USTHB University Algiers
Electronics and Computer engineering Faculty (FEI)
Signal Processing Laboratory
www.usthb.dz

Robust Language Identification of Noisy Texts: -Proposal of Hybrid Methods-

Keywords

Natural Language Processing; Text categorization; Automatic Language Identification; Noisy Text; Hybrid Approach.

Mr K. ABAINIA, **Dr S. OUAMOUR,** Prof H. SAYOUD

OUTLINE

Problem

- Growth of the database sizes leads to an expensiveness in access to information and extracting knowledge.
- Knowledge extraction requires NLP/ Computational Linguistics.
- All approaches need to know the input language in advance.
- Language identification = process of identifying automatically the language of a given text.

Difficulty

Main difficulty = noisy texts (forum texts);

- **Citations in other languages;**
- **Typing errors:** sometimes people do not revise their messages before posting them in the forum;
- **Urls:** may contain links to web sites, videos, images,…etc;
- **Tags:** texts may contain a tag of persons, events or different objects.
- **Abbreviations:** some people prefer to abbreviate nouns and use the SMS writing style.
- **Unaccented characters:** in the accented language (like French, Italian…etc) some people forget or do not accentuate the accented characters (e.g. “é” in French is written “e” without accent).
- **Insignificant characters:** texts may contain some insignificant characters to explain sentiments (like emotions).

OUTLINE

Language Corpus

✓ First Dataset DLI-32

- Collection of forum texts in 10 languages;
- 100 texts encoded with UTF-8 encoding;

✓ Second Dataset DLI-32

- Collection of forum texts in 32 languages;
- 320 texts encoded with UTF-8 encoding;
- unbalanced in term of text sizes (number of words)
- Several closed languages (e.g. Malay/Indonesian and Norwegian/Danish)

II. CORPUS

✓ First Dataset

- Concerned languages are:

1. French,
2. English,
3. Arabic,
4. Russian,
5. German,
6. Italian,
7. Greek,
8. Spanish,
9. Persian,
10. Chinese.

✓ Second Dataset

- Concerned languages are:

1. French,
2. English,
3. Arabic,
4. Russian,
5. German,
6. Italian,
7. Greek,
8. Spanish,
9. Persian,
10. Chinese,
11. Turkish,
12. Finnish,
13. Hebrew,
14. Portuguese,
15. Roman,
16. Polish,
17. Hungarian,
18. Dutch,
19. Irish,
20. Swedish,
21. Latin,
22. Icelandic,
23. Hindi,
24. Czech,
25. Malaysian,
26. Bulgarian,
27. Norwegian,
28. Albanian,
29. Urdu,
30. Thai,
31. Indonesian,
32. Danish.

OUTLINE

Three Basic Language Identification Methods

✓ **First Method**

CBA (Character Based identification Algorithm)

✓ **Second Method**

WBA (Word Based identification Algorithm)

✓ **Third Method**

SCA (Special Character based identification Algorithm)

1. CBA

It is based on the computation of the character frequencies.

However, it has 2 drawbacks :

Problem of confusion in case of :

1. languages having the same character set (e.g. Indonesian and Malay…etc);
2. Inclusion of language characters in others (e.g. Arabic and Persian).

CBA – Solution

Follow an order to classify the languages.

CBA – Algorithm

- Load the language characters;
- Read the text file (UTF-8 encoding);
- Remove insignificant characters
(i.e. 1 2 3 4 5 6 7 8 9 0 “) (, . ; ! ? [ ] = : _ * { } # + @)
- Remove multiple word separators;
(i.e. white space, line feed and carriage return)
- Compute the sum of char frequencies :

$$Sum = \sum_{j=1} freq_{ij}$$

where i is the ith language and j is the jth character.

