Collaboratively Constructed Linguistic Resources for Language Variants and their Exploitation in NLP Applications – the case of Tunisian Arabic and the Social Media

Fatiha Sadat
University of Quebec in Montreal
201 President Kennedy, Montreal, QC, Canada
sadat.fatiha@uqam.ca

Fatma Mallek
University of Quebec in Montreal
201 President Kennedy, Montreal, QC, Canada
mallek.fatma@uqam.ca

Rahma Sellami
Sfax University, Sfax, Tunisia
rahma.sellami@fsegs.rnu.tn

Mohamed Mahdi Boudabous
Sfax University, Sfax, Tunisia
mehdiboudabous@gmail.com

Atefeh Farzindar
NLP Technologies Inc.
52 LeRoyer Street W, Montreal, Canada
farzindar@nlptechnologies.ca

Abstract

Modern Standard Arabic (MSA) is the formal language in most Arabic countries. Arabic Dialects (AD) or daily language differs from MSA especially in social media communication. However, most Arabic social media texts have mixed forms and many variations especially between MSA and AD. This paper aims to bridge the gap between MSA and AD by providing a framework for the translation of texts of social media. More precisely, this paper focuses on the Tunisian Dialect of Arabic (TAD) with an application on automatic machine translation for a social media text into MSA and any other target language. Linguistic tools such as a bilingual TAD-MSA lexicon and a set of grammatical mapping rules are collaboratively constructed and exploited in addition to a language model to produce MSA sentences of Tunisian dialectal sentences. This work is a first-step towards collaboratively constructed semantic and lexical resources for Arabic Social Media within the ASMAT (Arabic Social Media Analysis Tools) project.

1 Introduction

The explosive growth of social media has led to a wide range of new challenges for machine translation and language processing. The language used in social media occupies a new space between structured and unstructured media, formal and informal language, and dialect and standard usage. Yet these new platforms have given a digital voice to millions of user on the Internet, giving them the opportunity to communicate on the first truly global stage – the Internet (Colbath, 2012).

Social media poses three major computational challenges, dubbed by Gartner the 3Vs of big data: *volume, velocity, and variety*[1]. Natural Language Processing (NLP) methods, in particular, face further difficulties arising from the short, noisy, and strongly contextualised nature of social media. In order to address the 3Vs of social media, new language technologies have emerged, such as the identification and definition of users' language varieties and the translation to a different language, than the source.

[1] http://en.wikipedia.org/wiki/Big_data

Proceedings of the Workshop on Lexical and Grammatical Resources for Language Processing, pages 102–110,
Coling 2014, Dublin, Ireland, August 24 2014.

Furthermore, language in social media is very rich with linguistic innovations, morphology and lexical changes. People are not only socially connected across the world but also emotionally and linguistically (Sadat, 2013).

The importance of social media stems from the fact that the use of social networks has made everybody a potential author, which means that the language is now closer to the user than to any prescribed norms. Thus, considerable interest has recently been focused on the analysis of social media in order to create or enrich NLP tools and applications. There are, however, still many challenges to be faced depending on the used language and its variants.

This paper deal with Arabic language and its variants for the analysis of social media and the collaborative construction of linguistic tools, such as lexical dictionaries and grammars and their exploitation in NLP applications, such as translation technologies.

Basically, Arabic is considered as morphologically rich and complex language, which presents significant challenges for NLP and its applications. It is the official language in 22 countries spoken by more than 350 million people around the world[2]. Moreover, Arabic language exists in a state of diglossia where the standard form of the language, Modern Standard Arabic (MSA) and the regional dialects (AD) live side-by-side and are closely related (Elfardy and Diab, 2013). Arabic has more than 22 variants, refereed a as dialects; some countries share the same dialects, while many dialects may exist alongside MSA within the same Arab country. Arabic Dialects (AD) or daily language differs from MSA especially in social media communication. However, most Arabic social media texts have mixed forms and many variations especially between MSA and AD.

