Computación y Sistemas
ISSN: 1405-5546
computacion-y-sistemas@cic.ipn.mx
Instituto Politécnico Nacional
México

Turki Khemakhem, Ines; Jamoussi, Salma; Ben Hamadou, Abdelmajid
POS Tagging without a Tagger: Using Aligned Corpora for Transferring Knowledge to Under - Resourced Languages
Computación y Sistemas, vol. 20, núm. 4, 2016, pp. 667-679
Instituto Politécnico Nacional
Distrito Federal, México

ISSN 2007-9737

POS Tagging without a Tagger: Using Aligned Corpora for Transferring Knowledge to Under-Resourced Languages

Ines Turki Khemakhem, Salma Jamoussi, Abdelmajid Ben Hamadou

University of Sfax, MIRACL Laboratory,
Tunisia

ines_turki@yahoo.fr, {salma.jammoussi, abdelmajid.benhamadou}@isimsf.rnu.tn

Abstract. Almost all languages lack sufficient resources and tools for developing Human Language Technologies (HLT). These technologies are mostly developed for languages for which large resources and tools are available. In this paper, we deal with the under-resourced languages, which can benefit from the available resources and tools to develop their own HLT. We consider as an example the POS tagging task, which is among the most primordial Natural Language Processing tasks. The task is importatn because it assigns to word tags that highlight their morphological features by considering the corresponding contexts. The solution that we propose in this research work, is based on the use of aligned parallel corpus as a bridge between a rich-resourced language and an under-resourced language. This kind of corpus is usually available. The rich-resourced language side of this corpus is annotated first. These POS-annotations are then exploited to predict the annotation on the under-resourced language side by using alignment training. After this training step, we obtain a matching table between the two languages, which is exploited to annotate an input text. The experimentation of the proposed approach is performed for a pair of languages: English as a rich-resourced language and Arabic as an under-resourced language. We used the IWSLT10 training corpus and English TreeTagger [15]. The approach was evaluated on the test corpus extracted from the IWSLT08 and obtained F-score of 89%. It can be extrapolated to the other NLP tasks.

Keywords. POS tagging, alignment, parallel corpus, under-resourced languages.

1 Introduction

There are more than 6,000 languages in the world but only a very small number of them are sufficiently equipped with the required linguistic resources and basic NLP tools for developing an appropriate Human Language Technology (HLT) [1]. The remaining languages are under-resourced and the majority of the tools and resources available are developed for specific academic purposes but not useful for large scale applications. This is due to the fact that the development of a HLT requires a lot of financial means and human expertise.

However, it is thought that it is always possible to build an appropriate HLT for the under-resourced languages, by exploiting the available resources and tools developed for the rich languages, as we suggest in the remainder of this paper. We are interested in the part of speech (POS) tagging which is a well known task in natural language processing (NLP). It is among the most important tasks faced by the majority of NLP systems.

For English and some European languages, POS tagging has achieved performances that approximate human levels. The accuracy of these POS taggers reaches 98% [12].

In this paper, we propose a new approach for POS tagging texts for Arabic, as an example of an under-resourced language, without developing a specific tool.

The idea that we defend is to use an aligned parallel corpus as a bridge between a rich resourced language (i.e. source language) and an under-resourced language (i.e. target language) in order to predict POS tags for an input text. In the proposed approach, we annotate the source side of the aligned parallel corpus, and then we generate the tags for the target side by using word alignments. In this work, we made use of the

memory-based learning approach (corpus-based method) to train our target language POS-tagger.

The remainder of this paper is as follows. In the next section, we review the related work for POS tagging. Section 3 gives a brief description of some related studies for the POS tagging process. We describe in section 4 our approach for predicting tags for under-resourced languages. Section 5 gives a short overview of the complexity of the Arabic morphology and the faced challenges of Arabic text POS tagging. In this section, we present the data and tools used to implement the proposed approach (the case of Arabic). We present the corpus as well as tag-sets used in the experiments. Section 6 gives the evaluation results, which are discussed in Section 7. Finally, section 8 concludes and suggests possible directions for future work.

2 Related Work

We focus here on related work for POS tagging. [16] present POS tagging experiments conducted to identify methods which result in good performance with small data set available. The result of their experiments for Amharic showed that a memory-based POS tagger-generator and tagger is a good tagging strategy for under-resourced languages as the accuracy of the tagger is less affected as the amount of training data increases compared with other methods. Memory-based tagging is based on the idea that words occurring in similar contexts will have the same tag. It is developed using Memory-Based Learning (MBL), a similarity-based supervised learning which is an adaptation and extension of the classical k-Nearest Neighbor (k-NN).

[3] suggest a solution to partially overcome the annotated resource shortage in the Vietnamese languages by building a POS-tagger for an automatically word-aligned English-Vietnamese parallel Corpus. This POS-tagger made use of the Transformation-Based Learning (or TBL) method by starting with a simple (or sophisticated) solution to the problem (called baseline tagging), and step-by-step applying optimal transformation rules (which are extracted from an annotated training corpus at each step) to solve (change from incorrect tags into correct ones) the problem. To build POS-tag annotated training data for Vietnamese, the authors need an annotated corpus with as high as possible accuracy.

