COMPARATIVE STUDY OF ARABIC AND FRENCH STATISTICAL LANGUAGE MODELS

Karima Meftouh[1], Kamel Smaili[2]
[2]*INRIA-LORIA, Parole team, BP 101 54602 Villers Les, Nancy, France*

Mohamed Tayeb Laskri
[1]*Department of Informatic,Badji Mokhtar University, Annaba, Algeria*

Keywords: Statistical language modeling, Arabic, French, Smoothing technique, n-gram model, Vocabulary, Perplexity, Performance.

Abstract: In this paper, we propose a comparative study of statistical language models of Arabic and French. The objective of this study is to understand how to better model both Arabic and French. Several experiments using different smoothing techniques have been carried out. For French, trigram models are most appropriate whatever the smoothing technique used. For Arabic, the n-gram models of higher order smoothed with Witten Bell method are more efficient. Tests are achieved with comparable corpora and vocabularies in terms of size.

1 INTRODUCTION

Statistical techniques have been widely used in automatic speech recognition and machine translation over the last two decades (Kim and Khudanpur, 2003). Most of the success, therefore, has been witnessed in the so called "resource rich languages" for instance English and French. More recently there has been an increasing interest in languages such as Arabic.

Arabic has a rich morphology characterized by a high degree of affixation and interspersed vowel patterns and roots in word stems, as shown in section 2. As in other morphologically rich languages, the large number of possible word forms entails problems for robust language model estimation.

In the present work, we investigate a comparative study of Arabic and French n-gram models performances. In our knowledge this kind of study has never been done and we would like to investigate the differences between these two languages over their respective n-gram models.

The n-gram models model natural language using the probabilistic relationship between a word to predict and the $(n-1)$ previous words.

The organization of the paper is as follows. We first give an overview of Arabic and French languages (section 2 and 3). We pursue by a description of n-gram models (section 4) and the used corpora (section 5). We then compare the performances of Arabic models with French ones (section 6) and finally we conclude.

2 AN OVERVIEW OF ARABIC

Arabic, one of the six official languages of the United Nations, is the mother tongue of 300 million people (Egyptian demographic center, 2000). Unlike Latin-based alphabets, the orientation of writing in Arabic is from right to left. The Arabic alphabet consists of 28 letters and can be extended to ninety by additional shapes, marks and vowels. Each letter can appear in up to four different shapes, depending on whether it occurs at the beginning, in the middle, at the end of a word, or alone. Table 1 shows an example of the letter < ف /"f" > in its various forms. Letters are mostly connected and there is no capitalization.

Meftouh K., Smaili K. and Tayeb Laskri M. (2009).
COMPARATIVE STUDY OF ARABIC AND FRENCH STATISTICAL LANGUAGE MODELS.
In *Proceedings of the International Conference on Agents and Artificial Intelligence*, pages 156-160
DOI: 10.5220/0001537501560160

Table1: The letter < ف/"f" > in its various forms.

Isolated	Beginning	Middle	End
ف	فـ	ـفـ	ـف

Arabic is a Semitic language. The grammatical system of Arabic language is based on a root-and-pattern structure and considered as a root-based language with not more than 10000 roots and 900 patterns (Hayder and al., 2005).The root is the bare verb form. It is commonly three or four letters and rarely five. Pattern can be thought of as template adhering to well-known rules.

Arabic words are divided into nouns, verbs and particles. Nouns and verbs are derived from roots by applying templates to the roots to generate stems and then introducing prefixes and suffixes (Darwish, 2002). Table 2 lists some templates (patterns) to generate stems from roots. The examples given below are based on the root درس/ < drs >.

Table 2: Some templates to generate stems from the root درس/ < drs >. C indicate a consonant, A a vowel.

Template	Stem
فعل CCC	درس < drs >/ Study
فاعل CACC	دارس < dArs >/ Student
مفعول mCCwC	مدروس < mdrws >/ Studied

Many instances of prefixes and suffixes correspond to entire words in other languages. In table 3, we present the different components of a single word وكررتها which corresponds to the phrase "and she repeats it".

Table 3: An example of an Arabic word.