This paper describes our efforts to create linguistic resources and translation tool for TDA to MSA. First, a bilingual TDA-MSA lexicon and a set of TDA mapping rules for the social media context are collaboratively constructed. Second, these tools are exploited in addition to a language model extracted from MSA corpus, to produce MSA sentences of the Tunisian dialectal sentences of social media. The rule-based translation system can be coupled with a statistical machine translation system from MSA into any language, example French or English to provide a translation from TDA to French or English of original Tunisian dialectal sentences of social media.

This paper is organized as follows. In Section 2, we present some related works to this research. Section 3 discusses the Tunisian Dialect of Arabic (TDA) and its challenges in social media context. In Section 4, we present the collaboratively construct linguistic tools for the social media. Section 5 presents some evaluations of the combined TDA-MSA rule-based translation and disambiguation system. Section 6 concludes this paper and gives some future extensions.

2 Related Work

There have been several works on Arabic NLP. However, most traditional techniques have focused on MSA, since it is understood across a wide spectrum of audience in the Arab world and is widely used in the spoken and written media. Few works relate the processing of dialectal Arabic that is different from processing MSA. First, dialects leverage different subsets of MSA vocabulary, introduce different new vocabulary that are more based on the geographical location and culture, exhibit distinct grammatical rules, and adds new morphologies to the words. The gap between MSA and Arabic dialects has affected morphology, word order, and vocabulary (Kirchhoff and Vergyri, 2004). Almeman and Lee (2013) have shown in their work that only 10% of words (uni-gram) share between MSA and dialects. Second, one of the challenges for Arabic NLP applications is the mixture usage of both AD and MSA within the same text in social media context. Recently, research groups have started focusing on dialects. For instance, Columbia University provides a morphological analyzer (MAGEAD) for Levantine verbs and assumes the input is non-noisy and purely Levantine (Habash and Rambow, 2006b).

Given that DA and MSA do not have much in terms of parallel corpora, rule-based methods to translate DA-to-MSA or other methods to collect word-pair lists have been explored. Abo Bakr et al. (2008) introduced a hybrid approach to translate a sentence from Egyptian Arabic into MSA. This hybrid system consists of a statistical system for tokenizing and tagging, and a rule-based system for the construction of diacritized MSA sentences. Al-Sabbagh and Girju (2010) described an approach of

[2] http://en.wikipedia.org/wiki/Geographic_distribution_of_Arabic#Population

mining the web to build a DA-to-MSA lexicon. Salloum and Habash (2012) developed Elissa, a dialectal to standard Arabic tool that employs a rule-based translation approach and relies on morphological analysis, morphological transfer rules and dictionaries in addition to language models to produce MSA paraphrases of dialectal sentences.

Using closely related languages has been shown to improve MT quality when resources are limited. In the context of Arabic dialect translation, Sawaf (2010) built a hybrid MT system that uses both statistical and rule-based approaches for DA-to-English MT. In his approach, DA (but not TDA) is normalized into MSA by performing a combination of character- and morpheme-level mappings. They then translated the normalized source to English using a hybrid MT or alternatively a Statistical MT system.

Very few researches were reported on Tunisian variant of Arabic or any other Maghrebi variant. Hamdi et al. (2013) presented a translation system between MSA TDA verbal forms. Their approach relies on modeling the translation process over the deep morphological representations of roots and patterns, commonly used to model Semitic morphology. The reported results are aat 80% recall in the TDA into MSA and 84% recall in the opposite direction. However, the translation process was highly ambiguous, and a contextual disambiguation process was therefore necessary for such a process to be of practical use. Boudjelbene et al. (2013a, 2013b) described a method for building a bilingual dictionary using explicit knowledge about the relation between TDA and MSA and presented an automatic process for creating Tunisian Dialect corpora. However, their work focused on verbs mainly in order to adapt MAGEAD morphological analyser and generator of arabic dialect to TDA (Hamdi et al., 2013). Also, they developed a tool that generates TDA corpora and enrich semi-automatically the dictionaries they built. Experiments in progress showed that the integration of translated data improves lexical coverage and the perplexity of language models significantly. Their research was very pertinent for TDA but did not consider the mixture form of social media corpora.