In a previous work [18], a solution for disambiguation of the output of the Arabic morphological analyzer was presented. This method was used to help select the proper word tags for translation purposes via word-aligned bitext.

[17] used the similar automatic word alignments to show that additional token constraints can be projected from a resource rich source language to a resource-poor target language. The authors explored several models that combine token constraints with type constraints extracted from different source languages to achieve the POS tagging task.

3 POS Tagging Process

POS tagging has its importance in various areas of Natural Language Processing such as Text-to-Speech, information retrieval, parsing, information extraction and linguistic research [4, 9]. It is among the most difficult NLP tasks. It consists in assigning POS tags to words expressing their syntactic features in their corresponding contexts.

Several approaches have been proposed to construct automatic POS tagging, where the notable ones are rule-based, stochastic, and memory-based approaches.

The initial approaches used to address POS tagging are rule-based ones. It was based on two-stage architecture. The first stage used a dictionary to assign each word a list of potential POS. The second stage used large lists of hand-written desambiguisation rules to winnow down this list to a single part-of-speech for each word. The set of rules must be properly written and checked by human experts.

After the 1980, the stochastic (Statistical) approach came into existence and gained more popularity. It requires less work and cost than the rule-based approach. It uses a training corpus to pick up the most probable tag for a word. It is based on building a statistical language models and estimating parameters [17]. Some of these statistics parameters are lexical probability (the probability that a certain word appears with a

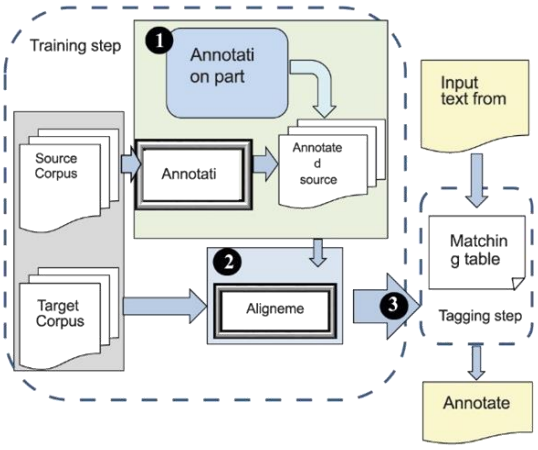

Fig. 1. The proposed POS-tagging system

certain tag) and contextual probability (the probability that a tag is followed by another). Statistical approaches mostly use Hidden Marcov Models (HMMs) for POS tagging.

Finally, in a memory-based approach, a set of cases is kept on memory. Each case consists of a word with its preceding and following context, and the corresponding category for that word in that context. A new sentence is tagged by selecting for each word in the sentence and its context, the most similar case in memory, and extrapolating the category of the word from these nearest neighbors. A memory-based approach has features of both learning rules based and stochastic taggers.

In this paper, we present a memory-based learning approach to POS tagging which combines the attractive properties of stochastic and rule-based approaches. The approach in this basic form is computationally expensive, however each new word in context that has to be tagged, has to be compared to each pattern kept in memory.

The proposed approach does not use the classic scheme that we have just described, which requires complex computations and higher costs. It is significantly simpler and faster to implement.

In our study, we present a method to benefit from rich language POS taggers in order to annotate under-resourced languages. It is based on the use of aligned parallel corpus. We exploit a POS tagger to annotate a rich language side. These POS-annotations were then exploited to predict tags of an under-resourced language via the word alignment.

4 Proposed Approach for Tagging without POS Tagger

4.1 Required Resource and Tool

The proposed approach needs the following resources:

– A parallel corpus composed of sentences in the source language (i.e. the rich language) aligned with sentences in the target language (i.e. under-resourced language).
– POS tagger for the rich language to annotate the source side of the parallel corpus.
– Alignment tool: the alignment indicates the mapping from source sentence words to target sentence words.

4.2 Main Step

The tagging process is performed in three main steps:

– The words on source side of the parallel corpus are combined with their respective POS and the words on target side are kept unchanged.
– This obtained bilingual corpus is automatically a word aligned by the alignment toolkit. After this alignment step, we obtain one model table containing annotated target word aligned with source word with an alignment probability.
– The obtained table is sorted out and the probability that corresponds to the same target word and the same POS of source word is added. Then, the resulting probabilities are sorted out, and the POS which corresponds to the maximum probability is selected.

Finally, a matching table is got, where each line from this table refers to the corresponding target word in the corpus and its POS projected from the source word. This table is used to annotate an input text.