French	Arabic	English
et	و	And
répéter	كرر	Repeat
elle	ت	She
la	ها	It

Arabic contains three genders (much like English): masculine, feminine and neuter. It differs from Indo-European languages in that it contains three numbers instead of the common two numbers (singular and plural). The third one is the dual that is used for describing the action of two people.

3 THE FRENCH LANGUAGE

French is a descendant of the Latin language of the Roman Empire, as are languages such as Portuguese, Spanish, Italian, Catalan and Romanian.

The French language is written with a modern variant of the Latin alphabet of 26 letters. French word order is Subject Verb Object, except when the object is a pronoun, in which case the word order is Subject Object Verb.

French is today spoken around the world by 72 to 160 million people as a native language, and by about 280 to 500 million people as a second or third language (Wikipedia, 2008).

French is mostly a second language in Africa. In Maghreb, it is an administrative language and commonly used though not on an official basis in the Maghreb states, Mauritania, Algeria, Morocco and Tunisia.

In Algeria, French is still the most widely studied foreign language, widely spoken and also widely used in media and commerce.

4 N-GRAM MODELS

The goal of a language model is to determine the probability of a word sequence $w_1^n, P(w_1^n)$. This probability is decomposed as follows:

$$P(w_1 \ldots w_n) = \prod_{i=1}^{n} P(w_i|h) \quad (1)$$

The most widely-used language models are n-gram models (Stanley and Goodman, 1998). In n-gram language models, we condition the probability of a word w_i on the identity of the last $(n-1)$ words w_{i+1-n}^{i-1}.

$$P(w_i / w_1^{i-1}) = P(w_i / w_{i+1-n}^{i-1}) \quad (2)$$

The choice of n is based on a trade-off between detail and reliability, and will be dependent on the available quantity of training data (Stanley and Goodman, 1998).

5 DATA DESCRIPTION

Currently, the availability of Arabic corpora is somewhat limited. This is due to the relative recent interest for Arabic applications.

For our experiments, the corpora used for Arabic are extracted from the *CAC corpus* compiled by Latifa

Al-Sulaiti within her thesis framework (Al-Sulaiti, 2004). Texts were collected from three main sources: magazines, newspapers and web sites.
For French, the models were trained on corpora extracted from *Le Monde* French newspaper.

We decide to use corpora of identical sizes so that the results could be comparable. Therefore, each training corpus contains 580K words. For the test, each one is made of 33K words.

6 EXPERIMENTAL RESULTS

A number of Arabic and French n-gram language models are computed in order to study their pertinence for these languages. Several smoothing techniques are tested in order to find out the best model: Good-Turing (Katz, 1987), Witten-Bell (Witten and Bell, 1991) and linear (Ney and al., 1994). The vocabulary consists of the most frequent 3000 words.

Statistical language models are usually evaluated using the so called perplexity (P). It can be seen as the average size of the word set over which a word recognised by the system is chosen, and so the lower its value is the better is the model (Saraswathi and Geetha, 2007). The results obtained by using the computed models are listed in table 4 and 5.

Let us notice that the French models are definitely more powerful than those of Arabic. More exactly, Arabic language seems to be more perplex. This can be mainly explained by the fact that Arabic texts are rarely diacritized. Diacritics are short strokes placed above or below the preceding consonant. They indicate short vowels and other pronunciation phenomena, like consonant doubling (Vergyri, 2004). The absence of this information leads to many identical looking word forms (e.g. the form ktb **كتب** (write) can correspond to **كَتَب** *kataba,* **كُتُب** *kutub*, …) in a large variety of contexts, which decreases predictability in the language model.

In addition, Arabic has a rich and productive morphology which leads to a large number of probable word forms. This increases the out of vocabulary rate (37.55%) and prevents the robust estimation of language model probabilities.

Let us also notice that for French, trigram models are the most appropriate whatever the smoothing technique used. For Arabic, it seems that n-gram models of higher order could be more efficient. This observation is confirmed by the values given in Table 6.

Table 4: performance of Arabic n-gram models in terms of perplexity (P) and entropy (E).

N	Good Turing		Witten Bell		Linear	
	P	E	P	E	P	E
2	326.14	8.35	310.17	8.28	346.68	8.44
3	265.03	8.05	240.41	7.91	292.07	8.19
4	233.97	7.87	204.44	7.68	261.84	8.03

Table 5: performance of French n-gram models in terms of perplexity (P) and entropy (E).