CBA – Algorithm

If the highest sum = Arabic sum then return Arabic language
Else pass to the next test:

1. Persian
2. Urdu
3. Bulgarian
4. Russian
5. French
6. Italian
7. Irish
8. Spanish
9. Portuguese
10. Albanian
11. Czech
12. Finnish
13. Hungarian
14. Swedish
15. German
16. Norwegian
17. Danish
18. Icelandic
19. Turkish
20. English
21. Dutch
22. Indonesian
23. Malay
24. Latin
25. Roman
26. Polish

2. WBA – Algorithm

Similar to CBA algorithm;

Instead using language characters we use common words.

1. Words collected over the web (works of other researchers).

2. Words obtained from a training process.

It has 2 additional steps:

1. Separate contracted words (e.g. l'eau becomes "l" and "eau");

2. Extract words from the text using word delimiters.

2. WBA – Algorithm

- Load the language common words;
- Read the text file (UTF-8 encoding);
- Remove insignificant characters
- Separate contracted words (e.g. l'eau becomes "l" and "eau") :

replace the slash, subtraction sign and apostrophe with white space.

- Extract words from the text using word delimiters.
- Compute the sum of common word frequencies :

$$Sum = \sum_{j=1} freq_{ij}$$

where i is the ith language and j is the jth common word.

- Return the class that has the highest **Sum**.

WBA – Drawbacks