Our POS-tagging system can be briefly described in figure 1 as follows:

Our approach of POS-tagging is effective and it reaches a competitive accuracy compared to other powerful POS-taggers as we will see in section 5. The algorithm for our POS-tagger is described as follows:

Algorithm for training step:

T_s={ws_1,ws_2,ws_3,….ws_n}
T_t={wt_1,wt_2,wt_3,….wt_m}
For i=1 to n do
ws_i = concat(ws_i,$tags_i$)
alignment_training(T_s,T_t)

/* Merge the probability P_{ik}=P(wt_j/wsk) for the same target word wt_j , and select $tags_i$ (POS of s_i) matching with the highest probability p_{ik} */

For i=1 to m do
begin
For k=1 to nb_tag do
P_{ik}=add(p_{ik}(wt_i/ws_k))
For k=1 to nb_tag do
$tags_i$= maximum(p_{ik})
Return $tags_i$
End

Algorithm for tagging input text:

Let w_1 w_2 w_3….w_d input sentence
For i=1 to d do
Begin
For j=1 to m do
If (w_i==wt_j) **then** Return ($tags_j$)
End

5 Application to the Arabic Language

5.1 Arabic as an Under-Resourced Language

Despite being a widely spoken language, Arabic is considered as a typical example of an under-resourced language. Indeed, it has few publicly available tools and resources, apart from a few notable exceptions, such as the Arabic Penn Treebank which is very expensive (Penn Tree Bank was invested over 1 million dollars) [11, 13]. In particular, Arabic NLP lacks resources such as annotated corpora, lexicons, machine-readable dictionaries in addition to fully automated fundamental NLP tools such as tokenizers, POS taggers and parsers [7]. The small number of available tools and corpora are mainly developed for specific academic purposes. Some reasons for this lack of resources may be the complex morphology, the absence of diacritics (vowels) in written text and the fact that Arabic does not use capitalization as we will discuss in detail in the following sub-section [6].

5.2 Challenges for Arabic Languages

Arabic is considered as one of the oldest languages in the world. It is ranked the fifth language among the widely used languages these days. Arabic words are written as a series of letters, in which the letters of a single word strung together to form it.

Arabic differs from other languages syntactically, morphologically, and semantically which makes it one of the most difficult languages for written and spoken language processing [16].

The Arabic language presents many challenges: it is both morphologically rich and highly ambiguous with a rich set of suffixes. Inflectional and derivational productions introduce a big number of possible word forms. In Arabic, articles, prepositions, pronouns, etc. can be affixed to adjectives, nouns, verbs and particles to which they are related.

Arabic grammarians categorized Arabic words into four main part-of-speech classes. These classes are: nouns, proper names, verbs and particles [16].

In Arabic language, there are two genders: masculine and feminine. In western languages words are singular or plural, but in Arabic language, the words could be singular, dual or plural. The dual represents a total of two nouns, pronouns, verbs or adjectives.

The plural form in western languages is obtained by adding the letter ”s” to the end of the word, whereas in Arabic, the plural form is of two types: regular and broken.

It seems that there are some aspects about the nature of Arabic that have slowed down the progress in Arabic NLP compared to the accomplishments in English and in other European languages. These include the following:

– The absence of capitalization in Arabic makes it hard to identify proper names.

– Arabic is highly inflectional and derivational, which makes morphological analysis a very complex task.
– Diacritics (vowels) are, most of the time, omitted from the Arabic text, which makes it hard to infer the word's meaning and therefore, it becomes difficult to POS tagging the text.

Unlike languages such as English and French, the Arabic morphological analysis is a particularly difficult step due to large graphics ambiguities of the Arabic word. An Arabic word can sometimes correspond to a whole English sentence (Example: the Arabic word "أتتذكّروننا" corresponds in English to: "Do you remember us").

The aim of a morphological analysis step is to recognize word composition and to provide specific morphological information about it. For Example: the word "يعرفون" (in English: they know) is the result of the concatenation of the prefix "ي" indicating the present and suffix "ون" indicating the plural masculine of the verb "عرف" (in English: to know). The morphological analyzer must determine for each word the list of all its possible morphological features.

As is the case with many other NLP applications, most of the activities are concerned with the English language. This is due to the fact that language resources are readily available.

Different techniques of POS tagging models have been implemented and performed for English language. On the contrary, only a small amount of work has been done for Arabic language. The structure of Arabic language is different from the English one, so it is not possible to apply available methods directly for Arabic.

Despite the differences between English and Arabic morphology, we show how it is possible to exploit the available English tools and resources to create POS tagger for Arabic.

5.3 Used Resources

To implement the proposed approach for tagging Arabic texts, we need the following resources and tools:

– A bilingual (English-Arabic) sentence aligned parallel corpus.
– Morphological analyzer for English words.
– Alignment tool: GIZA.

5.3.1 Aligned Parallel Corpus

The aligned parallel corpus used is composed of 177020 sentences. It consists of 73793 English words and 182088 Arabic words (121900 segmented Arabic words). To build the aligned parallel corpus, we have combined several corpora:

– The train part of the IWSLT10: a training corpus of 19972 sentence pairs;
– The dev6 dataset, made up of 489 sentences, which corresponds to the IWSLT07 development data;
– The dev1 dataset, made up of 506 sentences, which corresponds to the CSTAR03 development data;
– The dev2 dataset, made up of 500 sentences, which corresponds to the IWSLT04 development data;
– The dev3 dataset, made up of 506 sentences, which corresponds to the IWSLT05 development data;
– The train part of the IWSLT14: a training corpus of 155 047 sentence pairs.

The statistical information related to the used parallel corpora is detailed in table 1.

