même si le contenu peut bien évidemment aussi intéresser un public plus large, ayant tout simplement un intérêt pour le TAL. L'autrice détaille avant tout les techniques d'extraction d'informations (c'est-à-dire l'extraction d'informations ciblées à partir de textes non structurés, pour remplir des bases de données structurées). Ce domaine est particulièrement intéressant pour le Web sémantique, qui a précisément pour but de produire des bases de connaissances et/ou d'utiliser des bases de connaissances dans le cadre d'applications « intelligentes », s'adaptant au contexte ou pouvant, par exemple, engager une forme de dialogue avec l'utilisateur. Du coup, le livre laisse volontairement de côté certains domaines importants du TAL, comme la recherche d'informations ou la traduction automatique, qui sont a priori *jugés moins pertinents dans ce contexte. L'ouvrage est ainsi homogène, cohérent et très progressif. Le titre est toutefois un peu large : ce qui est visé, c'est une compréhension locale et limitée du contenu textuel ; il existe d'autres approches de la compréhension, plus ambitieuses, mais évidemment beaucoup moins opérationnelles, qui ne sont pas prises en considération dans cet ouvrage.*

Le livre est divisé en quatre parties. La première partie est consacrée à la présentation des techniques de reconnaissance d'entités nommées dans les données textuelles. La deuxième porte sur la notion de corpus et la recherche de séquences (n-grammes) dans des textes monolingues ou bilingues. La troisième explore des questions d'annotation de corpus (annotation morphosyntaxique et syntaxique) ; sur cette base, la section présente ensuite différentes méthodes de calcul de similarité entre séquences textuelles, et enfin le processus qui consiste à relier les formes de surface aux formes normalisées que l'on peut trouver dans des ressources comme des thésaurus ou des référentiels (techniques dites de « liage d'entité » ou « *entity linking* »). Enfin, la quatrième partie aborde les liens entre syntaxe et sémantique, et la question des relations entre entités au sein du texte. Le livre comprend ensuite trois annexes : un aperçu du Web sémantique pour le lecteur qui ne maîtriserait pas ce domaine ; des informations sur les plates-formes et les outils de TAL disponibles ; des listes de relations typiques, souvent utilisées dans les applications de Web sémantique. Enfin, le livre comprend un glossaire de deux cents termes couramment utilisés en TAL.

Le contenu de l'ouvrage est très pédagogique. Chaque notion nouvelle est introduite, expliquée et replacée dans le contexte plus large de l'ouvrage. Plusieurs

algorithmes classiques du domaine sont présentés (sous forme de pseudo-code) et expliqués en détail. Le lecteur est constamment invité à vérifier les résultats et mener des expériences par lui-même (en utilisant par exemple les ressources détaillées dans l'annexe 3). Chaque chapitre est accompagné d'exercices permettant au lecteur de vérifier qu'il a bien compris le contenu, les enjeux et les problèmes susceptibles de se poser. Chaque chapitre inclut enfin des lectures complémentaires permettant d'aller plus loin que ce qui est présenté dans le livre. De ce point de vue, on peut dire que le livre est clair, bien écrit, et remplit parfaitement sa fonction ; il est progressif et amène le lecteur pas à pas à aborder des questions de recherche plus ambitieuses (de la recherche d'entités à la recherche de relations entre entités, en passant par la question du « liage » avec une base de données). La focalisation de l'ouvrage sur le domaine de l'extraction d'informations (recherche et analyse des entités du texte, recherche de relations entre entités en vue de compléter des bases de données structurées) semble aussi pertinente, même si cela amène l'autrice à laisser de côté de nombreux autres domaines qui auraient aussi pu être pertinents (comme la recherche d'informations sémantiques, qui entretient pourtant aussi des liens étroits avec le monde du Web sémantique).