Shaalan (2010) presented a rule-based approach for Arabic NLP and developed a transfer-based machine translation system of English noun phrase to Arabic. Their research showed that a rapid development of rule-based systems is feasible, especially in the absence of linguistic resources and the difficulties faced in adapting tools from other languages due to peculiarities an the nature of Arabic language.

In real-life practise, a company named Qordoba[3] launched social media translation service for Arabic in general. However, no demonstration or freely available version was found online. Furthermore, a new Twitter service automatically translates tweets from some Arabic language variants to English. However, this translation tool is not 100% accurate[4].

3 The Tunisian Dialect of Arabic and its Challenges in Social Media

Tunisian, or Tunisian Arabic[5] (TDA) is a Maghrebi dialect of the Arabic language, spoken by some 11 million people in coastal Tunisia. It is usually known by its own speakers as Derja, which means dialect, to distinguish it from Standard Arabic, or as *Tunsi*, which means Tunisian. In the interior of the country it merges, as part of a dialect continuum, into Algerian Arabic and Libyan Arabic.

The morphology, syntax, pronunciation and vocabulary of Tunisian Arabic are quite different from Standard or Classical Arabic. TDA, like other Maghrebi dialects, has a vocabulary mostly Arabic, with significant Berber substrates, and many words and loanwords borrowed from Berber, French, Turkish, Italian and Spanish. Derja is mutually spoken and understood in the Maghreb countries, especially Morocco, Algeria and Tunisia, but hard to understand for middle eastern Arabic speakers. It continues to evolve by integrating new French or English words, notably in technical fields, or by replacing old French and Spanish ones with Standard Arabic words within some circles. Moreover, Tunisian is also closely related to Maltese, which is not considered to be a dialect of Arabic for sociolinguistic reasons.

An exemple is the following sentences in Tunisian Dialect of Arabic (TDA) in social media, as presented in Figure 1. The underlined words (also in red) cannot be analyzable by MSA morphological analyzers, and thus need their own TDA analysis. Moreover, there are some words (in blue) expressed

[3] http://www.wamda.com/2013/06/qordoba-launches-new-social-media-translation-service
[4] http://www.neurope.eu/article/twitter-launches-arabic-translation-service
[5] http://en.wikipedia.org/wiki/Tunisian_Arabic

in languages other than Arabic, French in this case. At least three morphological tools are needed for this short text that is very common in social media.

It is often assumed that for the country in question, users of social media will use mostly if not always, the native language. However, this isn't always the case. Languages will be mixed up with up to three languages in a single "tweet" or blog. A Tunisian user of social media can involve the following languages or their variants: 1) native tongue (Arabic dialect), 2) MSA (example for greetings), 3) English and 4) the Colonial country's language, which is French in our case. In the case of Tunisian, French words/numbers may be used that sound like an Arabic word. An accurate machine translation for social media should manage this level of complexity, especially when ones add numerical characters and an ever-changing Lexicon of words.

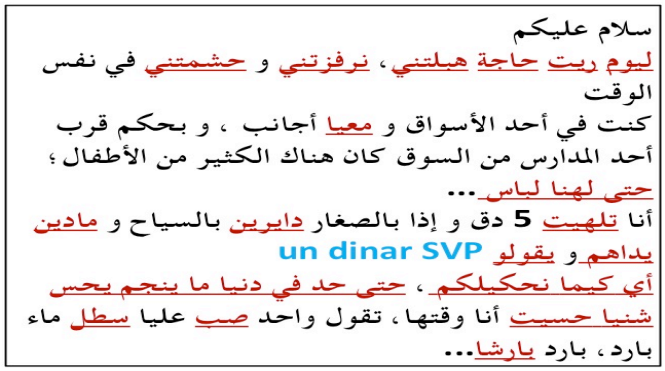
سلام عليكم
ليوم ريت حاجة هبلتني، نرفزتني و حشمتني في نفس الوقت
كنت في أحد الأسواق و معيا أجانب ، و بحكم قرب أحد المدارس من السوق كان هناك الكثير من الأطفال؛ حتى لهنا لباس...
أنا تلهيت 5 دق و إذا بالصغار دايرين بالسياح و مادين يداهم و يقولو un dinar SVP
أي كيما نحكيلكم، حتى حد في دنيا ما ينجم يحس شنيا حسيت أنا وقتها، تقول واحد صب عليا سطل ماء بارد، بارد يارشا...

