Sentence Boundary Detection for Transcribed Tunisian Arabic

Inès Zribi
ANLP Research group,
MIRACL Lab.
University of Sfax, Tunisia
ineszribi@gmail.com

Inès Kammoun
ANLP Research group,
MIRACL Lab.
University of Sfax, Tunisia
kammoun91ines@yahoo.fr

Mariem Ellouze
ANLP Research group,
MIRACL Lab.
University of Sfax, Tunisia
mariem.ellouze@planet.tn

Lamia Hadrich Belguith
ANLP Research group,
MIRACL Lab.
University of Sfax, Tunisia
l.belguith@fsegs.rnu.tn

Philippe Blache
Aix-Marseille University
& CNRS LPL, 13100,
Aix-en-Provence, France
philippe.blache@lpl-aix.fr

Abstract

In this paper, we study the problem of detecting sentence boundary in transcribed spoken Tunisian Arabic. We compare and contrast three different methods for detecting sentence boundaries in transcribed speech. The first method uses a set of handmade contextual rules for identifying the limit of sentences. The second method aims to classify words into four classes according to their position in a sentence. Both methods are based only on lexical and some prosodic information such as silent and filled pauses. Finally, we develop two techniques to mix the results of the two proposed methods. We show that sentence boundary detection system can improve the accuracy of a POS tagger system developed for tagging transcribed Tunisian Arabic.

1 Introduction

Automatic or manually transcription, generally, produces a set of texts that represent the contents of a speech. Transcripts need some more structuring or segmentation to be used in different spoken language processing systems (*e.g.,* speech summarization, speech translation, syntactic parsing, etc.), for which sentence is the basic unit. However, it is difficult to find speech sentences because of the absence of punctuation marks in the transcripts, which occur at sentences boundaries in most written languages. Moreover, sentences in spontaneous speech are ill-formed, and sentence boundaries are indistinct (Akita et al., 2006). Therefore, Sentence Boundary Detection (SBD) of transcripts is the preliminary step for multiple Natural Language Processing (NLP) applications.

Dialectal Arabic (DA) poses multiple challenges to SBD task due to the absence of resources. In addition, boundaries of dialectal sentences are related to different lexical cues (connectors such as امّا AmA "but", coordination conjunctions و w "and", etc.), which do not always present borders of sentences (Belguith et al., 2005).

We address, in this paper, the problem of SBD of manually transcribed Tunisian Arabic (TA). We present methods that exploit lexical and some prosodic cues for detecting the boundaries of TA sentences.

This paper is structured as follows: We first review some previous related work (Section 2). In Section 3, we present an overview of TA. We, then, highlight the challenges of SBD for TA (Section 4). Section 5 is devoted to presenting our data. We, then, present our methods (Section 6). In Section 7, we give the evaluation results. Finally, we conclude with a discussion of future work.

2 Related Works

Numerous techniques are used to recognize sentences boundaries for different spoken languages where several are based on statistical approaches using machine-learning techniques.

Jamil et al. (2015) have presented a supervised *Adaboost* classifier for SBD of spontaneous spoken Malay language. Their system is based on seven prosodic features, rate-of-speech and volume.

Beeferman et al. (1998) have developed *CYBERPUNC* that inserts punctuation in the transcripts of an automatic speech recognition system. Their system is solely based on lexical information. It relies on a trigram language model and a straightforward application of the Viterbi algorithm.

Using decision tree and hidden Markov modeling techniques, Shriberg et al. (2000) have combined prosodic cues with word-based approaches, and have evaluated performance on two speech corpora. Obtained results show that the probabilistic combination of prosodic and lexical information give the best result over English's task speech segmentation into sentence and topic units.

Akita et al. (2006) have tested two different techniques: statistical language model (SLM) and support vector machines (SVM) for SBD of spontaneous Japanese. In the SLM-based technique, they have used linguistic likelihoods and occurrence of pause to find sentence boundaries. They have, also, integrated heuristic patterns of end-of-sentence expressions to suppress false alarms. The SBD performed by an SVM-based text chunker (Akita et al., 2006) is based only on lexical and pause information.