Cannot extract Chinese words from the text because they are not spaced by white spaces like other languages.

Solution

Hybrid with Character Based Identification.

3. SCA – Description

This algorithm uses the special characters characterizing each language.
It is decomposed into two classification steps:

- The **first** step consists in classifying the language into a global class of languages.

Class 1: [Chinese]

Class 2: [Greek]

Class 3: [Arabic, Persian, Urdu]

Class 4: [Russian, Bulgarian]

Class 5: [English, French, Italian, Spanish, Portuguese, German, Turkish, Finnish, Roman, Polish, Hungarian, Dutch, Irish, Swedish, Latin, Icelandic, Czech, Malaysian, Albanian, Norwegian, Indonesian, Danish]

Class 6: [Hebrew]

Class 7: [Hindi]

Class 8: [Thai]

SCA – Description

- The **second** classification step consists in determining exactly what language the text belongs to:

✓ Each language is defined by a set of special characters.

✓ If the language is not characterized by any special character, we apply CBA algorithm to the languages of this class .

SCA – Algorithm

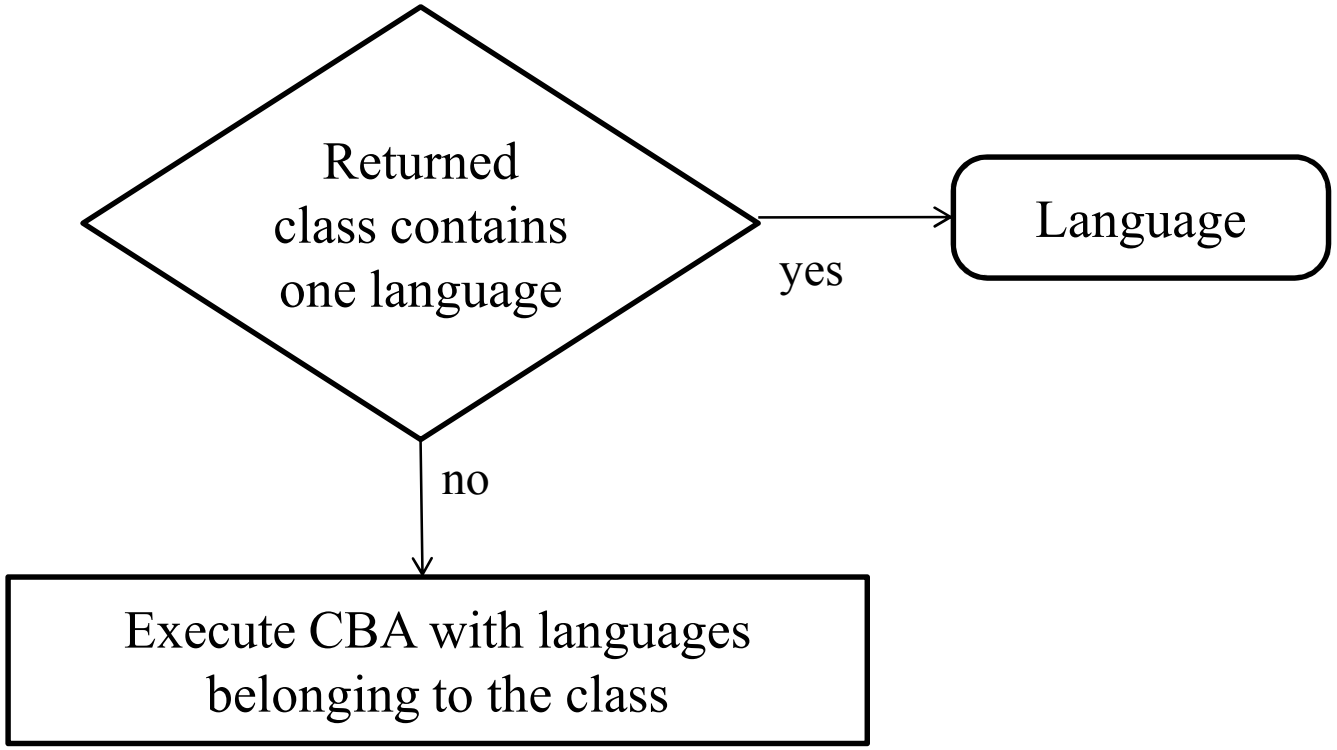
- Load the class languages, class characters and language characters
- Read the text file (UTF-8 encoding)
- Remove insignificant characters
- Remove multiple word separators
- Compute the sum of char frequencies