5.3.2 Treetagger Tool

We use treetagger, a supervised part-of-speech tagger, to determine POS tags for the English side of the parallel corpus and then the same corpus is used to make the predictions on the Arabic side. The treeTagger was developed at the University of Stuttgart by Helmut Schmid [15]. It supports multiple languages (Bulgarian, Dutch, English, French, German, Italian, Spanish, Russian) and can achieve good POS-tagging performances.

Unlike other probabilistic tagging tools, which have difficulty in estimating small precise probabilities of limited amounts of training data, TreeTagger prevents the sparse data problem by using a binary decision tree. It determines the appropriate size of the context used to estimate transition probabilities. TreeTagger is very fast and has great support for misspelled words as well as words non-existing in the lexicon [14]. The tag set proved by treetagger is given in Table 2.

Table 1. Parallel corpus description

Resources	Number of sentences	Number of English words	Number of Arabic words	Number of segmented Arabic words
IWSLT10	19972	7296	18149	14001
Dev1	506	184	459	301
Dev2	500	182	454	255
Dev3	506	179	450	320
Dev6	489	178	428	290
IWSLT14	155047	65774	162148	106733
Total	177020	73793	182088	121900

Table 2. The treetagger tag set

Tag	***Designation***	**Tag**	**Designation**	**Tag**	**Designation**
JJ	*Adjective*	NP/ NPS	proper noun singular / proper noun, plural	PP	personal pronoun
RB	Adverb	NN/ NNS	noun singular / noun plural	PP$	possessive pronoun
CC	Coordinating conjunction	VBP	imperfect verb	IN	Preposition/subord. conj.
DT	Determiner	VBN	passive verb	UH	Interjection
FW	foreign word	VBD	perfect verb	WP	wh-pronoun
CD	cardinal number	RP	particle	WRB	wh-abverb
SENT	End punctuation				

5.3.3 Alignment Tool: GIZA++

GIZA++ [14] is part of the statistical machine translation toolkit used to train IBM Model 1 to Model 5 (Brown et al., 1993) and the Hidden Markov Model (HMM) [14].

With the help of Expectation-Maximization (EM) algorithm, final word alignment results can be obtained after GIZA++ trains the parallel corpus several iterations from two directions (source to target language and vice-versa). Various heuristics, such as grow-diag-final [10] can be applied to obtain a better symmetrical alignment from those two directions.

5.4 Tagging Process

Arabic POS-tagging is implemented by two morphological processing steps. We first apply word segmentation to Arabic text. Then, we use the treetagger tool to extract POS-tags of the English words in the aligned parallel corpus.

EN: the light was red
AR: الاشارة كانت حمراء
⇩
EN: the_DT light_NN was_VBD red_JJ
AR: ال اشارة كانت حمراء

An example of partially-annotated parallel corpora is given below.

The English-Arabic alignment was trained with GIZA++ toolkit. It is illustrated in Figure 4, where

each Arabic morpheme is aligned to one or zero annotated English word. Next, English POS tagging result is used to identify Arabic side tags via word-alignments in order to form a new Arabic training corpus annotated with POS-tags. Word alignment results can be used to build phrase table. Table 3 shows an example of extracting phrases (refer to consecutive sequence of words) according to the English-Arabic word alignment in Figure 2.

After training, we end up with a model file containing aligned Arabic words followed by annotated English words with alignment probabilities and we sort the obtained file. Finally, we apply mapping rules to extract Arabic POS-tags of the Arabic side in the aligned parallel corpora as mentioned in section 5. Figure 3 presents an example of the resulting file and figure 4, 5 and 6 present the steps we have applied.

Arabic and English are two different languages. This fact will exhibit tagset mismatches, due to morphological difference between English and Arabic. In addition, the tagsets of these two languages are different. Due to characteristics of each language, we must use two different tagsets for POS-tagging.

In this first experimentation, we apply a complex 1-n matching (see table 4).

For English, the set of possible POS provided by TreeTagger is the verb, proper name, noun, adjective, adverb, conjunction, pronoun, preposition, etc. For Arabic, we made use of four categories of words: noun, verb, proper name, conjunction. POS tagging of Arabic words aligned with English words tagged by adjective or noun will be replaced by the morpho-syntactic feature: noun: "اسم". While, the POS tagging: adverb, conjunction, or prepostion will be replaced by the Arabic morpho-syntactic feature: particle: "أداة".

The mapping table of Arabic-English POS-tags is presented in Table 4.

The result of POS tagging of the Arabic sentence "إنها أمطرت بالأمس و لكن الجو جميل اليوم" is in table 5.

In the second experiment, we made use of the Penn TreeBank tagset for Arabic: noun, verb adverb, pronoun, conjunction, adjective, proper name, interjection, cardinal number.