Pour ce qui est des limites de l'ouvrage, on regrettera juste que l'accent soit essentiellement mis sur les techniques symboliques ou statistiques simples, et que les approches à base d'apprentissage artificiel soient le plus souvent à peine évoquées dans les lectures complémentaires, sans que ce choix ne soit véritablement expliqué. Les approches par apprentissage sont pourtant aujourd'hui la norme pour ce qui concerne la reconnaissance des entités nommées, et on regrettera ainsi que ni le format d'annotation BIO (*begin*, *inside*, *outside*), ni les algorithmes de reconnaissances de séquences (par exemple les CRF, *Conditional random fields*, pourtant omniprésents pour la reconnaissance des entités nommées) ne soient présentés. Quoi que l'on pense de ces techniques, elles ont acquis un poids tel qu'il semble difficile de les passer sous silence dans un ouvrage sur le TAL focalisé sur les techniques d'extraction d'informations. Dans le chapitre sur la similarité sémantique, la notion de plongements de mots (*word embeddings*) est rapidement évoquée, mais elle aurait aussi pu être davantage détaillée. Il est vrai que celle-ci a peut-être moins d'importance dans une perspective d'application dans le monde du Web sémantique que pour le TAL au sens large, encore que cela serait à prouver.

En conclusion, on appréciera un ouvrage très pédagogique et pertinent pour le lecteur souhaitant se former aux domaines abordés par l'ouvrage. On peut regretter le peu de détails concernant les techniques récentes d'apprentissage artificiel, mais le lecteur pourra s'appuyer sur les lectures complémentaires indiquées dans l'ouvrage (très complet de ce point de vue !), pour aller plus loin et s'initier par lui-même aux techniques les plus récentes. Notons enfin que l'ouvrage aborde un large choix de notions, dans la mesure où la reconnaissance des liens entre entités implique une analyse morphosyntaxique, syntaxique et même sémantique. Ce livre permet donc d'établir un pont solide entre traitement automatique des langues et

ingénierie des connaissances, et répond, à ce titre, à un besoin certain, que ce soit dans le monde universitaire ou dans le monde industriel.

Shay COHEN. Bayesian Analysis in Natural Language Processing. *Morgan & Claypool publishers*. 2016. 246 pages. ISBN 978-1-62705-873-5.

Lu par **Jose G. MORENO**

Université Paul Sabatier Toulouse III – IRIT

L'ouvrage est organisé en huit chapitres et une annexe. Il est d'un très haut niveau du point de vue formel et très riche en commentaires scientifiques. C'est aussi une source complète de références qui constitue un état de l'art de chaque sujet traité. Dans certains cas, quand l'auteur ne développe pas complètement un sujet, il indique des références précises pour trouver des informations complémentaires. Cependant, même si l'ouvrage est très complet, il requiert une connaissance élevée de l'inférence bayésienne dans son ensemble. Il est très adapté à des étudiants en doctorat sur le sujet ou des sujets plus proches de l'apprentissage automatique que du traitement automatique des langues (TAL). Le résumé en fin de chaque section ne résume pas forcément la section, et donne plus des informations pertinentes que manquantes. Le revers de ce livre est qu'il ne constitue pas un matériel pédagogique adapté pour un cours introductif de l'inférence bayésienne, alors qu'il sera clairement utile dans un cours avancé.

Le chapitre 1 est un rappel sur la probabilité et les statistiques relatives au TAL bayésien. Il aborde des concepts de base tels que les variables aléatoires, l'indépendance entre les variables aléatoires, l'indépendance conditionnelle et les valeurs attendues des variables aléatoires ; les statistiques bayésiennes sont présentées brièvement, en montrant la façon dont elles diffèrent des statistiques fréquentistes (ou statistiques classiques). La majeure partie de ce chapitre peut être omise si le lecteur possède une connaissance de l'inférence bayésienne. Cependant, il est recommandé de jeter un coup d'œil pour identifier les notations, revoir des concepts aussi bien en informatique qu'en statistique et identifier les exemples de TAL pour les chapitres suivants dans lesquels les concepts sont utilisés dans des contextes réels sans donner trop d'exemples.

Le chapitre 2 présente l'analyse bayésienne en TAL à l'aide de deux exemples : (1) le modèle d'allocation latente de Dirichlet (LDA) et (2) la régression bayésienne pour le texte. Ce chapitre donne également un aperçu général du sujet. L'histoire générative est présentée et sera utilisée dans différents chapitres jusqu'au dernier. Le modèle graphique (*plate notation*) est présenté pour le modèle LDA, mais n'est pas trop utilisé dans les chapitres suivants.