Few researchers have investigated SBD of modern standard Arabic (MSA) textual data. Nevertheless, it is still not addressed for DA. These systems are based on lexical information such as conjunctions, punctuation marks, and other lexical items.

Belguith et al. (2005) have used contextual rules for developing the system *STAr* that is able to segment Arabic text in paragraphs and sentences. The rules are mainly based on punctuation marks, conjunctions and other connectors. Belguith et al. (2005) have used some collection of newspaper articles and school books for extracting rules.

Chaibi et al. (2014) and Keskes et al. (2012) have exploited (Belguith et al., 2005)'s method for segmenting Arabic texts in clauses and minimal discursive units.

A statistical approach is tested by (Khalifa et al., 2011) to segment Arabic text into sentences. They have proposed semantic based segmentation method that classifies the connector و w[1] "and" into their rhetoric roles. Khalifa et al. (2011) have trained a SVM classifier using syntactic and semantic features. According to the meaning of the connector, the generated model can segment Arabic texts.

Keskes et al. (2013) have tested a Maximum Entropy (ME) for classifying word in three different classes. Each class represents the position of the word in the minimal discursive unit. They have proved that typographical, lexical and morphological features are enough for detecting minimal discursive unit.

The SBD of transcribed MSA is addressed by (Elshafei et al., 2007). They have developed a system based on hidden Markov models (HMM) that accepts an oral sentence and its orthographic transcription, and generates its phonemic transcription and the segmentation information of sentence. The system is trained using a corpus of Arabic TV news and is validated against manually segmented speech sentences (Elshafei et al., 2007).

3 Tunisian Arabic

Tunisian Arabic (TA) is a dialect of the North African (i.e., the Maghreb) dialects spoken in Tunisia (Zribi et al., 2014). It is considered a low variety given that it is neither codified nor standardized even though it is the mother tongue and the variety spoken by all the population in daily usage (Saidi, 2007). Approximately eleven million people speak at least one of the many regional varieties of TA (Zribi et al., 2014).

There are many differences as well as similarity points between TA and MSA in different levels. In order to compare these two varieties of Arabic language, we focus on four levels (i.e. the phonological level, the morphological level, the lexical level and the syntactic level).

3.1 The phonological level

The vocalic system of TA is reduced (Tilmatine, 1999). Some short vowels are neglected, especially if they are located in the last position of the word (Mejri et al., 2009). The MSA verb شرب /šariba/ "he drank" is pronounced /šrib/ in TA. We note the deletion of the vowels at the first and the last position of the verb. TA has, also, a long vowel /e:/ which does not exist in MSA (Zribi et al., 2014). Moreover, the consonant system includes some phonetic differences (Mejri et al., 2009). In some cases, the Arabic consonant ق /q/ is pronounced /g/. The MSA word بقرة bqrħ /baqara/ "cow" is pronounced in TA /bagra/. In addition, some consonants in TA have multiple pronunciations. For

[1] Transliteration is coded with Buckwalter transliteration. For more details about it, see (Habash et al., 2007).

example, the MSA consonants غ γ /γ/ and ج j /j/ can be pronounced in TA respectively as /x/ or /γ/ and /j/ or /z/.

3.2 The morphological level

The main difference between MSA and TA is on the affix level. We notice the presence of new dialectal affixes and the deletion of others. Dual suffixes ان An and ين yn are generally absent in TA. They are replaced by the numeral زوز[2] zwz "two" located after or before the plural form of the noun. However, some words in TA can be agglutinated to the suffix ين yn to express duality. In verb conjugation, TA is characterized by the absence of the dual (feminine and masculine) and the feminine in the plural. It has seen many simplifications in its affixation system (Ouerhani, 2009). Indeed, new affixes have appeared. The first one is the negation clitic. It is agglutinated to the last position of the verb that must be preceded by the negation particle مَا mA (*e.g.,* مَاكليتش mA klytš "I don't eat") (Mejri et al., 2009). The interrogation prefix of MSA أَ Â is transformed in TA into the suffix شي šy (*e.g.,* خرجشي, xrjšy, "Did he go out?"). Likewise, the future prefix سـ s- is replaced by the particle (بَاش bAš "will"). In addition, we note the absence of the dual clitics in TA.

