TECHLIMED@QALB-Shared Task 2015: a hybrid Arabic Error Correction System

Djamel Mostefa **Jaber Abualasal** **Omar Asbayou**

Mahmoud Gzawi **Ramzi Abbes**

Techlimed 42, rue de l'Université Lyon, France
{firstname.lastname}@techlimed.com

Abstract

This paper reports on the participation of Techlimed in the Second Shared Task on Automatic Arabic Error Correction organized by the Arabic Natural Language Processing Workshop. This year's competition includes two tracks, and, in addition to errors produced by native speakers (L1), also includes correction of texts written by learners of Arabic as a foreign language (L2). Techlimed participated in the L1 track. For our participation in the L1 evaluation task, we developed two systems. The first one is based on the spell-checker Hunspell with specific dictionaries. The second one is a hybrid system based on rules, morphology analysis and statistical machine translation. Our results on the test set show that the hybrid system outperforms the lexicon driven approach with a precision of 71.2%, a recall of 64.94% and an F-measure of 67.93%.

1 Introduction

Spell checking is an important task in Natural Language Processing (NLP). It can be used in a wide range of applications such as word processing tools, machine translation, information retrieval, optical character recognition etc. Automatic error correction tools on Arabic are underperforming in comparison with other languages like English or French. The lack of appropriate resources (e.g. publicly available corpora and tools) and the complexity of the Arabic language can explain this difference. Arabic is a challenging language for any NLP tool for many reasons. Arabic has a rich and complex morphology compared to other languages. Short vowels are missing in the texts but are mandatory from a grammatical point of view. Moreover, they are needed to disambiguate between several possibilities of words. Arabic is a rich language. It is characterised by its great number of synonyms and is a highly agglutinative, inflectional and derivational language that uses clitics (proclitics and enclitics). Arabic has many varieties. Modern Standard Arabic represents the variety of the news and formal speech. Classical Arabic refers to religious and classical texts. Dialectal Arabic has no standard rules for orthography and is based on the pronunciation. Therefore, a same word can be written using many different surface forms depending on the dialectal origin of the writer. Another very popular way of writing Arabic on the Internet and the social media like Facebook or Tweeter is to use "Arabizi", a Latinized form of writing Arabic using Latin letters and digits (Aboelezz 2009).

For our participation in this second QALB Shared Task, we tried to improve the systems we have developed for the first edition (Mostefa 2014). The first approach is a lexicon driven spell checker using Hunspell (Hunspell 2007). The second approach is a hybrid system based on correction rules, morphological analysis and statistical machine translation.

The paper is organized as follows: section 2 gives an overview of the automatic error correction evaluation task and resources provided by the organizers; section 3 describes the systems we have developed for the evaluations; and finally in section 4 we discuss the results and draw some conclusion.

Proceedings of the Second Workshop on Arabic Natural Language Processing, pages 161–165,
Beijing, China, July 26-31, 2015.

2 Task description and language resources

The QALB-2015 shared task (Rozovskaya 2015) is an extension of the first QALB shared task (Mohit 2014) that took place in 2014. QALB-2014 addressed errors in comments written to Aljazeera articles by native Arabic speakers (Zaghouani 2014).This year's competition includes two tracks, and, in addition to errors produced by native speakers, also includes correction of texts written by learners of Arabic as a foreign language (L2) (Zaghouani 2015). The native track includes Alj-train-2014, Alj-dev-2014, Alj-test-2014 texts from QALB-2014. The L2 track includes L2-train-2015 and L2-dev-2015. This data was released for the development of the systems. The systems were scored on blind test sets Alj-test-2015 and L2-test-2015.
The Alj-train-2014 set is made of 20,428 sentences for 1.1 M tokens.
The Alj-dev-2014 and Alj-test-2014 includes each around 1k sentences for 50k tokens
Finally, a test set Alj-test-2015 of 920 sentences for 48k tokens with no gold standard has to be corrected automatically for the evaluation campaign.
The evaluation metric is performed by comparing the gold standard with the hypothesis using the Levenshtein edit distance (Levenshtein 1966) and the implementation of the M2 scorer (Dahlmeier 2012). Then for each sentence Precision, Recall and F-measure are calculated.