$$Sum = \sum_{j=1} freq_{ij}$$

where i is the ith language and j is the jth common word.

- Return the class that has the highest **Sum**.

SCA – Algorithm



SCA - Drawbacks

Problem of confusion in case of :

1. languages having the same character set (i.e. Indonesian and Malay…etc);

2. Inclusion of language characters in others (i.e. Arabic and Persian).

OUTLINE

Hybrid Methods

HA1

Sequential Combination

(CBA/ WBA)

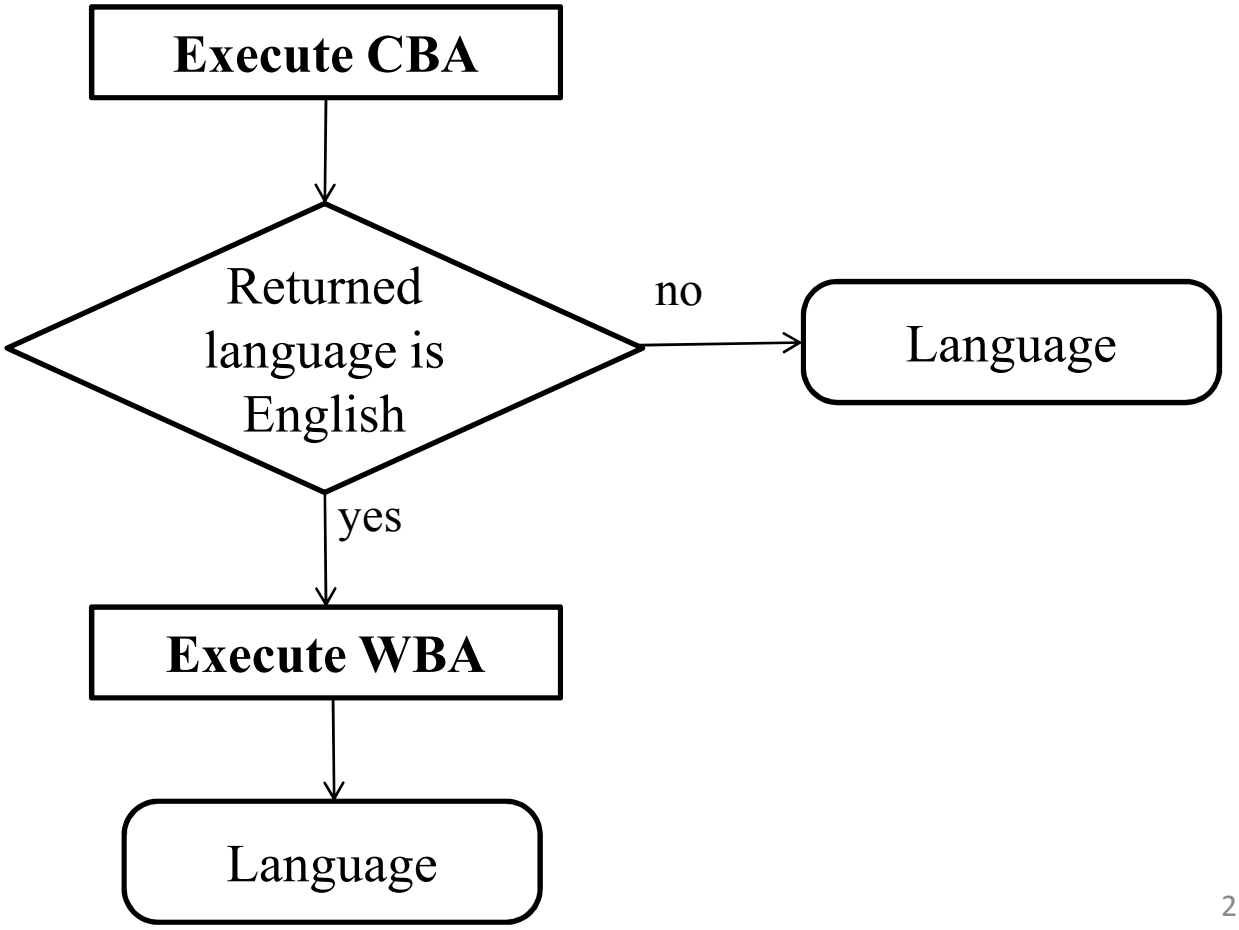
HA2

Parallel Combination

(CBA/ WBA)