3.3 The lexical level

TA is distinguished by the presence of words from several other languages. The presence of these languages mainly occurred due to historical facts. We find in Tunisia a significant amount expressions and words from European languages such as Spanish, French, Italian, Turkish and even Maltese (*e.g.,* قطوس qTws "cat" is of Maltese origin; كوجينة kwjynħ "kitchen" is of Italian origin; بلَاصة blASħ "place" and بَاكو bAkw "package" are derived from French language). In addition, TA has several words from the vocabulary of the Berber language (*e.g.,* برنوس, brnws, "traditional clothes") (Zribi et al., 2014).

In addition to all these borrowed terms, which have been integrated in the TA morpho-phonology, Tunisians code switch often in daily conversations, particularly from French (*e.g.,* ça va ? "Okay?", désolé "sorry", rendez-vous "meeting", etc.). All these expressions and words are used without being adapted to TA phonology.

3.4 The syntactic level

The syntactic differences between MSA and TA are minors. The MSA word order is generally VSO (Verb subject Object) especially in verbal sentences. But in TA, the preferred word order is SVO (Mahfoudhi, 2002). The VSO and VOS orders are also used in TA.

4 Challenges in TA Sentence Boundary Detection

Arabic language characteristics.

SBD is a challenging task for Arabic language that is characterized by the absence of capital letters and the boundaries of sentences are not generally marked with punctuation marks. We often find a paragraph in Arabic language, which has only one full stop. Boundaries of Arabic sentences are strongly related to conjunctions and other lexical expressions. These lexical cues are not necessarily present sentence limits. They have other discursive functions. For example, the interjection (بَاهي, bAhy, "OK") can be used as an adjective that means "good".

Spoken language characteristics.

The spoken form of the Arabic language presents other challenges for the task of SBD. Firstly, the transcripts are usually not punctuated. Similarly, linguists interested in speech quickly deserted the notion of sentence (Tellier et al., 2010). We have to define, first, the term *sentence*. In TA oral, we can detect several types of sentences: well-formed sentences, incomplete sentences, and sentences containing disfluent segments. The incomplete sentences are very frequent in oral. The disfluency, also, affects the structure of the sentences by involving several elements of different nature in a sentence. Truncated words, filled pauses, silent breaks, repetitions, etc. affect the syntactic structure of the sentence. So, it is necessary to define the units of statement that we suggest detecting its boundary.

Tunisian Arabic characteristics.

TA is a spoken variety of Arabic that Tunisians code switch between MSA and French language.

[2] We follow the CODA-TUN convention (Zribi et al., 2014) when writing examples of words in TA.

The massive use of words from foreign languages and code switching engender in certain cases a loss of the syntactic structure of sentences. Indeed, TA is characterized by an irregularity in the word order in the sentence. We can express a single sentence with several syntactical structures: Subject-Verb-Object (SVO), Verb-Subject-Object (VSO) and Object-Verb-Subject (OVS) (Mahfoudhi, 2002). The mix of language (MSA, TA, and French) and the free word order for TA increase the difficulty of SBD.

Consider the English sentences: "*It is true that we are today... It is a day of celebration, but we have to work...*". These sentences can be translated into the following sentence:
(c'est vrai امَا c'est le jour de la fête اليوم احنَا اللّي نخدموَا يلزمنَا, c'est vrai Allly AHnA Alywm c'est le jour de la fête AmA ylzmnA nxdmwA).

The translated sentence is composed of the French phrase (*c'est vrai*, "it is true"), the French sentence (*c'est le jour de la fête*, "it is a day of celebration") and a set of TA words. SBD of such a sentence, which is very frequent in daily speech of Tunisians, is very difficult. Indeed, in French grammar, the expression (*c'est*, "it is") always marks the beginning of a new sentence. However, this expression can be used anywhere in TA sentence. The first occurrence of the expression (*c'est*) introduces the start of a new sentence, but it is not the case for the second occurrence.