3 System description

3.1 Rule-based system

For the rule-based system, we used the spell-checker Hunspell (Hunspell 2007) with different dictionaries and affix files.
The structure of Hunspell uses two files to define the spell checking of a language. The first file is a dictionary file that contains a stem list of the language. The second file is an affix file that maps the lemmas with their affixes. Affixes in Hunspell are divided into prefixes and suffixes, infixes are only included in the stems and spell checked in terms of proximity in lexical morphemes.
Dictionary and affix file in Hunspell is similar to the one depicted in Table 1 and Table 2

لدن/36	
نعفن/290	1246
تعفنان/273	1246

Table 1 Example of Hunspell dictionary

AF Tbcc # 36
PFX Tb 0 و .
SFX cc 0 ي .

Table 2 Example of Hunspell affix file

The dictionary contains the minimal words which are mapped with the affix rules.
The affix file contains mainly prefix and suffix rules that apply to the words of the dictionary. For instance, the rule of prefixation /Tb/ in Table 2 creates the word-form ولدن (wldn) while the rule of suffixation /cc/ creates ولدني (wldny).

For the evaluation, we used Hunspell with a modified version of the Hunspell Arabic dictionary and affix files version 3.2 (Ayaspell 2008).
We obtained a precision of 56.64% and a recall of 19.78% for an F-measure of 29.32% on the development set.
The results seem to be low but we have to consider that Hunspell does not correct the punctuation errors; many errors in the data include punctuation errors (around 30%).

3.2 Hybrid system based on SMT

For the second approach, we combined Statistical Machine Translation (SMT) system with morphological output of MADAMIRA (Pasha 2014) and some automatic rules to correct the text.
We build three different SMT systems based on the Moses toolkit (Koehn 2007) with different input for training the phrase-based translation models.
For the first system (Tech-1), we used the output of MADAMIRA morphological analyzer and the corrected texts to train a MADAMIRA/correct translation model. We used the text from the Alj-train-2014 data and apply corrections to build a parallel MADAMIRA/correct text corpus of 20,428 sentences and we train a phrase based translation model. The Alj-dev-2014 data is used for Moses to tune the translation models.
The second system (Tech-2) is the same as the previous one, but we added Alj-dev-2014 in the training data and used Alj-Test-2014 as development data for tuning the translation models.
The third system (Tech-3) uses the original erroneous text instead of the MADAMIRA output to build a parallel error/correct text corpus and we train a phrase based model. As for Tech-1, the Alj-dev-2014 data is used for Moses to tune the translation models.

For the word alignment, we used GIZA++ (Och 2003).

For the language model, we used corpora of newspapers publicly available and collected by Techlimed. The sources are coming from the Open Source Arabic Corpora (Saad 2010) (20M words), the Adjir corpus (Abdelali 2005) (147M words) and other corpora we collected from various online newspapers for a total of 300M words. The language model was created with the IRSTLM toolkit (Federico, 2008).

SMT System	TECH-1	TECH-2	TECH-3
MADAMIRA	Yes	Yes	No
Training data	Alj-train-2014	Alj-train-2014+Alj-dev-2014	Alj-train-2014
LM data	300 M	300 M	300 M
Rule correction	Yes	Yes	Yes

Table 3 System component description

For each system, we then applied the following rules:

- Convert eastern Arabic digits (٠١٢٣٤٥٦٧٨٩) into western Arabic digits (0 1 2 3 4 5 6 7 8 9).
- Separate numbers from word.
- Add a final stop at all sentence with no final punctuation.
- Remove duplicated punctuation marks, for instance ". !" ➔"!" or "!!!!"➔"!".

The results obtained on the development data (Alj-test-2014) and the evaluation set (Alj-test-2015) are given in the Table *4* and Table *5*.