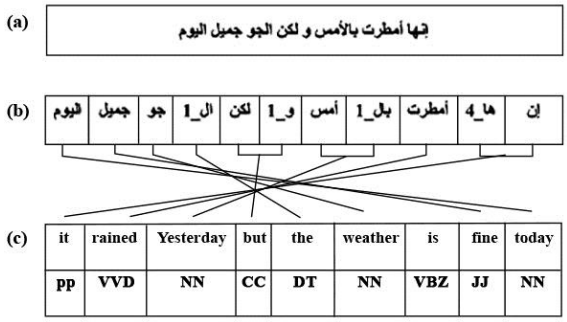

Fig. 2. An example of word alignments, (a) Original Arabic sentence, (b) segmented Arabic sentence, (c) English translation and its alignment with a morphological analysis

Table 3. Content of phrase table

إن ها_4 ||| it_PP
إن ها_4 أمطرت ||| it_PP rained_VVD
إن ها_4 أمطرت بال_1 أمس ||| it_PP rained_VVD yesterday_NN
إن ها_4 أمطرت بال_1 أمس و_1 لكن ||| it_PP rained_VVD yesterday_NN but_CC
أمطرت ||| rained_VVD
أمطرت بال_1 أمس ||| rained_VVD yesterday_NN
أمطرت بال_1 أمس و_1 لكن ||| rained_VVD yesterday_NN but_CC
أمطرت بال_1 أمس و_1 لكن ال_1 ||| rained_VVD yesterday_NN but_CC the_DT
أمطرت بال_1 أمس و_1 لكن ال_1 جو ||| rained_VVD yesterday_NN but_CC the_DT weathe_NN
بال_1 أمس ||| yesterday_NN
بال_1 أمس و_1 لكن ||| yesterday_NN but_CC
بال_1 أمس و_1 لكن ال_1 ||| yesterday_NN but_CC the_DT
بال_1 أمس و_1 لكن ال_1 جو ||| yesterday_NN but_CC the_DT weathe_NN
بال_1 أمس و_1 لكن ال_1 جو جميل ||| yesterday_NN but_CC the_DT weathe_NN is_VBZ fine_JJ
و_1 لكن ال_1 ||| but_CC the_DT
و_1 لكن ال_1 ||| but_CC
و_1 لكن ال_1 جو ||| but_CC the_DT weathe_NN
و_1 لكن ال_1 جو جميل ||| but_CC the_DT weathe_NN is_VBZ fine_JJ
و_1 لكن ال_1 جو جميل اليوم ||| but_CC the_DT weathe_NN is_VBZ fine_JJ today_NN
ال_1 جو جميل ||| the_DT weathe_NN is_VBZ fine_JJ
ال_1 جو جميل اليوم ||| the_DT weathe_NN is_VBZ fine_JJ today_NN
جميل اليوم ||| is_VBZ fine_JJ today_NN

Table 4. Mapping table of English-Arabic POS-tagset

Arabic POS	English POS
Noun	JJ, CD, NN, NNS
Verb	VBN / VBP / VBD
Proper name	NP, NPS
Conjunction	RB, CC, DT, RP, PP, PP$, IN

Beautiful_JJ	0,4545455	جميل
Buddy_NN	0,3333333	جميل
Mmm_NP	0,3333333	جميل
Considerate_JJ	0,2500000	جميل
pretty_JJ	0,2500000	جميل
we're_VV	0,2500000	جميل
last_VV	0,2000000	جميل
fine_RB	0,1818182	جميل
hilltop_NN	0,1666667	جميل
nice_JJ	0,1283784	جميل
pretty_RB	0,1052632	جميل
fine_NN	0,1000000	جميل
fancy_JJ	0,0909091	جميل
lovely_JJ	0,0714286	جميل
fine_JJ	0,0576923	جميل
Italian_JJ	0,0500000	جميل
Weather_NN	0,0285714	جميل
Food_NN	0,0067114	جميل
that's_NNS	0,0056180	جميل
been_VBN	0,0051282	جميل
good_JJ	0,0030211	جميل
is_VBZ	0,0015803	جميل
a_DT	0,0009750	جميل

Fig. 3. Example of the resulting table where probabilities are sorted

Beautiful_nom	0,4545455	جميل
Buddy_nom	0,3333333	جميل
Mmm_nompropre	0,3333333	جميل
Considerate_nom	0,2500000	جميل
pretty_nom	0,2500000	جميل
we're_verbe	0,2500000	جميل
last_verbe	0,2000000	جميل
fine_conjonction	0,1818182	جميل
hilltop_nom	0,1666667	جميل
nice_nom	0,1283784	جميل
pretty_conjonction	0,1052632	جميل
fine_nom	0,1000000	جميل
fancy_nom	0,0909091	جميل
lovely_nom	0,0714286	جميل
fine_nom	0,0576923	جميل
Italian_nom	0,0500000	جميل
Weather_nom	0,0285714	جميل
Food_nom	0,0067114	جميل
that's_nom	0,0056180	جميل
been_verbe	0,0051282	جميل
good_nom	0,0030211	جميل
is_verbe	0,0015803	جميل
a_conjonction	0,0009750	جميل