To conclude, the presence of many foreign words in the TA speech and the code switching phenomena improve the difficulties of SBD of TA.

5 Data

5.1 Presentation

In this work, we used a manually transcribed TA corpus, created by (Zribi et al., 2015), and labeled as "STAC". The corpus consists of about 42,388 words, and follows the CODA-TUN (Zribi et al., 2014) convention for writing TA words and OTTA guideline (Zribi et al., 2013) for annotating the phenomena of the oral. The corpus is morphosyntatic annotated and segmented into sentences. Speech text for each speaker is divided into many speech turns. Zribi et al. (2015) gathered the speech turn for each speaker in a unique text. They, then, segmented it in utterances. They, considered a sentence a semantically meaningful unit.

5.2 Preparation

In STAC Corpus, the experts have performed the segmentation manually. We have redone the segmentation of the corpus with two experts to validate the segmentation of sentences. We have calculated the inter-annotator agreement. The two experts achieved a Kappa coefficient rate of 0.86% indicating almost perfect agreement.

All types of annotations are removed from the corpus. We kept only annotations that mark incomplete words, filled pauses and named entities. We eliminated, also, all specific symbols from the corpus.

The STAC corpus is divided into three parts. The first part of the STAC corpus was used for training our methods. It is composed of 32,012 words and 6,133 sentences. The second part comprised 7,201 words and 1,215 sentences to test the different proposed approaches. The remaining part of the STAC corpus (440 sentences and 3,175 words) is used for development.

6 Our Methods

In this section, we describe three methods for SBD of TA. Our proposed methods belong to three approaches: rule-based, statistical and hybrid.

6.1 Rule-based method

Rule-based techniques are proposed for developing MSA SBD systems. The handmade rules are essentially based on punctuation marks, conjunctions and other connectors. We propose to apply this technique for segmenting TA transcripts. We have used *lexical items* (such as conjunctions and other markers) and *two simple prosodic features* (silent and filled pauses) for designing our SBD rules.

The *lexical markers* are in certain cases specific to oral. In others, they can be used in the written form of the dialect. We have classified our rules following this criterion. The role of our segmentation rules is to detect a word (or an expression) at the beginning of a sentence. The rules are, also, based on words belonging to the right and/or the left context. We call them contextual rules (CR).

Contextual rules follow the same structure as defined by (Belguith et al., 2005). They have the following form:

Left Context	Marker	Right Context
G	X	D

G, *X* and *D* present lexical items which can be the beginning of a sentence. X is a trigger marker. If the left context *G* and/or the right context *D* are present, then *X* or *D* can be the beginning of a sentence. The window size of right and left context is variable according to the number of words that compose the lexical markers.

We have extracted two sets of rules. The *first set* groups rules that detect sentence boundaries of the oral form of TA. These rules are based on oral *specific lexical items* and *prosodic features*. Indeed, silent pauses are located in 57.25% of the cases at the first position of sentences. In this case, the silent pause can be compared to a full stop in writing texts. However, in 42.75% of cases, silent pauses are in the right or the left context of the first position of the sentence. Filled pauses are also located in the last or the first position of sentences. Based on these two prosodic features, we have extracted six contextual rules.

Below (See Table 1) is an example of contextual rule based on a silent pause and some lexical features. If the trigger marker is equal to a silent break "#" and the left context belongs to this list of words, then, the break is a mark of the beginning of a sentence.

Left Context	Marker	Right Context
Interrogative Adverb : عَلاش ʕlAš "why", قدَاش qdAš "how much", etc. **Expression that marks time:** عَام كل kl ʕAm "every year", غدوة γdwħ "tomorrow", etc.	#	∅

Table 1: Contextual rule based on the silent pause.