N	Good Turing		Witten Bell		Linear	
	P	E	P	E	P	E
2	157.84	7.30	154.89	7.28	170.35	7.41
3	141.02	7.14	140.35	7.13	170.26	7.41
4	144.55	7.18	151.12	7.24	182.50	7.51

Table 6: performance of Arabic higher order n-gram models in terms of perplexity (P) and entropy (E).

N	Good Turing		Witten Bell		Linear	
	P	E	P	E	P	E
5	229.29	7.84	184.95	7.53	258.07	8.01
6	238.75	7.90	176.99	7.47	279.56	8.13
7	254.96	7.99	173.73	7.44	323.50	8.34
8	269.06	8.07	172.47	7.43	415.93	8.70
9	279.07	8.12	172.35	7.43	Inf	inf

Table 7: performance of French higher order n-gram models in terms of perplexity (P) and entropy (E).

N	Good Turing		Witten Bell		Linear	
	P	E	P	E	P	E
5	148.31	7.21	159.48	7.32	191.59	7.58
6	151.02	7.24	164.30	7.36	198.45	7.63
7	152.04	7.25	166.05	7.38	inf.	Inf
8	152.37	7.25	166.67	7.38		
9	152.65	7.25	166.87	7.38		

True enough the 5-gram models are the most efficient for Arabic, except with Witten Bell discounting method. For French, trigrams remain the most appropriate (see Table 7).

In order to summarize these results, we present them with the curve of Figure 1.

In general models, smoothed with Good Turing or Witten Bell, are the most appropriate. The linear smoothing technique provides infinite values from n= 9 for Arabic and n= 7 for French.

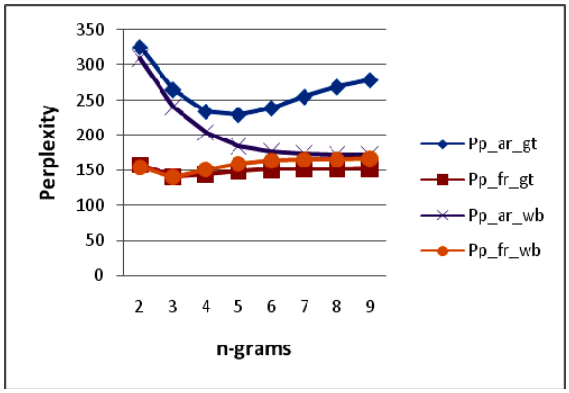

Figure 1: comparison of perplexities obtained for Arabic (ar) and French (fr) n-gram language models with Good Turing (gt) and Witten Bell (wb) smoothing techniques.

First, it should be noted that the variation in terms of perplexity is very important from an Arabic model to another. By against for French, the change is very small.

Good Turing technique gives the best perplexity values for French (Pp_fr_gt). Arabic models smoothed with Witten Bell are the most efficient (Pp_ar_wb). The perplexity stop decreasing only with this smoothing technique and from n = 8. Note also that with this value of n and only with Witten Bell smoothing, models performances for both languages are close.

6.1 Influence of the Size Vocabulary

To strengthen these results, we have carried out various experiments by varying the size of the training vocabulary. Figure 2 gives the perplexity values of the most efficient models of Arabic and French.

Once again trigram models with Good Turing smoothing (Pp_3gram_fr_gt) are most effective for French whatever the vocabulary size.

For Arabic, the n-gram models smoothed with Witten bell which are the most effective whatever the size of the vocabulary.

It is worth noting also that the change in the size of the vocabulary has a direct influence on the number of words Out Of Vocabulary (OOV) (see figure 3). But this increase in vocabulary size leads to a significant degradation of performances of language models (figure 2) especially Arabic ones.