System	P	R	F1
TECH-1	73.05	59.12	65.35
TECH-2	73.33	59.46	65.67
TECH-3	72.99	56.29	63.56

Table 4 Results on the development data (Alj-test-2014)

System	P	R	F1
TECH-1	71.08	64.74	67.76
TECH-2	71.20	64.94	67.93
TECH-3	69.99	60.41	64.85

Table 5 Results on the evaluation data (Alj-Test-2015)

The best system TECH-2 is obtained with the combination of MADAMIRA correction with the SMT system trained on 21k sentences and with correction rules. Table *6* describes the contribution of each component on the correction of TECH-2 on the evaluation data.

TECH-2	P	R	F1
MADAMIRA	80.33	39.98	53.39
+SMT	70.89	59.12	64.89
+Rule correction	71.20	64.94	67.93

Table 6 Performance of TECH-2 on the evaluation data (Alj-Test-2015) by component.

4 Error analysis and discussion

Some difficulties appear when we try to achieve and develop the automatic correction by spellchecker. These problems and difficulties are due not only to the complex morphological system of Arabic language, but also for many reasons, which concern the capacity of spellchecker system. The following list shows us types of problems and difficulties (the Buckwalter transliteration (Buckwalter 2002) is given for each Arabic word example).

Problem related to pronunciation similarities between the Hamza and Alif in some word such as إستقبل/إستعجال (<stEjAl/ <stqbl), which are respectively wrong versions of استقبل/استعجال. (AstEjAl/ Astqbl)

- Similar form problems leading to wrong word substitutions (i.e. incorrect substitution of words by one another): For example, words having similarities in form such as أن (>n) and إن (<n) are confused and ان (An), which does not exist in Arabic, is frequently used.
- Deverbal nouns ending ة/ـة: we notice that spellchecker does not respect Arabic forms of deverbal nouns, called Masdar in the Arabic grammatical tradition. As a result, it could not be able to correct words in which "ه/ـه" is wrongly used at the end of word position instead of ة/ـة (e.g. إبادة (<bAdp) having the deverbal form /?ifâlat/ إفالة (<fAlp) is written إباده/ابادة (AbAdh/ <bAdh).
- The morphosyntaxic information are not taken into consideration: the use of morphosyntaxic information makes our system capable of correcting nouns beginning with de morpheme "ال" (definite article) and ending by "ه/ـه" by substituting

this latter by "ة/ـة". These information allow us to apply rules such as المشكله (Alm$klh)➔المشكلة (Alm$klp).

- Plural nouns: broken plural (called also irregular plural) are not controlled by specific or respected rules in spellchecker system (e. g. both forms أفاعيل (>fAEyl) and أفعال (>fEAl) like اساطيل (AsATyl) and اطفال (ATfAl), wrong spelling of أساطيل (>sATyl) and أطفال (>TfAl), are not corrected by spellchecker system. The correct plural forms are أفاعيل (>fAEyl) and أفعال (>fEAl) instead of افعال (AfEAl) and افعال (AfEAl) where we do not respect the rule relative to the Hamza أ in the beginning of the broken plural form.
- Precision problems (homophony): a word in Arabic language may have different forms like سوريا (swryA) and سورية (swryp). But it has the same pronunciation. In such cases, both versions should be taken as correct.
- The spelling is influenced by dialectical language: e,g the use of انو (Anw) rather than أنه (>nh).
- The repetition of the same letter in a word: e.g الجزيييرة ;الجهاااد ; االمرسوم ; اللذي (Aljzyyyyrp; AljhAAAAd; AAlmrswm; Alll*y)
- The merging of two words: eg. ; اقتصادالبلد الثورةفسأحمل ; اقطعواالأمل (AqtSAdAlbld; Alvwrpfs>Hml; AqTEwAAl>ml).

5 Conclusion

This paper has reported on the participation of Techlimed in the 2015 QALB Shared Task on Automatic Arabic Error Correction. We developed two approaches, one based on Hunspell and the other based on a hybrid SMT system.

The best results were obtained with the hybrid SMT system which was able to deal with the punctuation mark corrections. We also tested a hybrid system by combining Hunspell and the SMT system but did not get better results than the SMT system. Our perspective is to include the DiiNAR lexical database (Abbès 2004) and also a large dialectal corpus to improve the results.