Figure 1. Example of Social media text including a mixture of MSA, TDA and French language

4 Collaboratively Constructed Linguistic Tools for TDA

This section describes our effort in collaboratively constructing some linguitic tools that help translate Tunisian text of social media into MSA and other target languages considering MSA as a pivot language. Among these tools, a bilingual TDA-MSA lexicon and a set of mapping rules that will be integrated in the rule-based translation system (TDA-into-MSA). Furthermore, a language Modeling of MSA will help disambiguate the many translation hypothesis and thus select the best phrasal translation in MSA.

```
<?xml version="1.0" encoding="UTF-8"?>
<lexiques>
<lexique num_l="1" dt="اثنين اثنَيْن اثْنيْن زوز ثْنَيْنْ ثْنِيْنْ ثْنين" msa="إثْنان" />
<lexique num_l="2" dt="أكْحَلْ اكحل " msa="أسود " />
<lexique num_l="3" dt="لكْحَلْ لكحل " msa="الأسود " />
<lexique num_l="4" dt="كحْلْ كحل " msa="سود سودا" />
<lexique num_l="5" dt="لكْحُلْ لكحل " msa="السود السودا" />
<lexique num_l="6" dt="كحْلة كحلة " msa="سوداء سوداء" />
<lexique num_l="7" dt="أزْرَقْ أزرق فاروزي فارُوزي" msa="أزرق" />
<lexique num_l="8" dt="زرْقة زرقة فاروزية فارُوزيّة" msa="زرقاء" />
<lexique num_l="9" dt="كِيفَاهْ كيفاش كيفاه كيفاش" msa="كيف" />
<lexique num_l="10" dt="علاش علاهْ لواش لواهْ لاش لاهْ لِيْنْ لَيْنْ علاش علاه لواش لواه لاش لاش لين" msa=
"لماذا" />
<lexique num_l="11" dt="آش باش شنوَا اَش إشْنُوة وَاشِي شِنْهُو شِنْهي شنُو شنو شنوا اش شنوة اشنوة
آشي اشني واشي شنهو" msa="ماذا" />
<lexique num_l="12" dt="فينُو وينُو فينو فينوا وينو وينوا" msa="أين_هو" />
<lexique num_l="13" dt="فِينِكْ وينِكْ فِينَكْ وينَكْ فينك وينك" msa="أين_انت" />
<lexique num_l="14" dt="هي_وين هي_فين فين_هي وين_هي وينهي وينهيَ فيني وِيني فَيْني وَيْني فَيْنْ
وِيني فين" msa="أين_هي" />
<lexique num_l="15" dt="أنا_وين وين_أنا فين_أنا ويني_أنا فيني_أنا فَيْني وَيْني انا_فِيْنْ فيني
فين أنا_فين" msa="أين_أنا" />
```

Figure 2. An example from the TDA-MSA lexicon database

4.1 The TDA-MSA Bilingual Lexicon

We have manually and collaboratively developed a bilingual TDA-MSA lexicon that contains around 1,600 source words in TDA and its corresponding translations in MSA, defined by a human expert. Furthermore, our research on some downloadded extracts from Tunisian blogs (around 6,000 words), showed a difference between verb morphology in TDA and that in MSA. We find that in TDA, the gender distinction is not marked. Similarly, we noticed the absence of the masculine and feminine dual in TDA.

In this phase, our aim was to build a bilingual lexicon of Tunisian nouns and verbs and their translations into MSA. Note that a term can be a noun, a verb, an adverb, etc. Furthermore, the most used imported words from other language than Arabic (Berbere, French, English, Turkish, Spanish, Maltese) and used in social media context were considered in this lexicon. These TDA-MSA couples are stored in an XML database. Figure 2 shows a bilingual TDA-MSA extract from the lexicon database, encoded in XML.