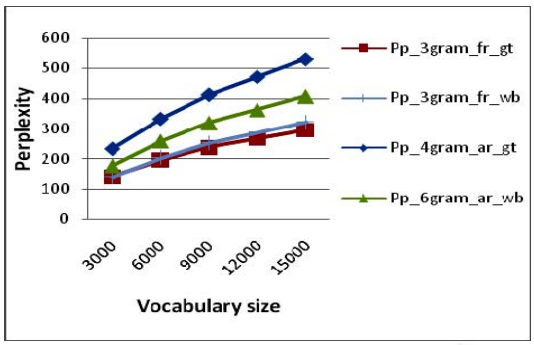

Figure 2: evolution of perplexity of Arabic (ar) and French (fr) n-gram models depending on the size of the vocabulary.

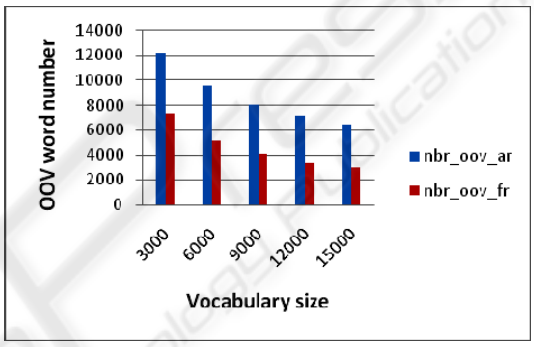

Figure 3: variation in the number of words OOV for Arabic (nbr_oov_ar) and French (nbr_oov_fr) depending on the size of the training vocabulary.

7 CONCLUSIONS

In this paper, we have investigated a comparative study of Arabic and French n-gram language models. Thus we have carried out various experiments using different smoothing techniques. For French, trigram models are most appropriate whatever the smoothing technique used. For Arabic, the n-gram models of higher order smoothed with Witten Bell method are more efficient. As in other morphologically rich languages, the large number of possible word forms entails problems for robust language model estimation. It is therefore preferable, for Arabic, to use morpheme like units instead of whole word forms as language modeling units (Meftouh and al., 2008).

REFERENCES

Meftouh, K., Smaili, K., Laskri, M.T. 2008. Arabic statistical modeling. In *JADT'08, 9e Journées*

internationales d'Analyse statistique des Données Textuelles. 12-14 Mars, Lyon, France.

Wikipedia, 2008. French language. http://en.wikipedia.org/wiki/french_language

Saraswathi, S., Geetha, T.V. 2007. Comparison of performance of enhanced morpheme-based language models with different word-based language models for improving the performance of Tamil speech recognition system. *ACM Trans. Asian language. Inform. Process.* 6, 3, Article 9.

Hayder K. Al Ameed, Shaikha O. Al Ketbi and al. 2005. Arabic light stemmer: A new enhanced approach. In *IIT'05, the Second International Conference on Innovations in Information Technology.*

Vergyri, D., Kirchhoff, K. 2004. Automatic Diacritization of Arabic for Acoustic Modeling in Speech Recognition, *COLING Workshop on Arabic-script Based Languages*, Geneva, Switzerland.

Al-Sulaiti, L. 2004. Designing and developing a corpus of contemporary Arabic. PhD thesis.

Kim, W., Khudanpur, S. 2003. Cross-Lingual lexical triggers in statistical language modelling. *Theoretical Issues In Natural Language Processing archive* Proceedings of the 2003 conference on Empirical methods in natural language processing, Volume 10

Darwish, K. 2002. Building a shallow Arabic morphological analyser in one day. *In Proceeding of the ACL workshop on computational approaches to Semitic languages.*

Egyptian Demographic center. 2000. http://www.frcu.eun.eg/www/homepage/cdc/cdc.htm

Stanley F. Chen, Goodman J. 1998. An empirical study of smoothing techniques for language modelling. *Technical report TR-10-98, Computer science group*, Harvard University, Cambridge, Massachusetts.

Ney H., Essen U. and Kneser R. 1994. On structuring probabilistic dependencies in stochastic language modeling. *Computer Speech and Language*, 8(1):1-38.

Witten I.T. and Bell T.C. 1991. The Zero-frequency problem: Estimating the probabilities of novel events in adaptive text compression. *IEEE Transactions on Information Theory*, 37(4):1085-1094.

Katz S.M. 1987. Estimation of probabilities from sparse data for the language model component of a speech recognizer. *IEEE Transactions on Acoustics, Speech and Signal processing*, 35(3): 400-401.