The *second set* of rules is more generic. It can be applied to the written form of TA. Rules conception is based on connectors, personal and relative pronouns, verbs, etc. Indeed, the syntactic structure of TA is very complex. Thus, we had difficulty in identifying patterns to detect the boundaries of sentences since the STAC corpus is an oral corpus with a high degree of spontaneity (95.65%). Sentences with simple structure present only 15.86% of our corpus. Below (See Table 2) is an example of rule that detects boundaries of sentences based on verbs.

Left Context	Marker	Right Context
∅	**Verb**	الَاقل علَى ʕlỳ AlAql "at least", أَيَا ÂyA "come on", لَا lA "no", وَاِلَا wAilA "otherwise", معنَاتهَا mʕnAthA "that is to say", لهنَا lhnA "here", etc.

Table 2: Contextual rule based on a verb.

This rule allows the detection of sentence that begins with a verb proceeded by an expression belonging to this list.

At the end, we have extracted in total 23 contextual rules. During the design of our rules, we have kept only rules that their precision is superior to 50%.

6.2 Statistical method

We have experimented with another approach for the SBD of TA. The task of SBD is converted into a word classification. We have proposed to classify words into four classes:

- "B-S" for marking the first word of the sentence,
- "I-S" for marking the word in the sentence,
- "E-S" for marking the last word of the sentence,
- and finally "S" for marking sentence composed of a single word.

We have built a classifier based on the rule-based classifier PART (Mohamed et al., 2012). Part is a partial decision tree algorithm, which is the development version of C4.5 and RIPPER algorithms (Mohamed et al., 2012). The main specialty of the PART algorithm is that does not need to do global optimization like C4.5 and RIPPER to generate exact rules, but it practiced separately and-conquer

strategy. For example, it builds a rule, and removes instances. It covers, and continues to create a recursive rules for the remaining of instances until there are no instances. PART builds a partial C4.5 decision tree in every iterative and makes the "best" leaf into a rule (Mohamed et al., 2012).

We have experimented with other classification methods included in the WEKA machine-learning tool[3]. However, PART gives the best results for our task.

The result of a classifier is strongly influenced by the set of defined features. In literature, the SBD task for spoken language is mainly related to two types of features: linguistic and prosodic features. The prosodic information (such as intonation, rhythm, etc.) is absent in our work. Thus, we have used two simple prosodic features that are silent and filled pauses. In the design of our features, we rely on linguistic features like adverbs, adjectives, verbs, etc. We note that we use lexicon lookup for determining words part-of-speech .

We have also used contextual features. To fix the window size, *n*, we have tested several contexts. We have experimented with *n=0*, *n=1*, and *n=2*. We show that *n=2* is the best configuration for our task.

Finally, we have used dynamic feature. It uses the class that is dynamically assigned to the two preceding words. Features given to PART are presented in Table 3. We note that the features take two possible values: *true* or *false*. They specify whether a word in the context belonging to the possible values set.

Features	Examples of the possible value
Silent Pause	#
Filled Pause	آ Ā "euh"
Expression marking the beginning of sentence	لكن lkn "but"
Conditional particle	ولأَن wlÂn "and because"
Discursive marker	معنَاتهَا mʕnAthA "that is to say"
Expression marking place	ثمة θmħ "there is"
The verb "want"	حبيت Hbyt "I want"
The verb "say"	يقول yqwl "he says"
Verb	TA verbs
Personal pronoun	آنَا ĀnA "I"
Verb "to be"	كَان kAn "he was"
Relative pronoun	اللّي Ally "that"
Demonstrative pronoun	هذية hðyħ "this"
Expression marking the time	كل عَام kl ʕAm "every year"
Interrogative adverb	علَاه ʕlAh "why"
Special expression	بصرَاحة bSrAHħ "honestly"
Greeting expression	عَالسلَامة ʕAlslAmħ "hello"

Table 3: Features for PART classifier.

6.3 Hybrid method

We have proposed to combine the result of the rule-based method and the statistical method. We have tested three different methods for combining the results of the two previous methods.

The *first method* consists of analyzing the transcripts using the contextual rules. The output of this step is a set of sentences. We have reanalyzed the longer sentences with the statistical model. We consider that a sentence is long only if the number of words is higher than 9. Nine words was chosen because it is the average number of words per TA sentence and nine gives us the best development results.

