International Arab Conference on Information Technology (ACIT'2016)

Word Boundary Detection in Tifinagh using MaxEnt and n-gram algorithms

Mohamed Biniz[1],Rachid El ayachi[2], Mohamed Fakir[3]
Laboratory of Information Processing and Decision Support (TIAD).
Faculty of Sciences and Technics, University Sultan MoulaySlimane.
mohamedbiniz@gmail.com[1],rachid.elayachi@usms.ma[2], fakfad@yahoo.fr[3],

Abstract:*This article proposes the use of Maximum Entropy and n-gram algorithms for the automatic segmentation of words in Tifinagh.The Maximum Entropy algorithm is a probability distribution widely used for a variety of natural language processing tasks, such as sentence boundary detection, part of speech, etc.The maximum entropy formulation has a unique solution that can be found by the Generalized Iterative Scaling algorithm.Our experimental results show that the model of maximum entropy using the approach based character considerably improves the quality of the segmentation of words in Tifinagh.*

Keywords:*Maxent, Tifinagh, WBD, Token, NLP.*

1. Introduction

Word Boundary Detection(WBD) is the most basic task in the field of treatment of written natural language.In most systems, this task is called tokenization and can be easily resolved by interpreting the word orthographic markers and end of a sentence, for example, white spaces.

Unlike English languages, many languages like Chinese, Japanese and Tifinagh, do not delimit words with white space. Where these are lacking, however, the word segmentation is a difficult task. Which makes the construction of a system for segmenting words is complicated by the fact that there is no standard definition of the word boundaries in Tifinagh.

For the Tifinagh language, syllable [1] [2] is the smallest linguistic unit, and a word consists of one or more syllables.It leads to a problem of ambiguity in determining word boundaries.There are two types of ambiguity namely the cross ambiguity and overlapping ambiguity.In the cross-ambiguity, a few syllables themselves have the meaning (can be a word), and their combination also has a meaning.And the overlap ambiguity occurs in a situation that when a syllable is combined with the preceding syllable or following syllable in a phrase generates words.

Besides ambiguity limits,the word segmentation in Tifinagh faces a problem in which there are many new words appearing in a document.These new words are usually the names that refer to individuals, location, abbreviation for foreign words, etc.

Many current approaches for other languages are suffering from lack of accurate inference sequences or knowledge handling difficulties derived from a corpus (this knowledge is not used, or used in a limited way, or used in a complicated manner, etc.).For example, the approach based on modelling N-gram [3] does not use the domain knowledge. Zhang&Al [4] use a hierarchical hidden Markov model to incorporate lexical knowledge.

A recent development in this area is Xue [5], in which the author uses a sliding window of maximum entropy to mark the Chinese characters in one of the four position tags, then convert these tags in a segmentation using rules.For the Tifinagh language, like most languages that have begun to be studied recently for Natural Language Processing (NLP), is still suffering from the scarcity of tools and language processing resources.Until the writing of this article does not find any work that deals word boundary detection.

This article is organized as follows: section 2 provides a brief description of a set of work in this field for other languages, Section 3 presents the methodology of the proposed approach. Then the integrated assessment procedure to test the performance of the developed system is described in Section 4, while Section 5 concludes the work.

2. Related Word

In the literature, the majority of the work was concentrated on the English [6] Japanese [7], Chinese [8]language.Their objectives are the segmentation of words and word identification tokens in a text in progress, or a large dictionary.When talking about token, there are two main approaches of segmentation

[9]: the word-based approach and character-based approach.

2.1 Word-Based Approach

The idea of this approach is to read the input sentences from left to right, and predict whether the current piece of continuous characters is a word token.When you find a word, the segmenters move and looking for a possible new word.There are various works for the prediction of problems based on words.For example, the work of Chen &Al [10] using the maximum matching algorithm (if a sequence of characters that appear in a dictionary, it is considered a candidate word).

2.2 Character-Based Approach

This approach [11] aims to classify characters according to their positions in the words, Segmentation can be treated as a sequence labelling problem,that is to sayassign labels to the characters in a sentence indicating whether a character in a word is a single character or there in the beginning, the middle or the end of a multi-character word.For word prediction, word tokens are deducted based on the character classes.