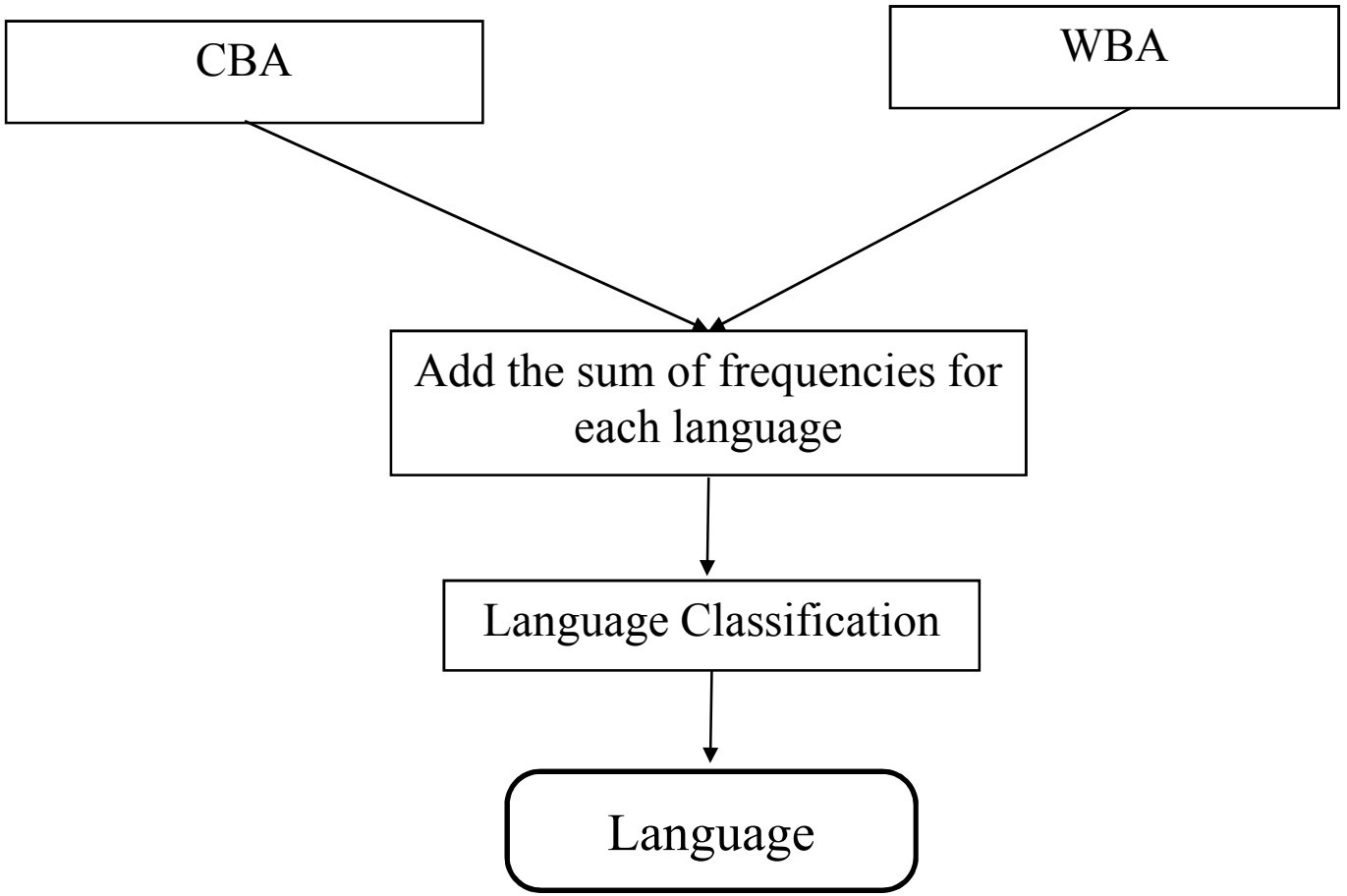
Hybrid Method - HA1

HA1 = **Sequential** combination between CBA and WBA



Hybrid Method – HA2

HA2 = **Parallel** combination between CBA and WBA



Hybrid Method – HA2

- Load the language characters and common words
- Read the text file (UTF-8 encoding)
- Remove insignificant characters
- Separate contracted words
- Remove multiple word separators
- Compute the sum of char frequencies for each language

$$SumC_i = \sum_{j=1} char\_freq_{ij}$$

- Compute the sum of word frequencies for each language

$$SumW_i = \sum_{j=1} word\_freq_{ij}$$

- Add the two previous sum for each language

$$Sum_i = SumC_i + SumW_i$$

- Follow the same order of tests of the CBA

OUTLINE

V. RESULTS

Table 1: Comparative scores of language identification with 10 languages.

	CBA	WBA *test 1*	WBA *test 2*	SCA	HA1 *test 1*	HA1 *test 2*	HA2 *test 1*	HA2 *test 2*
Identification score in %	100	90	90	100	100	100	100	100

- The WBA algorithm uses white space and line feed to extract words from the text, which causes errors in case of Chinese language.
- The identification score of WBA is 90%.
- The identification score of CBA, SCA and Hybrid algorithms is 100%.