4.2 Grammatical Mapping Rules for TDA

Our second collaboratively constructed linguistic tool, consists on a set of mapping rules that were checked by human experts. This set consists of some rules applied on verbs transformation in TDA and their corresponding translation into MSA. In final, we have defined a set of 226 mapping rules from TDA into MSA on verbs transformation. Figure 3 shows an extract of the defined rules, encoded in XML. Figure 4 shows an example of a verb in TDA and its translation ito MSA using rule number 171 of the collaboratively built set of mapping rules.

4.3 Automatic Rule-based TDA-MSA Machine Translation

We have developed a rule-based translation system that is able to translate any social media text in TDA into MSA. This rule-based translation system can be coupled with any statistical machine translation system from MSA to another language to provide a translation of original Tunisian dialectal sentences of social media from TDA to that other language.

Figure 5 shows the different steps used in the translation of any social media text from TDA into MSA. First, for each word in the TDA social media text, we proceed by searching in the TDA-MSA lexicon database for the corresponding translation of the TDA word. Mostly, TDA nouns and imported words from other languages than Arabic were included in the lexicon. Second, we proceed by searching in the database of mapping rules for the source verb in TDA and its corresponding MSA translation, as shown in Figure 4. Last, both word-by-word translation candidates are extracted from the lexicons and using the set of mapping rules; thus considered as translation hypothesis.

```
<?xml version="1.0" encoding="UTF-8"?>
<lexiques>
<lexique num_r="1" prefixe="ما" proclitique="ت" postfixe="ش" nprefixe="لا" nproclitique=
"ت" npostfixe="" />
<lexique num_r="2" prefixe="ما" proclitique="ت" postfixe="ش" nprefixe="لا" nproclitique=
"ت" npostfixe="ي"/>
<lexique num_r="3" prefixe="ما" proclitique="ت" postfixe="ش" nprefixe="لا" nproclitique=
"ت" npostfixe="ا و"/>
<lexique num_r="4" prefixe="ما" proclitique="ت" postfixe="وش" nprefixe="لا" nproclitique=
"ت" npostfixe="ا"/>
<lexique num_r="5" prefixe="ما" proclitique="ت" postfixe="وش" nprefixe="لا" nproclitique=
"ن" npostfixe="ن"/>
<lexique num_r="6" prefixe="" proclitique="ت" postfixe="ا و" nprefixe="" nproclitique="ت"
 npostfixe="ون"/>
<lexique num_r="7" prefixe="" proclitique="ت" postfixe="ا و" nprefixe="" nproclitique="ت"
 npostfixe="ا و"/>
<lexique num_r="8" prefixe="" proclitique="ت" postfixe="ا و" nprefixe="" nproclitique="ت"
 npostfixe="نا"/>
<lexique num_r="9" prefixe="" proclitique="ت" postfixe="ا و" nprefixe="" nproclitique="ت"
 npostfixe="ا"/>
<lexique num_r="10" prefixe="" proclitique="ت" postfixe="ا و" nprefixe="" nproclitique=
"ت" npostfixe="ن"/>
```

Figure 3. An example of some mapping rules from TDA to MSA

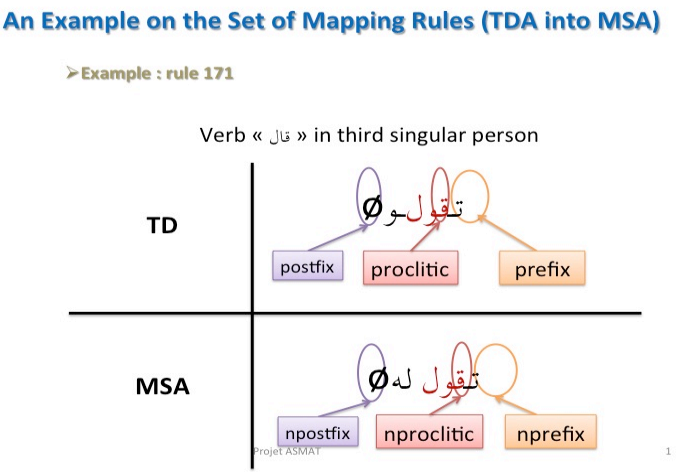

Figure 4. An example of the application of rule 171 for a verb in TDA and its translation into MSA