3. Methodology

In the proposed approach, the Word boundary detection was considered as the classification problem by using Maximum Entropy algorithm to identify the tokens in a corpus. Our approach is to classify all words in either a "yes" class (a word boundary) or "no". Class "yes" break points is similar to blank spaces in sentences in English. The sentence (□ □ □ □ □ □ □ □ □ □ □ □ □ □ □) Contains four tokens using the naive method of segmentation (segmentation by white spaces), but if we use the immediate constituent analysis [12], seven tokens is obtained (□ □ <s> □□ □ □ □ □ □ □ □ □ □ □ <s> □ <s>:).

The system developed in this work allows the detection of tokens of a text written in Tifinagh.It includes all the steps described in the following figure:

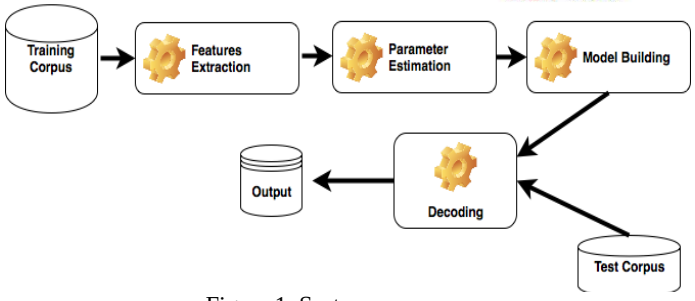

Figure 1: System process

First, the features extraction phase is interested in the release of features from a training corpus.Then, the maximum likelihood parameters are estimated in the second phase.Next, in the third phase a Maximum Entropy algorithm is adopted for model development.Finally, the decoding phase allows detection of tokens in a test corpus based on the constructed model.

Prior to the detection of word boundaries of a text, we define a set of features, including: character, word, punctuation. These characteristics differ from one language to another.The procedure of segmentation follows the following rule: two segments are separated by spaces.A segment that is followed by punctuation is considered probable end of the word (respecting the particular form of each language). The corpus is composed of texts of various styles: poetry, magazines, literature, etc.It should be noted that the corpus used Tifinagh reached 4035 words.

3.1 Features

At this level, we will describe the choice of parameters for the recognition of ends words to both approaches: word-based approach and character-based approach.

In the character-based models, the features are generally defined by the character information in the window n characters.Despite a number of interesting features that may be expressed, it is slightly less natural to encode information of a predicted word. On the contrary, taking the words as dynamic tokens in the word-based approach, it is very easy to define the symbolic features of a word.However, the word based segmenters have greater power of representation.Despite the lack of representative capacity, the characters based segmenters can use word-like functionality by consulting a dictionary.For example, if a string matches a word type, it has a very high probability to be a token word.For both approaches, the Li template feature[13] was used, as shown tables 1 and 2:

Table 1: template feature based character

Type	Features	Function
Character Unigram	C_0 , C_1	The single character features
Character Bi-gram	C_1C_0 , C_0C_1 , C_1C_2	The character bi-gram features

Transition	$T_1C_0T_0$, $C_1T_1C_0T_0$, $T_1C_0T_0C_1$	The character adding tag transition features

Table 2: template feature based word

Type	Features	Function
Word Unigram	W_0 , W_1	The single word features
Word Bi-gram	W_1W_0 , W_0W_1 , W_1W_2	The word bi-gram features
Transition	$T_1W_0T_0$, $W_1T_1W_0T_0$, $T_1W_0T_0W_1$	The word adding tag transition features

WhereT_n represents the borderline of each word, that is to say the end of a word or not.

3.2 Maximum Entropy Model

The results of the probability model are "yes" and "no" where "yes" denotes a potential word boundary, and "no" indicates that there is not a limit of a real word.