References

Abbès, Ramzi Dichy, Joseph and Mohamed Hassoun. "The architecture of a standard arabic lexical database: some figures, ratios and categories from the Diinar. 1 source program." *In Proceedings of the Workshop on Computational Approaches to Arabic Script-based Languages.* Association for Computational Linguistics, 2004. 15–22.

Abdelali, Ahmed. *http://aracorpus.e3rab.com/.* 2005. http://aracorpus.e3rab.com/ (accessed 2015).

Aboelezz, Mariam. "Latinised arabic and connections to bilingual ability." *In Papers from the Lancaster University Postgraduate Conference in Linguistics and Language Teaching.* 2009.

Ayaspell. *Ayaspell Arabic dictionary project.* 2008. http://ayaspell.sourceforge.net.

Buckwalter, Tim. *Buckwalter Arabic Morphological Analyzer Version 1.0.* 2002. http://catalog.ldc.upenn.edu/LDC2002L49 (accessed 06 2015).

Dahlmeier, Daniel and Ng, Hwee Tou. "Better evaluation for grammatical error correction." *In Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies.* Association for Computational Linguistics, 2012. 568–572.

Hunspell. *Hunspell.* 2007. http://hunspell.sourceforge.net/ (accessed 2015).

Koehn, Philipp Hoang, Hieu Birch, Alexandra Callison-Burch, Chris Federico, Marcello Bertoldi, Nicola Cowan, Brooke Shen, Wade Moran, Christine Zens, Richard Dyer, Christopher Bojar, Ondrej Constantin, Alexandra and Herbst. Evan. "Open source toolkit for statistical machine translation." *45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions.* Association for Computational Linguistics, 2007. 177-180.

Levenshtein, Vladimir. "Binary codes capable of correcting deletions, insertions and reversals." *In Soviet physics doklady.* 1966. volume 10, page 707.

Mohit, Behrang Rozovskaya, Alla Habash, Nizar Zaghouani, Wajd and Obeid, Ossama. "The First QALB Shared Task on Automatic Text Correction for Arabic." *Proceedings of EMNLP Workshop on Arabic Natural Language Processing.* Doha, Qatar, 2014.

Mostefa, Djamel Asbayou, Omar and Abbes, Ramzi. "TECHLIMED System Description for the Shared Task on Automatic Arabic Error Correction." *Proceedings of EMNLP Workshop on Arabic Natural Language Processing.* Doha, Qatar, 2014.

Och, Franz Joseph and Ney, Hermann. "A systematic comparison of various statistical alignment models." *Computational Linguistics.* 2003. 29(1):19–51.

Pasha, Arfath Al-Badrashiny, Mohamed El Kholy, Ahmed Eskander, Ramy Diab, Mona Habash, Nizar Pooleery, Manoj Rambow,

Owen and Roth, Ryan. "MADAMIRA: A Fast, Comprehensive Tool for Morphological Analysis and Disambiguation of Arabic." *Proceedings of LREC'2014.* Reykjavik,, 2014.

Rozovskaya, Alla Bouamor, Houda Habash, Nizar Zaghouani, Wajdi Obeid, Ossama and Mohit, Behrang. "The Second QALB Shared Task on Automatic Text Correction for Arabic." *Proceedings of ACL Workshop on Arabic Natural Language.* Beijing, China, 2015.

Saad, Ashour and Motaz, Wesam. "Osac: Open source arabic corpora." *In 6th ArchEng Int. Symposiums, EEECS.* 2010. volume 10.

Zaghouani, Wajdi Habash, Nizar Bouamor, Houda Rozovskaya, Alla Mohit, Behrang Heider, Abeer and Oflazer, Kemal. "Correction Annotation for Non-Native Arabic Texts: Guidelines and Corpus." *Proceedings of The 9th Linguistic Annotation Workshop.* Denver, Colorado, USA: Association for Computational Linguistics, 2015. 129-139.

Zaghouani, Wajdi Mohit, Behrang Habash, Nizar Obeid, Ossama Tomeh, Nadi Rozovskaya, Alla Farra, Noura Alkuhlani, Sarah and Oflazer, Kemal. "Large scale arabic error annotation: Guidelines and framework." *In Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14).* Reykjavik, Iceland: European Language Resources Association (ELRA), 2014.