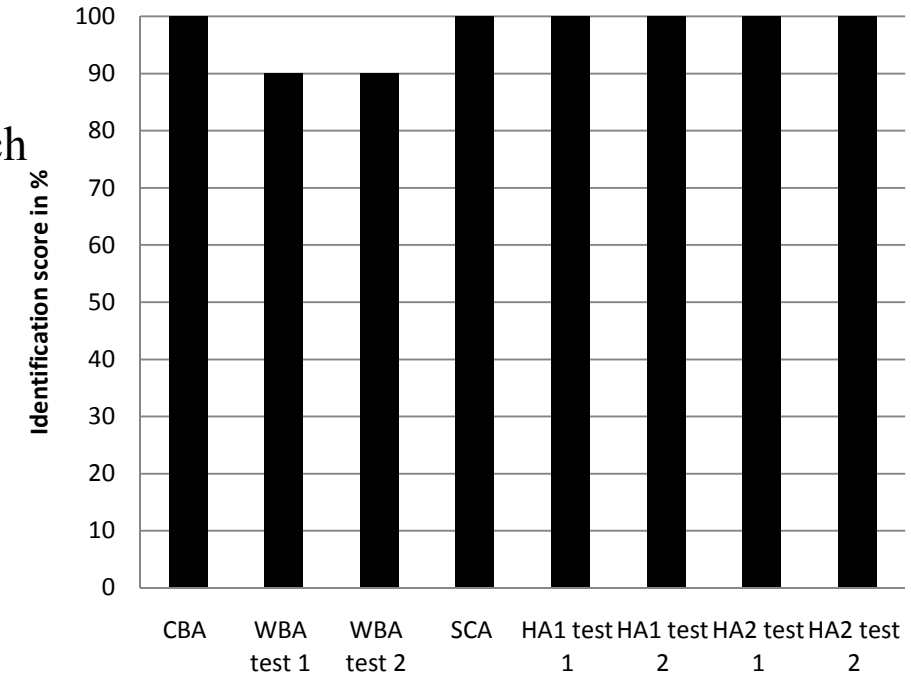
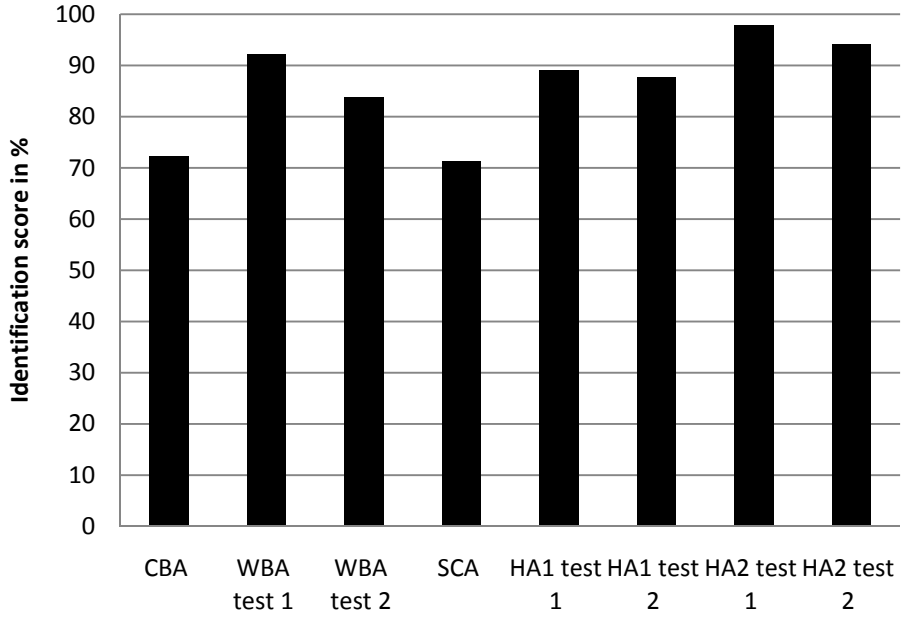


