Automatic Difficulty Classification of Arabic Sentences

Nouran Khallaf, Serge Sharoff
School of Languages, University of Leeds
Leeds, LS2 9JT, United Kingdom mlnak,s.sharoff@leeds.ac.uk

Abstract

In this paper, we present a Modern Standard Arabic (MSA) Sentence difficulty classifier, which predicts the difficulty of sentences for language learners using either the CEFR proficiency levels or the binary classification as simple or complex. We compare the use of sentence embeddings of different kinds (fastText, mBERT , XLM-R and Arabic-BERT), as well as traditional language features such as POS tags, dependency trees, readability scores and frequency lists for language learners. Our best results have been achieved using fined-tuned Arabic-BERT. The accuracy of our 3-way CEFR classification is F-1 of 0.80 and 0.75 for Arabic-Bert and XLM-R classification respectively and 0.71 Spearman correlation for regression. Our binary difficulty classifier reaches F-1 0.94 and F-1 0.98 for sentence-pair semantic similarity classifier.

1 Introduction

In the last century, measuring *text readability (TR)* has been undertaken in education, psychology, and linguistics. There appears to be some agreement that TR is the quality of a given text to be easy to comprehend by its readers in adequate time with reasonable effort (Cavalli-Sforza et al., 2018). Research to date has tended to focus on assigning readability levels to whole text rather than to individual sentences, despite the fact that any text is composed of a number of sentences, which vary in their difficulty (Schumacher et al., 2016). Assigning readability levels for a text is a challenging task and it is even more challenging on the sentence level as much less information is available. Also, the sentence difficulty is influenced by many parameters, such as, genre or topics, as well grammatical structures, which need to be combined in a single classifier. Difficulty assessment at the sentence level is a more challenging task in comparison to the better researched text level task, but the availability of a readability sentence classifier for Arabic is vital, since this is a prerequisite for research on *automatic text simplification* (ATS), i.e. the process of reducing text-linguistic complexity, while maintaining its meaning (Saggion, 2017).

We focus here on experiments aimed at measuring to what extent a sentence is understandable by a reader, such as a learner of Arabic as a foreign language, and at exploring different methods for readability assessment. The main aim of this paper lies in developing and testing different sentence representation methodologies, which range from using linguistic knowledge via feature-based machine learning to modern neural methods.

In summary, the contributions of this paper are:

1. We compiled a novel dataset for training on the sentence level;
2. We developed a range of linguistic features, including POS, syntax and frequency information;
3. We evaluated a range of different sentence embedding approaches, such as fastText, BERT and XLM-R, and compared them to the linguistic features;
4. We cast the readability assessment as a regression problem as well as a classification problem;
5. Our model is the first sentence difficulty system available for Arabic.

2 Corpora and Tools

2.1 Dataset One: Sentence-level annotation

This dataset was used for Arabic sentence difficulty classification. We started building our own dataset by compiling a corpus from three available source classified for readability on the document level along with a large Arabic corpus obtained by

Web crawling.

The first corpus source is the reading section of the **Gloss**[1] Corpus developed by the Defense Language Institute (DLI). It has been treated as a gold standard and used in the most recent studies on document level predictions (Forsyth, 2014; Saddiki et al., 2015; Nassiri et al., 2018a,b). Texts in Gloss have been annotated on a six level scale of the Inter-Agency Language Roundtable (IL), which has been matched to the CEFR levels according to the schema introduced by (Tschirner et al., 2015). Gloss is divided according to the four competence areas (lexical, structural, socio-cultural and discursive) and ten different genres (culture, economy, politics, environment, geography, military, politics, science, security, society, and technology).

The second corpus source is the **ALC** , which consists of Arabic written text produced by learners of Arabic in Saudi Arabia collected by (Alfaifi and Atwell, 2013). Each text file is annotated with a proficiency level of the student. We mapped these student proficiency levels to CEFR levels.

Our third corpus source comes from textbook **"Al-Kitaab fii TaAallum al-Arabiyya"** (Brustad et al., 2015) which was compiled from texts and sentences from parts one and two of the third edition but only texts from part three third edition. This book is widely used to teaching Arabic as a second language. These texts were originally classified according to the American Council on the Teaching of Foreign Languages (ACTFL) guidelines which we mapped to CEFR levels.

As these corpora have been annotated on the document level and not on the sentence level, we assigned each sentence to the document level in which it appears, by using several filtering heuristics, such as sentence length and containment, as well as via re-annotation through machine learning, see the dataset cleaning procedure below.

A counterpart corpus of texts not produced for language learners in mind is provided by I-AR, 75,630 Arabic web pages collected by wide crawling (Sharoff, 2006). A random snapshot of 8627 sentences longer than 15 words was used to extend the limitations of C-level sentences coming from corpora for language learners.

Table 1 shows distribution of the number of used sentences and tokens per each Common European Framework of language proficiency Reference [CEFR] Level. In principle we have data for 5-way (A1, A2, B1, etc), 3-way (A, B or C) and binary (A+B vs C) classification tasks, but here in this presentation, we focus on the 3-way and binary (simple vs complex) classification tasks.

