Statistical Modeling of Arabic Language

Karima MEFTOUH
Department of Informatics, Badji Mokhtar University
BP 12, Annaba, 23000, Algeria
karima.meftouh@univ-annaba.org

May 8, 2009

1 Join MEDAR

Human Language Technologies enable humans to communicate with computers and to use computers and the internet in a more natural way and in their own language, i.e. to participate in the information society in a totally natural way. Native speakers of languages that are not well served by language technology suffer from less access to information, and from less efficient tools, and higher productions costs for documents and translation [1]. Since 2001, Arabic language becomes a priority for several researchers through the world. this interest is probably due to a massive need of computer tools necessary to deal with the huge amount of Arabic data electronically available and, which is dramatically increasing daily. Some tools already exist for Arabic, but there is a long way to go before the Arabic tools reach the level which exists for the so called "resource rich languages" as English, French or Chinese.
this lack of perfection, in my opinion, is certainly due to lack of cooperation in the research community working on the Arabic language. So what we need is to consolidate efforts by sharing ideas, knowledge and resources to produce better tools.
precisely MEDAR encourages and facilitates the exchange and sharing of resources and knowledge:As much as possible, MEDAR partners work from present open-source technology and build on it. This helps the Arabic HLT community in terms of sharing knowledge and resources, helping to minimize costs. MEDAR also updates partners on state-of-the-art Arabic HLT so that duplication of work is minimized. this update is also provided by the organization of meetings between researchers. this contact is certainly very

rewarding especially, as was the case for this second meeting (MEDAR), if it includes different generations of researchers.I was personally very pleased to finally associate faces to the names of people we encounter certainly making a simple request "Arabic NLP". I was also honored by the little discussion I had with Mr. Josef Dichy who is probably one of the pioneers of Arabic language processing. I was also very pleased to meet Shereen Khoja, Nizar Habash, Mona Diab, Nasreedine Semmar, Kareem Darwish, ... and the list is too long.
To this list I'll add kindly Dorte Haltrup Hensen and Helene Mazo. Thank you greatly, you are really humanly very skilled.

2 Statistical modeling of Arabic

Within the Laboratory of Research in Informatics (LRI), we work since years on the Arabic language specially Arabic statistical modeling. Arabic has a rich morphology characterized by a high degree of affixation, interspersed vowel patterns and roots in word stems. As in other morphologically rich languages, the large number of possible word forms entails problems for robust language model estimation.
A statistical language model is used to build up sequence of words, classes or phrases which are considered as linguistically valid in accordance to a corpus without any use of external knowledge. To each linguistic event a probability is estimated to indicate its likelihood. An event is any potential succession of words.
The common model used in the literature is the well-known n-grams. A word is estimated in accordance to the $n-1$ previous words. To be efficient this model needs a huge amount of data to train all the required parameters. In this report, I will present an abstract of our last publication [2]. this work is part of my doctoral thesis directed by Prof. Laskri Mohamed Al-Tayeb and Prof. Kamel Smaili. It's 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.
First I give an overview of French language. I pursue by a description of n-gram models and the used corpora. Then the results of the comparison. at the end I will present my two supervisors Pr Laskri and Pr. Smaili.

2.1 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 [5]. 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.

2.2 N-gram Models

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

$$P(w_1^k) = \prod_{i=1}^{k} P(w_i / w_1^{i-1}) \tag{1}$$

The most widely used language models are n-gram models [6]. In n-gram language models, we condition the probability of a word on the identity of the last $n-1$ words.

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

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 [6]. Because of the sparseness data, in statistical language models, parameters have to be smoothed. The objective is to fine-tune probabilities to overcome the problem of missing data. Several methods exist in the literature, we can cite: Good-Turing [7], Witten-Bell [8], linear [9], Kneser-Ney [10].
The quality of a language model is estimated either by entropy or by perplexity. The perplexity is inspired from entropy and is given by the following formula:

$$PP = \frac{1}{\sqrt[n]{P(w_1, ..., w_n)}} \tag{3}$$

2.3 Data description

The development of corpora is an important research resource since Arabic needs some solid investigation based on large amounts of authentic material. At present, corpus-based research in Arabic lags far behind that of modern European languages. As far as we know, most studies on Arabic up to now have been based on rather limited data. For our experiments, the corpora used for Arabic are extracted from the CAC corpus compiled by Latifa Al-Sulaiti within her thesis framework [11]. 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.

2.4 Experimental results

Several Arabic and French n-gram language models are computed in order to study their pertinence for these languages. Few smoothing techniques are tested in order to find out the best model. The results obtained are listed in table 1 and table 2.