Le chapitre 3 traite d'une composante importante de la modélisation bayésienne : l'*a priori*. Les *a priori* les plus couramment utilisés dans le TAL bayésien sont présentés, tels que la distribution de Dirichlet, les antécédents non informatifs

(inappropriés et uniformes ou basés dans l'information de Fischer comme l'*a priori* de Jeffreys), la distribution normale et d'autres. Ces concepts sont primordiaux pour la suite des explications, en particulier la distribution de Dirichlet qui sera rappelée jusqu'à la fin du livre. Les positions données aux *a priori* sont d'une grande importance pour les modèles bayésiens et leur combinaison est possible dans des modèles hiérarchiques d'*a priori*.

Le chapitre 4 examine des idées qui regroupent les statistiques fréquentistes et bayésiennes à travers l'estimation de la distribution postérieure. Il détaille les approches permettant de calculer une estimation ponctuelle d'un ensemble de paramètres tout en maintenant une mentalité bayésienne.

Le chapitre 5 aborde l'une des principales approches d'inférence dans les statistiques bayésiennes : la chaîne de Markov Monte-Carlo (MCMC). Il détaille les algorithmes d'échantillonnage les plus couramment utilisés dans le TAL bayésien, tels que l'échantillonnage de Gibbs et l'échantillonnage de Metropolis-Hastings. Dans ce chapitre, le problème de convergence MCMC est abordé. La recommandation finale des auteurs est d'être prudent lors de la présentation des résultats avec la MCMC, car la convergence de la chaîne n'est pas une garantie et, dans la plupart des cas, il faut être très attentif pour ne pas manquer la bonne configuration du modèle.

Le chapitre 6 traite d'une autre approche d'inférence importante dans le TAL bayésien, à savoir l'inférence variationnelle. Il décrit l'inférence variationnelle du champ moyen, l'algorithme de maximisation des vraisemblances variationnelles et les méthodes d'échantillonnage. Ce sont des approches pour obtenir une inférence approximative. La différence principale entre ces méthodes est que la dernière est présentée comme un problème d'optimisation. Dans le chapitre 5, l'auteur rappelle que les modèles, type MCMC, ne font pas de l'optimisation sur une fonction unique, comme c'est le cas pour l'échantillonnage de Gipps, leur objectif consiste à trouver la chaîne qui correspond dans l'espace des possibilités. Dans ce cas, les itérations sont fixes sans garantie de convergence. Les méthodes d'inférence variationnelle sont une option viable pour atténuer les problèmes de convergence.

Le chapitre 7 traite d'une technique importante de modélisation en TAL bayésien, à savoir la modélisation non paramétrique. Les modèles non paramétriques comme le processus de Dirichlet et le processus de Pitman-Yor sont présentés. Ce chapitre est particulièrement intéressant pour les chercheurs des méthodes non supervisées. L'explication du processus de Dirichlet et ses relations avec l'algorithme k-moyennes sont claires et intéressantes.

Le chapitre 8 traite des modèles de grammaires formelles pour le TAL, comme les grammaires probabilistes non contextuelles et les grammaires synchrones. La façon de les cadrer dans un contexte bayésien est aussi abordée en détail, en particulier à l'aide de modèles tels que les grammaires d'adaptation, l'utilisation de hiérarchies dans le processus de Dirichlet et d'autres. Les lectures complémentaires recommandées par l'auteur sont une riche collection de références pour les chercheurs intéressés par les grammaires bayésiennes et leurs applications.

Commentaires

Pour résumer, il est important de préciser que ce livre est très technique et précis. Malgré sa qualité, il n'est pas recommandé aux débutants. Le lecteur doit avoir une bonne connaissance des modèles probabilistes pour pouvoir suivre l'auteur. Il ne faut pas s'attendre à des exemples explicatifs et il faut l'utiliser comme une référence technique, plutôt en apprentissage automatique qu'en TAL. Côté exercices, ceux-ci sont aussi très focalisés sur l'apprentissage automatique et ont peu ou rien à voir avec des exercices réels en TAL. En revanche, le livre comprend deux annexes qui fournissent des informations supplémentaires afin de faciliter sa lecture. Pour les novices, les annexes doivent être abordées avant de démarrer ce livre. Enfin, signalons que ce livre contient des conseils sur le plan théorique et pratique de l'inférence bayésienne. Par exemple, après une explication détaillée de l'inférence variationnelle, il énonce des méthodes récentes qui diminuent la charge théorique du développeur d'une nouvelle méthode sans propager des erreurs de différentiation (très fréquents parmi les débutants selon le chapitre 5).