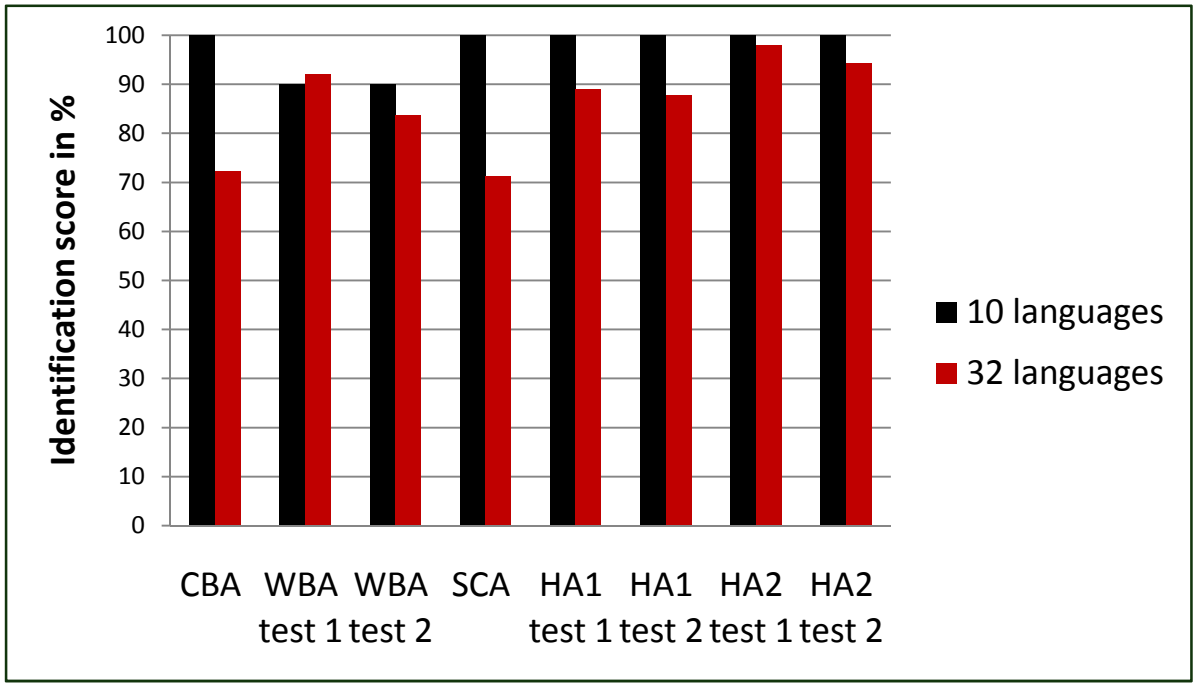
Table 2: Comparative scores of language identification with 32 languages.

	CBA	WBA *test 1*	WBA *test 2*	SCA	HA1 *test 1*	HA1 *test 2*	HA2 *test 1*	HA2 *test 2*
Identification score in %	72.40	92.19	83.85	71.35	89.06	87.85	97.92	94.27

• A difficulty in some languages, which is Due to the inclusion of some characters within other different languages, like Arabic with Persian and Urdu.



V. RESULTS



- Identification scores are reduced for all algorithms, which means that the number of languages impacts directly the accuracy.
- The accuracy of the WBA algorithm is slightly reduced compared with the previous results (10 languages).
- The accuracy of CBA and SCA algorithms are highly reduced on the second dataset (many languages sharing a set of characters).
- Test 1 is better than test 2.
- Parallel fusion HA2 gives the best results.

Table 3: Comparative performances of the hybrid methods with some universal tools: on 1st subset of dataset (10 languages).

	Google Translator	Microsoft Word	HA1 *test 1*	HA2 *test 1*
Identification score in %	98	90	**100**	**100**

- The two methods HA1 and HA2 (with test 1) provide high performances and seem to be better than Google Translator and Microsoft Word.

- We notice that Microsoft Word did not manage to recognize several documents, for instance; Persian texts were recognized as Arabic texts.