4.4 Language Modeling

The rule-based translation system is based on a word-by-word translation using the bilingual lexicon and the set of mapping rules. Thus, most of the time, one TDA sentence will have more than one possible translation. The language modeling (LM) of the target language (MSA) combined to the previous rule-based translation system will help disambiguate and select the best translation hypothesis in MSA.

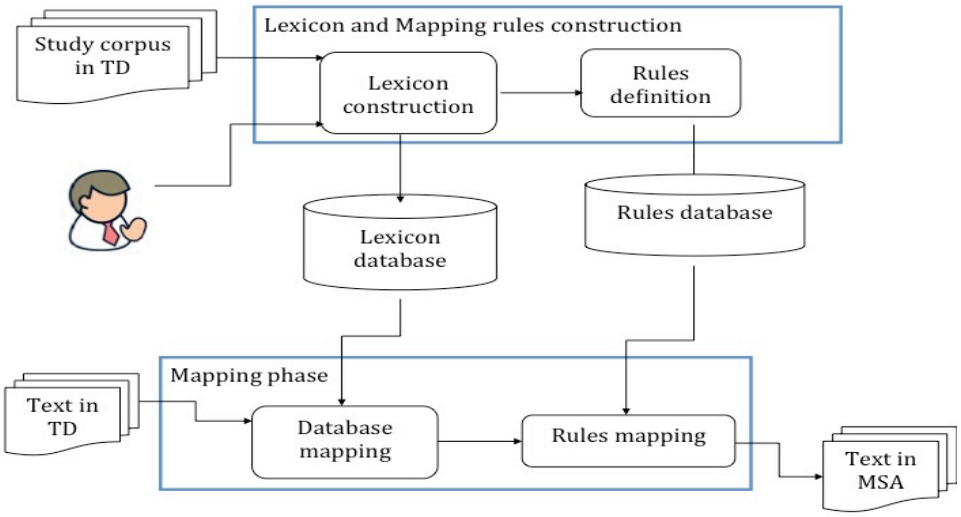

Figure 5. The rule-based translation approach for an automatic mapping from TDA to MSA

5 Evaluations

We have carried out some experiments and evaluations on the accuracy of the translation of TDA social media texts into MSA.

First, we collected manually a TDA corpus consisting of 6,000 words from some Tunisian forums and blogs. This corpus is very heterogeneous and multilingual, as many words are not in TDA but in MSA, French, English and sometimes using a certain style and form of social media, example using tweeter or SMS slangs). An extract of this corpus is presented in Figure 1.

For evaluation purposes, we considered a reference set of 50 phrases in TDA, translated manually into MSA. We also considered these 50 TDA phrases as the test set. Thus, we applied the proposed rule-based approach on this test set.

In order to combine adequately the rule-based translation approach to the language modeling (in MSA), we considered using the United Nation Arabic corpus to train a trigram language model. This training corpus contains around 50M words after cleaning the Latin content.

A preprocessing step is very crucial to any Arabic language processing. We considered tokenizing the MSA words using the D3 (Habash and Sadat, 2006a) scheme to overcome all problems of agglutination. The D3 scheme splits off clitics as follows: the class of conjunction clitics (w+ and f+), the class of particles (l+, k+, b+ and s+), the definite article (Al+) and all pronominal enclitics. These preprocessing are applied for both the hypothesis translation sentences and the training corpus, both in MSA. In addition to this preprocessing step, manual cleaning the MSA corpus of Latin contents was required. Thus, a trigram language model was implemented using the SRILM toolkit (Stolcke, 2002) on this training MSA corpus.