Ce livre est recommandé aux chercheurs (confirmés ou pas) qui veulent vérifier leur état de l'art, trouver une explication à certains résultats, trouver des sources d'inspiration ou une notation formelle pour présenter un nouveau modèle. Comme l'auteur le souligne dans ces derniers mots, il existe aujourd'hui une énorme possibilité, pour les chercheurs intéressés, à profiter des avantages du TAL bayésien et des approches neuronales en TAL. Le formalisme existant dans le TAL bayésien est clairement déficient dans le TAL neuronal, mais la performance de ce dernier est plus que remarquable. Ce livre est une belle présentation du point de vue bayésien pour le TAL, et le plus probable est qu'il restera une référence pendant un certain temps.

Résumés de thèses

Rubrique préparée par Sylvain Pogodalla

Université de Lorraine, CNRS, Inria, LORIA, F-54000 Nancy, France
sylvain.pogodalla@inria.fr

Maximin COAVOUX : mcoavoux@inf.ed.ac.uk

Titre : Analyse syntaxique automatique en constituants discontinus des langues à morphologie riche

Mots-clés : Traitement automatique des langues, analyse syntaxique automatique, arbres en constituants discontinus, systèmes de transitions, apprentissage profond, apprentissage multi-tâches.

Title: *Discontinuous Constituency Parsing of Morphologically Rich Languages*

Keywords: *Natural Language Processing, syntactic parsing, discontinuous constituency trees, transition systems, deep learning, multitask learning.*

Thèse de doctorat en Sciences du langage, UFR de linguistique, Laboratoire de Linguistique Formelle (LLF), UMR 7110, Université Paris Diderot, sous la direction de Benoît Crabbé (MC HDR, Université Paris Diderot). Thèse soutenue le 11/12/2017.

Jury : M. Benoît Crabbé (MC HDR, Université Paris Diderot, directeur), Mme Claire Gardent (DR, CNRS, LORIA, Nancy, rapporteur), M. Alexis Nasr (Pr, Aix-Marseille Université, rapporteur et président), M. Alexandre Allauzen (MC HDR, Université Paris-Sud, examinateur), M. Carlos Gómez Rodríguez (Profesor Contratado Doctor, Universidade da Coruña, La Corogne, Espagne, examinateur).

Résumé : *L'analyse syntaxique consiste à prédire la représentation syntaxique de phrases en langue naturelle sous la forme d'arbres syntaxiques. Cette tâche pose des problèmes particuliers pour les langues non configurationnelles ou qui ont une morphologie flexionnelle plus riche que celle de l'anglais. En particulier, ces langues manifestent une dispersion lexicale problématique, des variations d'ordre des mots plus fréquentes et nécessitent de prendre en compte la structure interne des mots-formes pour permettre une analyse syntaxique de qualité satisfaisante.*

TAL. Volume 58 — n° 3/2017, pages 131 à 133

*Dans cette thèse, nous nous plaçons dans le cadre de l'analyse syntaxique robuste en constituants par transitions. Dans un premier temps, nous étudions comment intégrer l'analyse morphologique à l'analyse syntaxique, à l'aide d'une architecture de réseaux de neurones basée sur l'apprentissage multitâches. Dans un second temps, nous proposons un système de transitions qui permet de prédire des structures générées par des grammaires légèrement sensibles au contexte telles que les LCFRS (*Linear Context-Free Rewriting System*). Enfin, nous étudions la question de la lexicalisation de l'analyse syntaxique. Les analyseurs syntaxiques en constituants lexicalisés font l'hypothèse que les constituants s'organisent autour d'une tête lexicale et que la modélisation des relations bilexicales est cruciale pour désambiguïser. Nous proposons un système de transition non lexicalisé pour l'analyse en constituants discontinus et un modèle de scorage basé sur les frontières de constituants et montrons que ce système, plus simple que des systèmes lexicalisés, obtient de meilleurs résultats que ces derniers.*

Yann MATHET : yann.mathet@unicaen.fr

Titre : Une contribution à la linguistique computationnelle et au TAL : de la sémantique de l'espace et du temps à l'annotation et aux mesures d'accord inter-annotateurs.