The *second method* is the opposite of the first method. It consists of applying, in the first step, the model generated by the statistical method. Then, we apply for the longer sentences contextual rules for segmenting them.

The *third method* consists of using the generated rules from the PART algorithm. We have suggested using simultaneously contextual rules and the generated rules by the algorithm PART for segmenting TA transcripts.

PART algorithm extracts a set of rules from the training corpus that classify words to four classes (B-S, I-S, E-S and S). These rules have the following form: "*if condition(s), then conclusion*".

We have chosen rules that classify words into "B-S" and "S". These rules can detect words at the first position of the sentence and word that presents a whole sentence. Only no redundant rules are

[3]http://www.cs.waikato.ac.nz/ml/weka/

selected. We have attributed to each rule a score. We have calculated it by applying the rule to a validation corpus composed of 440 sentences. This corpus is not used for generating rules. We have calculated the success rate for each rule. If its rate exceeds 75% we kept the rule. We remark that 40 rules attribute incorrectly class. All remaining rules are equal to handcrafted contextual rules.

Therefore, this method fails to integrate automatic generated and handcrafted rules. However, it shows that the automatic rule extraction can generate rules equal to handmade rules.

7 Evaluation

We look first at the performance of the three SBD methods proposed in this paper. We compare these methods against the baseline. Then, we test the effect of SBD methods on POS tagging of transcribed spoken TA.

7.1 Results and discussion

The evaluation metrics we use are recall, precision and F-measure. We have evaluated how well we could correctly segment TA transcripts. In this evaluation, we have compared our proposed methods: rule-based method (CR), statistical method (PART) and two hybrid methods (Hyb1 and Hyb2) against the baseline.

We have used STAr system (Belguith et al., 2005) as our baseline. STAr is SBD system designed for written form of MSA and it is based on a set of contextual rules. We have chosen STAr since some of its contextual rules are shared with TA. These rules are based primarily on the coordination conjunction (و, w, "and").

Table 4 lists the results of the different methods. We see that running statistical method alone gives us the best SBD results. We reported improvements up to 27.35% compared to the baseline. We see that the STAr system performs poorly on TA input. However, the precision value of the baseline is good (82.45%). This is due to the high number of TA sentences that begin with the coordination conjunction (و, w, "and"). The results given by rule-based method are lower than those of statistical method. Indeed, some of lexical markers are located far from sentences limits. This is due to the relative free order of some TA sentences. As well, some markers have other discursive functions that falsified the output of the application some contextual rules.

We turn now to analyze the hybrid methods results. The application of the first hybrid method has improved the recall value of the contextual rules. We notice an improvement of 4.11%. By against, it decreases the precision value (5.65%) compared to the method PART. We see that the application of PART algorithm followed by contextual rules downgraded the recall value. The value fell down from 72.42 to 66.00 (a decrease of 6.42%).

The second step of the two hybrid methods divides the long sentences into very small segments. This segmentation increases the number of sentences, but it decreases the accuracy of the SBD.

In conclusion, we note that the rule-based method and statistical method are powerful for the task of SBD. However, the higher increase (gain) has been observed in statistical method.

	Recall	**Precision**	**F-measure**
Baseline	40.98	82.45	54.75
CR	68.31	90.841	77.98
PART	**72.5**	**94.8**	**82.1**
Hyb1	72.42	89.15	79.92
Hyb2	66.00	73.91	69.73

Table 4: Comparison of the performance of the different SBD methods.

7.2 Extrinsic Evaluation: POS tagging of Tunisian Arabic

Part-of-speech tagging task (POS tagging or POST), is the process of marking up a word in a text (corpus) as corresponding to a particular part of speech, based on both its definition and its context — *i.e.*, its relationship with adjacent and related words in a phrase, sentence, or paragraph[4]. Thus, the detection of sentence (in writing language) and utterance boundaries (in spoken language) is considered one of the necessaries preliminary steps. Indeed, SBD for written languages is trivial due to the presence of punctuation marks and capital letters notably on Indo-European languages. Contrariwise, it is not trivial for spoken languages, specifically for spoken Arabic dialect.