Next, we extracted all possible trigrams from the preprocessed MSA hypotheses translations and we computed the probability that these trigrams extracted appear in the MSA corpus based on the language model. A probability for each hypothesis translation is computed based on a trigram language model (LM). The hypothesis translation that has the highest probability is considered as the best translation.

Evaluations of the best translation sentence from TDA to MSA against the reference sentence in MSA were completed using the BLEU metric for automatic machine translation (Papineni et al., 2002). Our experiment produced a score of 14.32 BLEU. This low score could be related to our rule-based translation approach that is word-based and to the high number of unknown words in our source test file in other language variants than TDA. Adopting a phrasal translation and solving the problem of unknown words should be more effective.

Unfortunately, we could not found an available TDA-MSA test and reference files to conduct better evaluations in machine translation and social media context.

6 Conclusion and Future Work

Social media has become a key communication tool for people around the world. Building any NLP tool for texts extracted from social media is very challenging and daunting task and always be limited by the rapid changes in the social media. Considering an Arabic social media text is much more challenging because of the dominant use of English, French and other languages which intend to bring more problems to solve.

This paper presents our effort to create linguistic resources such as a bilingual lexicon, a set of grammatical mapping rule and a ruel-based translation and disambiguation system for the translation of any social media text from TDA into MSA. A language modeling of MSA is used in the disambiguation phase and the selection of the best translation phrase.

As for future work, we intend to enlarge the set of words in the TDA-MSA lexicon as well as the set of mapping rules. We intend to develop more grammatical rules for not only verbs but also adjectives and nouns. Furthermore, it would be interesting to build a parallel or comparable TDA-MSA corpus by selecting the most pertinent sources of social media and mining the web. A phrase-based statistical machine translation can be built using this parallel/comparable corpus and coupled to the rule-based translation system.

What we presented in this draft is a research on exploiting social media corpora for Arabic in order to analyze them and exploit them for NLP applications, such as machine translation within the scope of the ASMAT project.

Reference

Hitham Abo-Bakr, Khaled Shaalan, and Ibrahim Ziedan. 2008. A Hybrid Approach for Converting Written Egyptian Colloquial Dialect into Discretized Arabic. *Proceedings of the 6th International Conference on Informatics and Systems* 2008. Cairo University.

Rania Al-Sabbagh and Roxana Girju. 2010. Mining the Web for the Induction of a Dialectical Arabic Lexicon. *Proceedings of the 7th International Conference on Language Resources and Evaluation* LREC 2010. Valletta, Malta, May 19-21, 2010.

Khalid Almeman and Mark Lee. 2013. Automatic Building of Arabic Multi Dialect Text Corpora by Bootstrapping Dialect Words. *In Communications, Signal Processing, and their Applications* ICCSPA 2013. Sharjah, UAE, Feb.12-14, 2013.

Rahma Boujelbane, Mariem Ellouze Khemekhem, Siwar BenAyed, and Lamia Hadrich Belguith. 2013. Building Bilingual Lexicon to Create Dialect Tunisian Corpora and Adapt Language Model. *Proceedings of the 2nd Workshop on Hybrid Approaches to Translation,* ACL 2013. Sofia, Bulgaria.

Rahma Boujelbane, Mariem Ellouze Khemekhem, and Lamia Hadrich Belguith. 2013. Mapping Rules for Building a Tunisian Dialect Lexicon and Generating Corpora. *Proceedings of the International Joint Conference on Natural Language Processing*. Nagoya, Japan.

David Chiang, Mona Diab, Nizar Habash, Owen Rambow, and Safiullah Shareef. 2006. Parsing Arabic Dialects. *Proceedings of the European Chapter of ACL* EACL 2006.

Sean Colbath. 2012. Language and Translation Challenges in Social Media. *Proceedings of AMTA 2012, Government presentations*. Submitted by Raytheon BBN Technologies. Oct. 28th to Nov. 1st, 2012. San Diego, USA.

Mona Diab and Nizar Habash. 2007. Arabic Dialect Processing Tutorial**.** *Proceedings of HLT-NAACL, Tutorial Abstracts* 2007: 5-6.