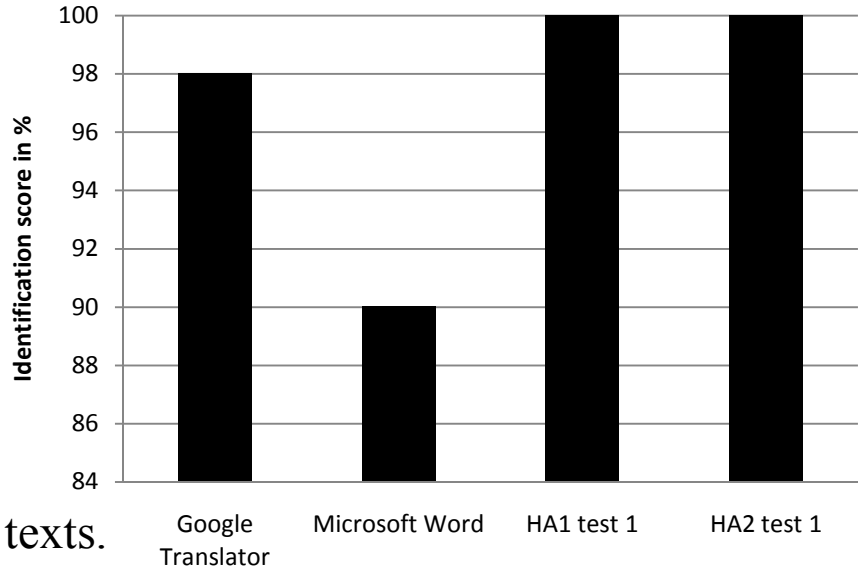
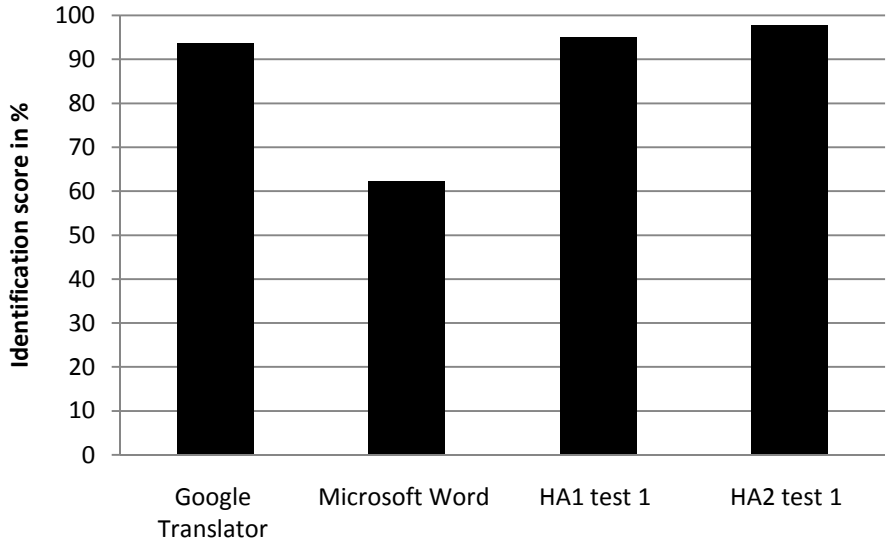


Table 4: Comparative performances of the hybrid methods with some universal tools: on 2nd subset of dataset without Latin and Urdu.

	Google Translator	**Microsoft Word**	**HA1** *test 1*	**HA2** *test 1*
Identification score in %	93.67	62.22	95	97.78

Same observations as table 3.



OUTLINE

CONCLUSION

- This research work deals with language identification of noisy texts (forum texts).
- Three basic methods and two hybrid methods have been proposed.
- Identification results are quite interesting (100%) on a subset of 10 languages, except for WBA;
- However, on the large subset (32 languages), the different scores are reduced, due to the presence of several closest languages.
- The parallel hybrid approach (HA2), gives an identification score of **100%** with 10 languages and a score of **97.78%** with 32 languages.
- Performances of HA2 are better than those of HA1, which shows that the parallel fusion is more interesting.
- Results show that the proposed hybrid techniques are more accurate than Google Translator and Microsoft Word.
- In the overall, the proposed approaches seem to be quite interesting and should be used efficiently to recognize the language of noisy texts.

FUTURE WORK

This investigation is a part of a global system of language identification. The second part of this system consists in using character n-gram (bigram, trigram) with some similarity distances and different classifiers.

Thank You

Mr K. ABAINIA : abainia@hotmail.fr

Dr S. OUAMOUR : siham.ouamour@uni.de

Prof H. SAYOUD : halim@sayoud.net