Table 1: performance of Arabic n-gram models.

n	$Good-turing$		$Witten-bell$		$Linear$	
	$Perplexity$	$Entropy$	$Perplexity$	$Entropy$	$Perplexity$	$Entropy$
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 2: performance of French n-gram models.

n	$Good-turing$		$Witten-bell$		$Linear$	
	$Perplexity$	$Entropy$	$Perplexity$	$Entropy$	$Perplexity$	$Entropy$
2	157.4	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

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 [12]. The absence of this information leads to many identical looking word forms (e.g. the form ktb كتب (write) can correspond to كَتَبَ (wrote), كُتُب (books),) in a large variety of contexts, which decreases predictability in the language model. In addition, Arabic has a rich and productive morphology that 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 3.
True enough the 5-gram models are more efficient for Arabic whatever the smoothing technique. Specially, Witten Bell discounting method seems to be the most powerful for this language. These results are not confirmed for French (Table 4).

Table 3: Performance of Arabic higher order n-gram models in terms of perplexity and entropy.

n	$Good-turing$		$Witten-bell$		$Linear$	
	$Perplexity$	$Entropy$	$Perplexity$	$Entropy$	$Perplexity$	$Entropy$
5	229.29	7.84	184.95	7.53	258.07	8.01
6	23.75	7.95	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	415.3	8.70
9	279.07	8.12	172.35	7.43	inf	inf

In order to summarize these results, we illustrate 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.

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

Table 4: Performance of French higher order n-gram models in terms of perplexity and entropy.

n	$Good-turing$		$Witten-bell$		$Linear$	
	$Perplexity$	$Entropy$	$Perplexity$	$Entropy$	$Perplexity$	$Entropy$
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		

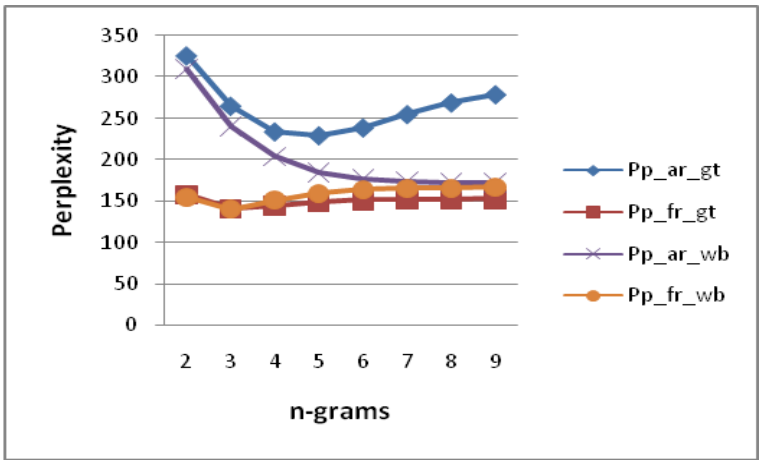

Figure 1: Comparison of perplexities obtained for Arabic (ar) and French (fr) n-gram language models with Good Turing (gt) and Witten Bell (wb) smoothing techniques.

values for French (Pp-fr-gt). Arabic models smoothed with Witten-Bell are the most efficient (Pp-ar-wb). The perplexity stops 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.

2.4.1 Influence of the vocabulary size

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)

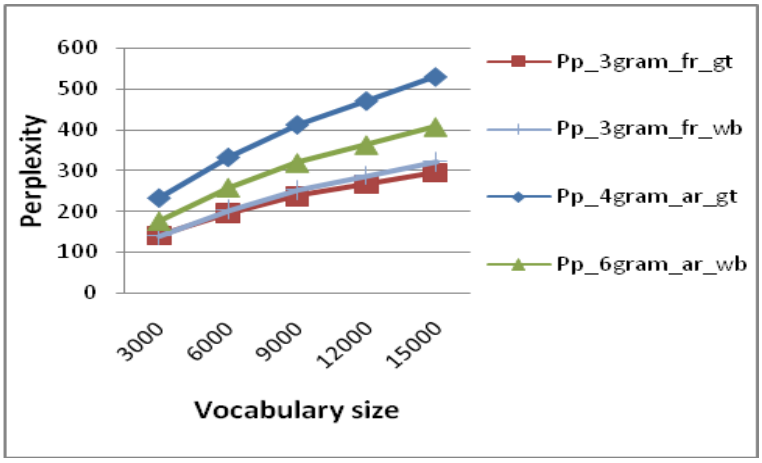

Figure 2: Evolution of perplexity of Arabic (ar) and French (fr) n-gram models depending on the size of the vocabulary.

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.

2.5 Conclusion

A comparative study of Arabic and French n-gram language models was investigated. Thus various experiments have been carried out 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 [13].