Fig. 4. Example of application of mapping rules

fine_conjonction	0,1818182	جميل
pretty_conjonction	0,1052632	جميل
a_conjonction	0,0009750	جميل
Beautiful_nom	0,4545455	جميل
Buddy_nom	0,3333333	جميل
Considerate_nom	0,2500000	جميل
pretty_nom	0,2500000	جميل
hilltop_nom	0,1666667	جميل
nice_nom	0,1283784	جميل
fine_nom	0,1000000	جميل
fancy_nom	0,0909091	جميل
lovely_nom	0,0714286	جميل
fine_nom	0,0576923	جميل
Italian_nom	0,0500000	جميل
Weather_nom	0,0285714	جميل
Food_nom	0,0067114	جميل
that's_nom	0,0056180	جميل
good_nom	0,0030211	جميل
Mmm_nompropre	0,3333333	جميل
we're_verbe	0,2500000	جميل
last_verbe	0,2000000	جميل
been_verbe	0,0051282	جميل
is_verbe	0,0015803	جميل

Fig. 5. Probabilities corresponding to the same POS are added

Conjonction	0,2880564	جميل
Nom	1,9968758	جميل
Nompropre	0,3333333	جميل
verbe	0,4567085	جميل

Fig. 6. POS which corresponds to the maximum probability is selected

Table 5. An example of POS tagging of Arabic words via a word-aligned pair of sentences in English-Arabic corpora

English word	It	Rained	Yesterday	But	The	weather	is	fine	today
English tag	PP	VVD	NN	CC	DT	NN	VBZ	JJ	NN
Arabic word	إن ها_4	أمطرت	بال_1 أمس	و_1 لكن		ال_1 جو		جميل	اليوم
Arabic tag	conj	verb	noun	conj	conj	noun	verb	noun	noun

Table 6. Mapping table of English-Arabic POS-tagset

Arabic POS	Noun	Verb	proper name	Adjective	Adverb	Pronoun	Conj-unction	Interj-ection	Cardinal number
English POS	NN NNS	VBN VBP VBD	NP NPS	JJ	RB, RP,IN	WP, PP, PP$, DT	CC	UH	CD

Table 7. An example of POS tagging of Arabic words via word-aligned pairs of sentences in English-Arabic corpora

English word	It	rained	yesterday	but	the	weather	is	fine	today
English tag	PP	VVD	NN	CC	DT	NN	VBZ	JJ	NN
Arabic word	إن ها_4	أمطرت	بال_1 أمس	و_1 لكن		ال_1 جو		جميل	اليوم
Arabic tag	Pronouns	Verb	Nouns	Conju-nctions	Pronouns	Nouns	Verb	Adje-ctives	Nouns

Table 8. Testing datasets

Test dataset	Number of sentences	Number of Arabic words	Number of segmented Arabic words
dev7	507	1105	800

Table 9. Results of POS tagging of Arabic words from aligned parallel corpora

Correct tags	Incorrect tags	Unknow word	Precision	Recall	F-Measure
671	108	21	84 %	96 %	89 %

For English, the set of possible POS provided by TreeTagger is used. The mapping table of Arabic-English POS-tags can be seen in table 6.

The result of POS tagging of the Arabic sentence "إنها أمطرت بالأمس و لكن الجو جميل اليوم" is as table 7.

6 Experimental Results

In our experimentation, we used the dev7 dataset, made up of 507 sentences, which corresponds to the IWSLT08 development data. It contains 1105 Arabic words (800 segmented Arabic words) as seen in table 8. Our systems were tested on this test corpus.

In the first experiment, the result of Arabic POS-tagging, using four categories of words, is given as in table 9.

In order to evaluate our approach, we need a part-of-speech tagger to compare the results. We use MADA [8], a supervised part-of-speech tagger, to determine POS tags for Arabic. MADA is a system for Morphological Analysis and Disambiguation for Arabic. MADA produces for each input word, a list of analyses specifying every possible morphological interpretation of that word, covering all the morphological features of the word (diacritization, POS, lemma, and 13 inflectional and clitic features). MADA then applies a set of models, Support Vector Machines (SVMs) and N-gram language models, to produce a prediction, per word in context, for different morphological features, such as POS, lemma, gender, number or person. A ranking component scores the analyses produced by the morphological analyzer using a tuned weighted sum of matches with the predicted features. The top-scoring analysis is chosen as the predicted interpretation for that word in context. This analysis can then be used to deduce a POS for the word. Table 10 compares the performance of our approach with the results obtained by MADA.

Our system achieves an accuracy of 84% compared to 91% obtained by a supervised part-of-speech tagger, MADA. It is quite natural to find an error reduction of 7% compared to MADA, since this latter uses a large annotated corpora to predict tags compared with our system that uses a very reduced one.

In the second experiment, which use the Penn TreeBank tagset for Arabic, the annotated Arabic side in the aligned parallel corpora is used to make the predictions on the test datasets. The result of Arabic POS-tagging is given in table 11.

In this refined statistical, our system obtains an accuracy of 75 % compared to 76 % obtained by MADA. We observed a degradation in F-measure, in the order of 5% for our system and 9% for MADA. This degradation is due to morphological difference between English and Arabic.

7 Discussion of Results

The discussion of the results is presented in two sections. First presents the discussion related to the first experiment, where we use four categories of words for Arabic POS-tagging. The second is the discussion devoted to the second experiment, which use the Penn TreeBank tagset for Arabic POS-tagging.