[1] https://gloss.dliflc.edu/

CEFR	**Old**		**New**	
	S	T	S	T
A	8661	187225	9030	195343
B	5532	126805	5083	117825
C	8627	287275	8627	287275
Total	22820	601305	22740	600443

Table 1: (S)sentences and (T)tokens available per each CEFR Level in the two versions of the corpus

Dataset cleaning: In our initial experiments we noticed unreliable sentence-level assignments in the training corpus. Therefore, we decided to improve the quality of the training corpus by an error analysis strategy introduced by Di Bari et al. (2014), which is based on detecting agreement between classifiers belonging to different Machine Learning paradigms. The cases when the majority of the classifiers agreed on predicting a label while the gold standard was different were inspected manually by a specialist in teaching Arabic. In our Dataset cleaning experiment we used the following classifiers: SVM (with the rbf kernel), Random Forest, KNeighbors, Softmax and XgBoost using linguistic features discussed in Section 3, trained them via cross-validation and compared their majority vote to the gold standard.

We modified the error classification tags introduced by Di Bari et al. (2014) as follows:

Wrong if the classifiers have wrongly labelled the data, and the gold standard is correct.

Modify if the classifiers are correct and we need to modify the gold standard.

Ambiguous if we consider both either label is possible based on different perspectives.

False is an added label which represent the disagreement between the gold standard and the classifiers, when neither is correct.

For each sentence, five different predictions are assigned. Compared to the gold standard CEFR-label, the classifiers agreed in predicting 10204 instances. Then what we need to consider is when all classifiers agree on the predicted label and it contradicts with the gold standard's one. In that matter, the classifiers agreed on 1943 sentence clas-

sification. We manually investigated random sentences and assigned the error classification tags. We found that the main classification confusion was in Level B instances. The analysis results as in Table 4 show the distribution of categories where each error type occurred. In the end, 380 instances had to be assigned to lower level (usually from B to A).

2.2 Dataset Two: Simplification examples

A set of simple/complex parallel sentences has been compiled from the internationally acclaimed Arabic novel **"Saaq al-Bambuu"** (Al-Sanousi, 2013) which has an authorized simplified version for students of Arabic as a second language (Familiar, 2016). We assume that a successful classifier should be able to detect sentences in the original text that require simplification. Dataset Two consists of 2980 parallel sentences Table 2.

Level	Sentence	Token
Simple A+B	2980	34447
Complex C	2980	46521
Total	5690	80968

Table 2: Number of Sentences and Tokens available per each CEFR Level in Dataset two

3 Features and extraction methods

We work with following groups of features in Table 3: Part of speech tagging features (POS-features); Syntactic structure features (Syntactic-features); CEFR-level lexical features; Sentence embeddings.

3.1 Linguistic features

While the sentence-level classification task is novel, we borrowed some features from previous studies of text-level readability (Forsyth, 2014; Saddiki et al., 2015; Nassiri et al., 2018a,b). We decided to exclude the sentence length from the feature set, as this creates an artificial skew in understanding what is difficult: more difficult writing styles are often associated with longer sentences, but it is not the sentence length which makes them difficult. Specifically, many long Arabic sentences contain shorter ones, which are connected by conjunctions such as ' /wa /= and'. According to the experience of language teachers such sentences do not present problems for the learners.

3.1.1 The POS-features

[Table 3 features (1-21)], these features represent the distribution of different word categories in the sentence, and the morpho-syntactic features of these words. According to Knowles and Don (2004), Arabic lemmatization, unlike that of English, is an essential process for analysing Arabic text, because it is a methodology for dictionary construction. Therefore, we used the Lemma/Type ratio instead of Word/Type ratio. Adding features represents the different verb types (Verb pseudo, Passive verbs, Perfective verbs, Imperfective verbs and 3rdperson). As conjunction is one of the important features in representing sentence complexity in Arabic (Forsyth, 2014), we used the annotated discourse connectors introduced by Alsaif (2012) by splitting this list into 23 simple connectors and 56 complex connectors referring to non-discourse connectors and discourse connectors respectively. For POS-features extraction we used MADAMIRA a robust Arabic morphological analyser and part of speech tagger (Pasha et al., 2014).

3.1.2 Syntactic features

Features (22-27) from Table 3 provide some information about the sentences structures and number of phrases as well as phases types. These features are derived from a dependency grammar analysis. Because dependency grammar is based on word-word relations, it assumes that the structure of a sentence consists of lexical items that are attached to each other by binary asymmetrical relations, which is known as dependency relations. These relations will be more representative for this task. We used CamelParser (Shahrour et al., 2016) a system for Arabic syntactic dependency analysis together with contextually disambiguated morphological features which rely on the MADAMIRA morphological analysis for more robust results.

3.1.3 CEFR-level lexical features

Features (28-34) from Table 3 are used to assign each word in the sentence with an appropriate CEFR level. For this, we created a new Arabic word list consisting of 8834 unique lemmas labelled with CEFR levels. This list was a combination of three frequency