Heba Elfardy and Mona Diab. 2013. Sentence-Level Dialect Identification in Arabic. *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics,* ACL 2013, Sofia, Bulgaria. 2013.

Nizar Habash and Fatiha Sadat. 2006. Arabic Pre-processing Schemes for Statistical Machine Translation. *Proceedings of the Human Language Technology Conference of the NAACL.* Companion volume: 49–52. New York City, USA.

Nizar Habash and Owen Rambow. 2006. MAGEAD: A Morphological Analyzer and Generator for the Arabic Dialects. *Proceedings of the 21st International Conference on Computational Linguistics and 44st Annual Meeting of the Association for Computational Linguistics*: 681–688, Sydney, Australia.

Ahmed Hamdi, Rahma Boujelbane, Nizar Habash, and Alexis Nasr. 2013. The Effects of Factorizing Root and Pattern Mapping in Bidirectional Tunisian - Standard Arabic Machine Translation. *Proceedings of MT Summit* 2013, Nice, France.

Ahmed Hamdi, Rahma Boujelbane, Nizar Habash, and Alexis Nasr. 2013. Un Système de Traduction de Verbes entre Arabe Standard et Arabe Dialectal par Analyse Morphologique Profonde. *Proceedings of TALN* 2013, Nantes, France.

Hanaa Kilany, Hassan. Gadalla, Howaida Arram, Ashraf Yacoub, Alaa El-Habashi, and Cynthia McLemore. 2002. Egyptian Colloquial Arabic Lexicon. *LDC catalog number LDC99L22*.

Katrin Kirchhoff, Jeff Bilmes, Sourin Das, Nicolae Duta, Melissa Egan, Gang Ji, Feng He, John Henderson, Daben Liu, Mohamed Noamany, Pat Schone, Richard Schwartz, and Dimitra Vergyri. 2003. Novel Approaches to Arabic Speech Recognition: Report from the 2002 Johns Hopkins Summer Workshop. *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing.* Hong Kong, China.

Katrin Kirchhoff and Dimitra Vergyri. 2004. Cross-dialectal Acoustic Data Sharing for Arabic Speech Recognition. *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, 2004.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei J. Zhu. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation. *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*: 311–318. Philadelphia, USA.

Fatiha Sadat. 2013. Arabic social media analysis for the construction and the enrichment of NLP tools. *In Corpus Linguistics* 2013. Lancaster University, UK. Jul. 22-26, 2013.

Hassan Sajjad, Kareem Darwish and Yonatan Belinkov. 2013. Translating Dialectal Arabic to English. *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*: 1–6. Sofia, Bulgaria, Aug. 4-9 2013.

Wael Salloum and Nizar Habash. 2011. Dialectal to Standard Arabic Paraphrasing to Improve Arabic-English Statistical Machine Translation. *Proceedings of the First Workshop on Algorithms and Resources for Modelling of Dialects and Language Varieties*. Edinburgh, Scotland.

Wael Salloum and Nizar Habash. 2012. Elissa: A Dialectal to Standard Machine Translation System. *Proceedings of Coling 2012*: Demonstration Papers: 385-392, Mumbai, India.

Hassan Sawaf. 2010. Arabic Dialect Handling in Hybrid Machine Translation. *Proceedings of the Conference of the Association for Machine Translation in the Americas* AMTA 2010. Denver, Colorado.

Khaled Shaalan. 2010. Rule-based Approach in Arabic Natural Language Processing. *International Journal on Information and Communication Technologies*, 3(3).

Andreas Stolcke. 2002. SRILM—An Extensible Language Modeling Toolkit. *Proceedings of ICSLP*, 2002.

Rabih Zbib, Erika Malchiodi, Jacob Devlin, David Stallard, Spyros Matsoukas, Richard Schwartz, John Makhoul, Omar F. Zaidan, and Chris Callison-Burch. 2012. Machine translation of Arabic Dialects. *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics*: Human Language Technologies, Montreal, Canada.