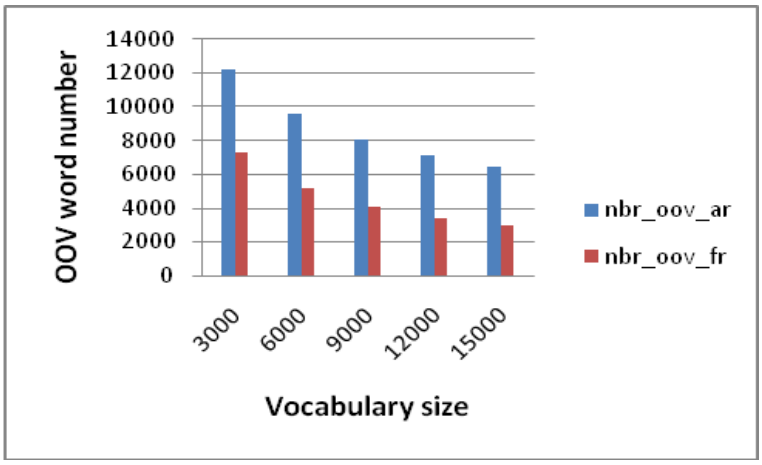

Figure 3: Variation in the number of words OOV for Arabic (nbr-oov-ar) and French (nbr-oov-fr) depending on the size of the training vocabulary.

3 My supervisors

In this section, I will present information about my thesis'directors: Pr. Mohamed Tayeb Laskri and Pr. Kamel Smaili.

1. Mohamed Tayeb Laskri was born at Annaba (Algeria) in 1958. He obtained a 3rd cycle doctorate on computer science (France, 1987) He obtained his PhD degree from Annaba University (Algeria, 1995). Pr. Laskri is head of the research group in artificial intelligence within LRI laboratory. His current research takes the reasoning in artificial intelligence like field of application privileged in particular in image processing, multi-agents systems, engineering of the Human-machine interfaces and automatic natural language processing.

2. Kamel Smaili Professor at university of Nancy since 2002, obtained a PHD from the same university on 1991. He is a member of LORIA-France lab. He is a leader of a research group working on statistical language modeling and speech-to-speech translation. He defended an HDR (Habilitation diriger la recherche) on 2001. His research interest since 20 years concerns statistical language modeling for speech recognition and since 2000 he oriented his research to speech-to-speech translation. He proposed several original ideas: retrieving phrases based on class-phrases, purging statistical language models from impossible events, Cache-features language model, multilingual triggers .

. . He participated to several European and French projects concerning speech recognition: COCOS, MULTIWORKS, COST, MIAMM, IVOMOB (RNRT project). He advised more than 9 PHD students and participated to 20 PHD committees through the world. He took part to several program committees: Eurospeech, ICSLP, ICASSP, SIIE, TAIMA, TAL, Computer speech and language, Speech communication . . . He published his research in more than 55 international conferences and journals and in more than 20 francophone conferences and journals.

References

[1] Bente Maegaard, Khalid Choukri, Chafik Mokbel and Mustafa Yaseen: Language Technologie For Arabic. NEMLAR, Center for Sprogteknolo gi, University of Copenhagen, July 2005, http://www.nemlar.org.

[2] K. Meftouh, K. Smaili and M.T. Laskri: Comparative study of Arabic and French statistical language models. Proceedings of the international conference on agents and artificial intelligence ICAART'09, Porto, Portugal, 2009.

[3] Hayder K.Al Ameed, Shaikha O.Al Ketbi et al.: Arabic light stemmer: Anew anhanced approach. in proc. of the Second Internationnal Conference on Innovations in Informations Technology (IIT'05), 2005.

[4] Kareem Darwish: Building a shallow Arabic morphological analyser in one day. In proceedings of the ACL workshop on computational approches to semitic languages, Philadelphia, PA, 2002.

[5] Wikipedia, 2008. French language. *http* : *//en.wikipedia.org/wiki/frenchlanguage.*

[6] Stanley F.Chen and J.Goodman: An empirical study of smoothing techniques for language modeling. Tech. report TR-10-98, Computer science group, Harvard University, Cambridge, Massachusettes, August 1998.

[7] S.M. Katz: 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, 1987.

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

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

[10] R. Kneser and H. Ney: Improved backing-off for M-gram language modeling. Proceedings of the IEEE ICASSP, 1995.

[11] Latifa Al-Sulaiti: Designing and developing a corpus of contemporary Arabic. PHD thesis, University of Leeds, School of Computing, 2004.

[12] D. Vergyri and K. Kirchhoff: Automatic Diacritization of Arabic for Acoustic Modeling in Speech Recognition. Proceedings of the COLING Workshop on Arabic-script Based Languages, Geneva, Switzerland, 2004.

[13] K. Meftouh, K. Smaili and M.T. Laskri: Arabic statistical modeling. Proceedings of 9e Journées internationales d'Analyse statistique des Données Textuelles, 837-844, Lyon, France, 2008.