We present in this section the effect of sentence boundary detection on POS tagging of transcribed spoken TA. Here, we are evaluating a POS tagger for TA trained on the STAC corpus (Zribi et al., 2015). The proposed tagger is tested with three different training methods: the statistical method

[4]https://en.wikipedia.org/wiki/Part-of-speech_tagging

SVM (Vapnik, 1995) and two rule-based classifiers (Ripper (Cohen, 1995) and PART (Collins and Singer, 1999)). We compare the performance of this tagger when it is trained on a manually (HandSeg), automatically (AutSeg) and non-segmented (NoSeg) version of the STAC corpus. In order to make the best use of our corpus, we tested our POS tagger using a 10-fold cross-validation procedure. Table 5 shows the result of the evaluation.

We remark that the SBD system helps the TA POS tagger to improve its accuracy. We note that SVM and RIPPER performed better when the SBD system detects short sentences. The value of accuracy of our POS tagger trained on SVM has decreased from 61.78% (non-segmented corpus) to 63.66% (corpus segmented with the second method of hybridization). Likewise, the accuracy increases from 62.53% to 64.84% when Ripper is used for training the tagger. However, the PART algorithm works best with long sentences. We show that the best value is given by using non-segmented corpus.

		Ripper	**PART**	**SVM**
NoSeg		62.53	**71.88**	61.87
HandSeg		63.92	70.55	63.02
AutSeg	**PART**	61.69	66.58	61.04
	CR	**64.84**	70.65	63.04
	Hyb1	64.20	70.21	63.39
	Hyb2	63.92	68.22	**63.66**

Table 5: The accuracy values of the POS tagger trained and tested with a manually (HandSeg), automatically (AutSeg) and non-segmented (NoSeg) corpus.

8 Conclusion

In this paper, we have proposed three different methods for detecting Tunisian Arabic sentence boundaries. We have experimented a rule-based, statistical, and hybrid method. These different methods are based on linguistic and two simple prosodic cues. The proposed method has shown encouraging results.

As future work, we intend to add more prosodic features to improve the efficiency of our system. We also intend to realize an extrinsic evaluation of our system in some NLP applications dealing with the spoken form of Tunisian Arabic. Finally, we aim to expand the training and the test corpora to cover other types of TA sentences.

References

Yuya Akita, Masahiro Saikou, Hiroaki Nanjo, and Tatsuya Kawahara. 2006. Sentence boundary detection of spontaneous japanese using statistical language model and support vector machines. In *INTERSPEECH 2006 - ICSLP, Ninth International Conference on Spoken Language Processing, Pittsburgh, PA, USA, September 17-21, 2006*.

Doug Beeferman, Adam L. Berger, and John D. Lafferty. 1998. Cyberpunc: a lightweight punctuation annotation system for speech. In *Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP '98, Seattle, Washington, USA, May 12-15, 1998*, pages 689–692.

Lamia Hadrich Belguith, Leila Baccour, and Ghassan Mourad. 2005. Segmentation de textes arabes basée sur l'analyse contextuelle des signes de ponctuations et de certaines particules. In *TALN 2005*.

Anja Habacha Chaibi, Marwa Naili, and Samia Sammoud. 2014. Topic segmentation for textual document written in arabic language. *Procedia Computer Science*, 35:437 – 446. Knowledge-Based and Intelligent Information ; Engineering Systems 18th Annual Conference, KES-2014 Gdynia, Poland, September 2014 Proceedings.

William W. Cohen. 1995. Fast effective rule induction. In Morgan Kaufmann, editor, *In Proceedings of the Twelfth International Conference on Machine Learning*, pages 115–123.

M. Collins and Y. Singer. 1999. A simple, fast and effective rule learner. *Proceedings of the 16th National Conference on Artificial Intelligence (AAAI-99)*, pages 335–342.