Each token border potential word, we want to estimate a joint probability *p* of it and its surrounding context that considers itself a real word of the border probability distribution. we proceed by using a maximum entropy model defined by equation (1):

$$p(a|b) = \frac{1}{Z(b)} \prod_{j=1}^{k} \alpha_j^{f_j(a,b)} \qquad (1)$$

Where k is the number of features and Z (b) is a normalizing factor to ensure that$\sum_a p(a|b) = 1$. Z (b) is defined by the following formula:

$$Z(b) = \sum_a \prod_{j=1}^{k} \alpha_j^{f_j(a,b)} \qquad (2)$$

Context predicates considered useful for word boundary detection, we are encoded in the model using features. For example, a useful function could be:

$$f_j(a,b) = \begin{cases} 1 & if\ Prefix(b) = I and\ a = no \\ 0 & otherwise \end{cases} \qquad (3)$$

This feature will allow the model to discover that the word"□"rarely occurs as a word boundary.Therefore, the setting for that function will hopefully stimulate the probability if the prefix is "□".All functions that occur 10 or more times in the training data are stored in the model, and the model parameters are estimated with the "Generalized Iterative Scaling" algorithm [14].

3.3 Generalized Iterative Scaling Algorithm

The algorithm "Generalized Iterative Scaling", or GIS allows us to estimate the parameters ***p**** (maximum likelihood).It requires that the sum of the features is equal to a constant C for any (a,b)∈AxB, as shows in equation (4):

$$\sum_{j=1}^{k} f_j(a,b) = C \qquad (4)$$

In the case where this condition is not verified, we use the total training to choose C, which is defined by the equation (5):

$$C = \max_{a \epsilon A, B \epsilon T} \sum_{j=1}^{k} f_j(a,b) \qquad (5)$$

Then added a "correction" expressed by the equation (6):

$$f_l(a,b) = C - \sum_{j=1}^{k} f_j(a,b) \quad where\ l = k+1 \qquad (6)$$

For any (a, b) pair, note that *f (a, b)* is between 0 and C, where C may be greater than 1.In theory, a constant correction to demand the constraint in equation (5) for all (a, b) pairs should be derived from the space of possible events AxB.

However, one of the sets of the training is usually accurate in practice.

GIS Algorithm:

- Initialization:

$$\alpha_j^{(0)} = 1$$

- Iteration until convergence:

$$\alpha_j^{(n+1)} = \alpha_j^{(n)} \left[\frac{E_{\tilde{p}} f_j}{E_{p(n)} f_j}\right]^{\frac{1}{C}}$$

Such as : (7)

$$E_{p(n)}f_j = \sum_{a,b} \tilde{p}(b)p^{(n)}(a|b)f_j(a,b)$$

$$p^{(n)}(a|b) = \frac{1}{Z(b)} \prod_{j=1}^{l} (\alpha_j^{(n)})^{f_j(a,b)}$$

4. Experimental results:

To evaluate our system, we adopted three measures: Precision, Recall and F-measure [15].

Precision is defined as the total number of word boundaries properly extracted*Tn* on the total number of words extracted by the system*Nt*.

$$Precision = \frac{Tn}{Nt} \quad (8)$$

Recall is defined in equation (9) as the total number of word boundaries properly extracted*Nt* on the true limits of the total number existing in the corpus *Nc*.

$$Recall = \frac{Nt}{Nc} \quad (9)$$

F-measure is the harmonic mean of precision and recall, which is a good indicator of system performance.

$$F - measure = \frac{2 * Precision * Recall}{Precision + Recall} \quad (10)$$

In the experimental part, we use the implementations OpenNLP [16] Version 1.6.3 with specific adaptation to support the Tifinagh language. The corpus used contains in total 4035 words. This corpus is divided into two parts:

- Test Corpus
- Training Corpus

Corpus segmentation is done manually using the immediate constituent analysis [12].

The following tables show the results for both approaches:

Table 3: Results obtained using the based approach character

Training Corpus	Test Corpus	Precision	Recall	F-measure
1000	3035	0.59	0.54	0.56
2000	2035	0.64	0.67	0,65
3000	3035	0.78	0.84	0.80

Table 4: Results obtained using the word-based approach

Training Corpus	Test Corpus	Precision	Recall	F-measure
1000	3035	0.23	0.42	0.29
2000	2035	0.36	0.39	0.37
3000	3035	0.47	0.54	0.50

The results figured in Tables 3 and 4 clearly show the performance of our developed system. They show that the results obtained from the character-based approach indicate a significant increase in recognition performance for the three-test corpus. The best result is obtained by the third corpus with an F-measurement rate of 0.80.