Mots-clés : Annotation, mesures d'accord inter-annotateurs, sémantique, temps.

Title: *A Contribution to Computational Linguistics and Natural Language Processing: from the Semantics of Space and Time to Annotations and Agreement Measures*

Keywords: *Annotation, agreement-measures, semantics, time.*

Habilitation à diriger des recherches en Informatique, GREYC, UMR 6072, Université de Caen Normandie, sous la direction de Marc Spaniol (Pr, Université de Caen). Habilitation soutenue le 05/12/2017.

Jury : M. Marc Spaniol (Pr, Université de Caen, directeur), M. Jean-Yves Antoine (Pr, Université François Rabelais, Tours, rapporteur), Mme Delphine Battistelli (Pr, Université Paris X, rapporteur), M. Massimo Poesio (Pr, University of Essex, Royaume-Uni, rapporteur), M. Frédéric Landragin (DR, CNRS, Lattice, Montrouge, examinateur), M. Pierre Zweigenbaum (DR, CNRS, Limsi, Orsay, examinateur).

Résumé : *This study addresses two different questions in the fields of Computational Linguistics (CL) and Natural Language Processing (NLP): the question of how to model natural language semantics, especially in space and time paradigms, and the question of how to annotate corpora. These seemingly different questions are tied by the fact that when studying how to model some linguistic phenomena, for instance in semantics, it is necessary to get annotated data related to these phenomena, first to get inspiring examples of what is really studied, and second to assess our models by confronting their productions with reference annotations. Precisely, because of these*

ties, my research domain has progressively widened from pure semantics questions to questions about annotation.

In my PhD thesis, I addressed spatio-temporal semantics as it appears in natural language. Most available models rely on so-called topological relations, where the very questions is "in what place is X located?" These models fail to render the semantics of many expressions which cannot be described in terms of being located into a place, nor in terms of going into (or getting out of) a place. For instance, the sentence "(the road / the car) circumvents the city" involves a complex relationship between the shape of the road or of the trajectory of the car and the city (in addition to a topological relation of exteriority). I introduced the limits of these models and proposed solutions. Subsequently, my work has been focusing more and more on the semantics of time, in collaboration with other computer scientists, and also a linguist. In particular, we have addressed the question of how repetition (iterative events) is conveyed in natural language, in such examples as: "every Thursday, they played cards. The game lasted about 2 hours", and how to model it. One of the main results of this study is that natural language is able to handle an iterative event as if it were a sole generic event. This is clearly visible in the second sentence by the use of the singular "the game" which surprisingly refers to a plurality of games. We have designed a model which accounts for that, and for a wide range of related phenomena.

At the same time, several collaborations in CL and NLP research projects led us to focus more and more on annotation process. In particular, the ANNODIS project consisted in creating and providing a discourse relations corpus, and made appear the need for new methods and tools to annotate texts. Together with a colleague, Antoine Widlöcher, we designed and developed a versatile annotation platform, namely Glozz, which not only fulfills the ANNODIS requirements, but also fits a wide range of projects worldwide. Producing annotations brings another question: how to make sure that annotations are valid? Consequently, we have studied the existing methods to assess annotations, and we found that most of them do not fit CL nor NLP purposes. In particular, CL and NLP mainly refer to linguistic streams (texts, videos), whereas most used assessment methods concern sets of independent items. As a consequence, in many cases, scholars do not use relevant measures to assess their annotations, which leads to strong biases in the results. Here again, we have proposed solutions with a new set of agreement measures, namely the Gamma family. Besides, this work goes along with a more general reflection on the principles of assessment methods, which is an additional contribution.

URL où le mémoire peut être téléchargé :
https://mathet.users.greyc.fr/