As shown in the evaluation section, the experimental results for our first experiment show that our system gives encouraging results, which is 84% for the precisión and 89% for the F-measure. The obtained results are promising and motivate the process of using the alignement to project the obtained POS from the English side to the Arabic side. This process allows exploiting the tools and resources available for the English language (i.e., well-resourced language), for the Arabic language processing, which lacks resources. In fact, English POS-tagging is provided by the TreeTagger annotation tool. The use of this latter results in a precision of 96.7%.

To assess the impact of the tagset on the tagging accuracy, a similar experiment was carried out using the same data but using an extended set of tags. The obtained results in terms of precision and F-measure are respectively 75% and 84 %. While the tagging accuracy decreases of 9%, the

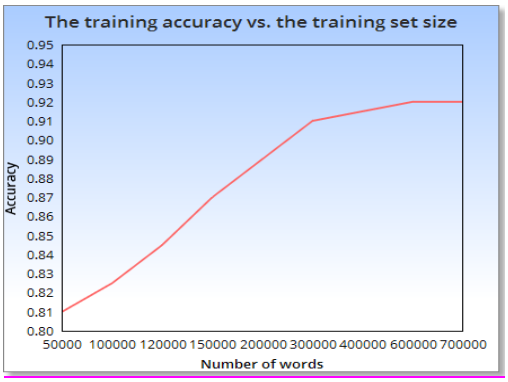

Fig. 7. The training accuracy according to the training set size

Table 10. Tagging accuracies of Arabic words

	Correct tags	Incorrect tags	Unknown word	Precision	Recall	F-measure
POS taggs of Arabic using aligned parallel corpora (four categories of Arabic words)	671	108	21	84%	96%	89%
MADA	726	74	0	91%	1	95%

Table 11. Results of POS tagging of Arabic words from an aligned parallel corpora

	Correct tags	Incorrect tags	Unknown word	Precision	Recall	F-Measure
POS taggs of arabic using aligned parallel corpora (Penn TreeBank tagset)	588	191	21	75 %	96 %	84 %
MADA	606	194	0	76 %	1	86 %

enlargement of the tagset allowed us to obtain a precision that is close to the precision of the supervised POS tagger MADA, which is 76%.

While our system uses very reduced annotated corpora compared to MADA, experimental results show significant performance of our system. This performance is due to the quality of alignment used and the performance of the TreeTagger annotation tool.

We argue that using resources from resource-rich languages to identify part-of-speech of under-resourced languages helps achieve a significant result. Our approach provides a fast method to extract the morphological features, which is not expensive to create and can save valuable time for informants to provide POS.

Our method enables a faster deployment of NLP systems than recruiting informants to generate a labeled training set, especially since, for many under-resourced languages reliable informants may not be readily available.

Despite the POS-tagging satisfaction, the results can not be fully satisfactory due to the following reasons:

– The result of automatic word-alignment is only 87% [5].
– The size of Arabic-English aligned parallel corpora is reduced.

Through the statistical figure 7 below, the change of accuracy as a function of the Arabic-English training set size can be seen as follows:

From figure 7, it is found that POS tagging accuracy increases with the increase of training data.

8 Conclusions

We have proposed a new approach for POS tagging without developing a specific tool. Our approach is based on the use of aligned parallel corpus as a bridge between a rich resourced language and an under-resourced language. We exploited a supervised POS taggers to annotate a rich-language side. These POS-annotations were then exploited to predict tags of under-resourced language via the word alignments.

We have experimented this method on Arabic-English aligned corpora. The Arabic POS-tagging is done in two steps: We first applied word segmentation to Arabic. We then used the treetagger tool to extract POS-tagging of the English words in the aligned parallel corpora.

The basic step consists in matching the data in the segmented Arabic corpora to the words concatenated with POS tags provided by

TreeTagger. The matching included the results of the alignment of words of the segmented Arabic corpus and the POS tagged English corpus.

Our method is evaluated in terms of precision, recall, and F-measure. The F-measure values exceed 84%. The results are very favorable and demonstrate the effectiveness of the use of resources from resource-rich languages to identify part-of-speech of under-resourced languages.

The result of POS-tagging of English plays a meaningful role in the building of the automatic training corpus for the Arabic tagging.

By making use of the language morphological differences between Arabic and English and the word-alignments in bilingual corpus, we are able to POS tagg Arabic text using the powerful English POS-taggers and the word alignment.

In future work we plan to apply our approach to a wider range of languages. Furthermore, we will exploit our approach to integrate POS in Arabic to English statistical machine translation (SMT) context for improving machine translation quality.