Unlike the word-based approach, the recognition performance remains "moderately acceptable". This difference is mainly due to the influence of word segmentation cross-ambiguity, such as word segmentation "□ □ □ □ □"gives a word by word-based approach, and two words ("□ □ □ □"and "□") by the character-based approach.

We also observe that the rate of word boundary detection performance for both approaches increases, so the training corpus increases.

5. Conclusion

We introduced a new approach to segmentation of words in Tifinagh. Our segmentation method uses two

approaches to extract the word features. Our approach is generic because it does not use any dictionary information. With the limited number of features they have, this segmentation algorithm responds to a wide variability between characters chaining configurations. The first system performance obtained show that it is necessary to consider the segmentation ambiguity effect when learning classifiers.

For the validation of the system, a set of experimental tests were carried out successfully on a corpus of Tifinagh built locally. The result finds deemed system performance.

References

[1] A. Boukous, *Phonologie de l'Amazighe*. Institut royal de la culture amazighe, 2009.

[2] A. Bounfour, *Terminologie grammaticale berbère (amazighe)*. Harmattan, 2009.

[3] L.-X. Tang, S. Geva, A. Trotman, and Y. Xu, 'A boundary-oriented Chinese segmentation method using n-gram mutual information', in *Proceedings of CIPS-SIGHAN Joint Conference on Chinese Language Processing*, 2010, pp. 234–239.

[4] H.-P. Zhang, L. Qun, C. Xue-Qi, Z. Hao, and I. Hong-Kui, 'Chinese lexical Analysis Using Hierarchical Hidden Markov Model', *Proceedings of the Second SIGHAN Workshop on Chinese Language Processing*, pp. 63–70, 2003.

[5] N. Xue and others, 'Chinese word segmentation as character tagging', *Computational Linguistics and Chinese Language Processing*, vol. 8, no. 1, pp. 29–48, 2003.

[6] P. Cairns, R. Shillcock, N. Chater, and J. Levy, 'Bootstrapping word boundaries: A bottom-up corpus-based approach to speech segmentation', *Cognitive Psychology*, vol. 33, no. 2, pp. 111–153, 1997.

[7] C. P. Papageorgiou, 'Japanese word segmentation by hidden Markov model', in *Proceedings of the workshop on Human Language Technology*, 1994, pp. 283–288.

[8] F. Peng, F. Feng, and A. McCallum, 'Chinese segmentation and new word detection using conditional random fields', in *Proceedings of the 20th international conference on Computational Linguistics*, 2004, p. 562.

[9] W. Sun, 'Word-based and character-based word segmentation models: comparison and combination', in *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, 2010, pp. 1211–1219.

[10] K.-J. Chen and S.-H. Liu, 'Word identification for Mandarin Chinese sentences', in *Proceedings of the 14th conference on Computational linguistics-Volume 1*, 1992, pp. 101–107.

[11] K. Wang, C. Zong, and K.-Y. Su, 'A character-based joint model for Chinese word segmentation', in *Proceedings of the 23rd International Conference on Computational Linguistics*, 2010, pp. 1173–1181.

[12] C. F. Hockett, *A Course in Modern Linguistics*. Macmillan, 1975.

[13] S. Li and C.-R. Huang, 'Word Boundary Decision with CRF for Chinese Word Segmentation.', in *PACLIC*, 2009, pp. 726–732.

[14] J. Darroch and D. Ratcliff, 'Generalized Iterative Scaling for Log-Linear Models', *The Annals of Mathematical Statics*, vol. 43, no. 5, pp. 1470–1480, Oct. 1972.

[15] D. M. Powers, 'Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation', 2011.

[16] R. M. Reese, *Natural Language Processing with Java*. Packt Publishing Ltd, 2015.