Moustafa Elshafei, Mohammad Ali, Husni Al-Muhtaseb, and Mansour Al-Ghamdi. 2007. Automatic segmentation of Arabic speech. In *Workshop on Information Technology and Islamic Sciences.*

Nizar Habash, Abdelhadi Soudi, and Timothy Buckwalter. 2007. *Arabic Computational Morphology: Knowledge-based and Empirical Methods*. Springer Netherlands.

Nursuriati Jamil, M.I. Ramli, and N. Seman. 2015. Sentence boundary detection without speech recognition: A case of an underresourced language. *Journal of Electrical Systems*.

Iskandar Keskes, Farah Benamara, and Lamia Hadrich Belguith. 2012. Clause-based discourse segmentation of arabic texts. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, may. European Language Resources Association (ELRA).

Iskander Keskes, Farah Benamara Zitoune, and Lamia Hadrich Belguith. 2013. Segmentation de textes arabes en unités discursives minimales. In *Conférence du Traitement Automatique des Langues Naturelles - TALN 2013*, pages pp. 435–449, Sables d'Olonne, FR. LINA - Laboratoire d'Informatique de Nantes Atlantique.

Iraky Khalifa, Zakareya Al Feki, and Abdelfatah Farawila. 2011. Arabic Discourse Segmentation Based on Rhetorical Methods. *International*

Journal of Electric and Computer Sciences IJECS-IJENS, 11(01):10–15, February.

Abdessatar Mahfoudhi. 2002. Agreement lost, agreement regained: A minimalist account of word order and agreement variation in arabic. *California Linguistic Notes*, XXVII(2).

Salah Mejri, Mosbah Said, and Inès Sfar. 2009. Pluringuisme et diglossie en tunisie. *Synergies Tunisie*, 1:53–74.

W. N. H. W. Mohamed, M. N. M. Salleh, and A. H. Omar. 2012. A comparative study of reduced error pruning method in decision tree algorithms. In *IEEE International Conference on Control System, Computing and Engineering (ICCSCE)*, pages 392–397, Nov.

Béchir Ouerhani. 2009. Interference entre le dialectal et le littéral en tunisie: Le cas de la morphologie verbale. In *Synergies Tunisie n1*, pages 75–84.

Darine Saidi. 2007. Typology of Motion Event in Tunisian Arabic. In *LingO*, pages 196–203.

Elizabeth Shriberg, Andreas Stolcke, Dilek Z. Hakkani-Tür, and Gökhan Tür. 2000. Prosody-based automatic segmentation of speech into sentences and topics. *Speech Communication*, 32(1-2):127–154.

Isabelle Tellier, Iris Eshkol, Samer Taalab, and Jean-Philippe Prost. 2010. Pos-tagging for oral text with crf and category decomposition. *Research in Computing Science*, 46:79–90.

Mohamed Tilmatine. 1999. Substrat Et Convergences: Le Berbère Et L'arabe Nord-Africain. *Estudios de dialectología norteafricana y andalusí*, pages 1–3.

Vladimir N. Vapnik. 1995. *The Nature of Statistical Learning Theory*. Springer-Verlag New York, Inc., New York, NY, USA.

Inès Zribi, Marwa Graja, Mariem Ellouze Khmekhem, Maher Jaoua, and Lamia Belguith Hadrich. 2013. Orthographic Transcription for Spoken Tunisian Arabic. In *14th International Conference CICLing 2013, Proceedings, Part I, Samos, Greece, March 24-30*, volume 7816 of *LNCS*, pages 153–163. Springer.

Inès Zribi, Rahma Boujelbane, Abir Masmoudi, Mariem Ellouze, Lamia Belguith Hadrich, and Nizar Habash. 2014. A Conventional Orthography for Tunisian Arabic. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'2014), Reykjavik, Iceland, May 26-31*, pages 2355–2361. ELRA.

Inès Zribi, Mariem Ellouze, Lamia Hadrich Belguith, and Philippe Blache. 2015. Spoken Tunisian Arabic Corpus " STAC": Transcription and Annotation. *Research in computing science*, 90.