References

1. **Besacier, L., Le, V.-B., Boitet, C., & Berment, V. (2006).** ASR Translation for Under-resourced Languages. *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP*, pp. 1221–1224.
2. **Brown, P., Della-Pietra, V., Della-Pietra, S., & Mercer, R. (1993).** The mathematics of statistical machine translation: parameter estimation. *Computational Linguistics*, Vol. 19, No. 1, pp. 263–311.
3. **Dien, D. & Kiem, H. (2003).** POS-Tagger for English-Vietnamese Bilingual Corpus. *Proceedings of the HLT-NAACL 2003 Workshop on Building and Using Parallel Texts: Data Driven Machine Translation and Beyond*.
4. **Dinesh, K. & Gurpreet, S. J. (2010).** Part of Speech Taggers for Morphologically Rich Indian Languages: A Survey. *International Journal of Computer Applications*, Vol 6, No. 5.
5. **Dinh, D. & Hoang, K. (2003).** POS-Tagger for English–Vietnamese Bilingual Corpus. *Workshop: Building and Using Parallel Texts: Data Driven Machine Translation and Beyond*, Edmonton, CA.,
6. **El-Haj, M. (2014).** Creating language resources for under-resourced languages: methodologies, and experiments with Arabic. *Language Resources and Evaluation*, Springer.
7. **El-Haj, M. & Kruschwit, C. F. (2010).** UsingMechanical Turk to Create a Corpus of Arabic Summaries. *Proceedings of the International Conference on Language Resources and Evaluation*.
8. **Habash, N., Rambow, O., & Rot, R. (2009).** MADA+TOKAN: A toolkit for Arabic tokenization, diacritization, morphological disambiguation, POS tagging, stemming and lemmatization. *Proceedings of the 2nd International Conference on Arabic Language Resources*.
9. **Jurafsky, D. & Martin, J. H. (2002).** *Speech and Language Processing an Introduction to Natural Language Processing*. Pearson Education Series.
10. **Koehn, P., Hoang, H., Birch, A., Callison-Burch C., Federico, M., Bertoldi, N., Cowa, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., & Herbst, E. (2007).** Moses: Open source toolkit for statistical machine translation. *Proceedings of the ACL-2007 Demoand Poster Sessions*, Prague, Czeck Republic, pp. 177–180.
11. **Maamouri, M., Bies, A., Jin, H., & Buckwalter, T. (2003).** *Arabic treebank: Part 1 v 2.0*. Linguistic Data Consortium, Philadelphia, lDC catalogue number: LDC2003T06.
12. **Manning, C. D. (2011).** Part-of-speech tagging from 97% to 100%: Is it time for some linguistics? *Proceedings of the 12th International Conference on Computational Linguistics and Intelligent Text Processing*, CICLing'11, pp. 171–189.
13. **Mourad, A. & Darwish, K. (2013).** Subjectivity and sentiment analysis of Modern Standard Arabic and Arabic microblogs. *Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, Association for Computational Linguistics, Atlanta, Georgia, pp. 55–64.
14. **Och, F. J. & Ney, H. (2003).** A systematic comparison of various statistical alignment models. *Computational Linguistics*, Vol 29, No 1.
15. **Schmid, H. (1994).** Probabilistic Part-of-Speech Tagging Using Decision Trees. *Proceedings of International Conference on New Methods in Language Processing*, Manchester, UK, pp. 44–49.
16. **Tachbelie, M. Y., Solomon, T. A., & Besacier, L. (2011).** Part-of-Speech Tagging for Under-

Resourced and Morphologically Rich Languages: The Case of Amharic. *Conference on Human Language Technology for Development*, Alexandria, Egypt.

17. **Tackstrom, O., Das, D., Petrov, S., McDonald, R., & Nivre, J. (2012).** Token and Type Constraints for Cross-Lingual Part-of-Speech Tagging. *Transactions of the Association for Computational Linguistics*, pp. 1–12.
18. **Turki-Khemakhem, I., Jamoussi, S. & Ben Hamadou, A. (2010).** Arabic morpho-syntactic feature disambiguation in a translation context. *Proceedings of SSST-4 Fourth Workshop on Syntax and Structure in Statistical Translation.*

Ines Turki Khemakhem is a researcher and assistant at Sfax University. She received her Ph.D. in computer Science in 2016. She is currently a research member in the MIRACL (Multimedia InfoRmation system and Advanced Computing Laboratory) laboratory. Her Ph.D. thesis aims to integrate morpho-syntactic and semantic information for statistical machine translation. Her main interest focuses on Arabic language processing.

Salma Jamoussi is a researcher and assistant-professor at Sfax University in the higher institute of computer science and multimedia. She received her Ph.D. in computer Science in 2004 from the Henri Poincaré University, France. She focuses her research on classification methods, data mining and natural language processing.

Abdelmajid Ben Hamadou obtained a doctorate degree in computer science from the University of Orsay (France) in November 1979 and a These d'Etat in Computer Science from the University of Tunis (Tunisia) in March 1993. He is presently Professor of Computer Science at the Higher Institute of Computer science and Multimedia, Sfax-university and member of the Research Laboratory MIRACL at the same institution. In July 2002 he was decorated by the President of the Tunisian Republic ("Merit in Education and Science") and in May 2009, he received from the Vice President of Syria the "Al-Kindi" Award ("the best computer science researcher"). Abdelmajid Ben HAMADOU has published more than 280 articles in journals and conferences and has supervized more than 40 doctoral theses. His research domains are: Natural Language Processing, semantic web, information retrieval/filtering and document summarizing.

Article received on 15/07/2016; accepted on 09/09/2016.
Corresponding author is Ines Turki Khemakhem.