# Statistical Classification Methods for Arabic News Articles

**Hassan Sawaf**†‡
h.sawaf@aixplain.de
sawaf@cs.rwth-aachen.de

**Jörg Zaplo**†
j.zaplo@aixplain.de

**Hermann Ney**‡
ney@cs.rwth-aachen.de

† AIXPLAIN AG
Monnetstrasse 18
D-52146 Würselen, Germany

‡ Computer Science Department VI
RWTH Aachen
Ahornstrasse 55
D-52056 Aachen, Germany

## Abstract

In this paper, we present experimental results on document clustering and classification achieved on the Arabic NEWSWIRE corpus using statistical methods. Arabic is a highly inflecting language. The methods presented here show to be very robust and reliable without morphological analysis.

## 1 Introduction

Text classification is a fundamental task in document processing, especially because the information flood is getting enormous. Various classification approaches are tested on languages like English, German, French and other European languages.

For many European languages, there are many rule-based and statistical approaches that can be used for all fields of information retrieval and knowledge management. For Arabic, there are only a few approaches, which are able to handle problems with the inflections that do not appear in other languages in that fashion.

In the system introduced here, we study the use of statistical methods for text analysis for Arabic, namely document clustering and text classification.

For any statistical document processing task the morphology of Arabic is a crucial problem. Morphological analyzer are therefore essential when using classification or clustering on word level.

The problem we have to deal with when using full-form words is the sparse data problem. This problem can be reduced by morphological analysis. A possible approach is the use of full morphological analysis based on linguistic knowledge and a complex set of rules, e.g. the system described in [Beesley 96].

Here, we use an alternative approach to cope with the morphological processing. The use of character *n*-grams, i.e. use of sub-word units, is a simplified statistical approach towards a morphological analyzer.

Practical experience shows that the use of pure statistical methods makes a system much more flexible, a working prototype can be built in a very short period of time and the improvement and maintenance of the system is easier.

## 2 Mathematical Approach

The main topic in this paper will be the classification of documents or topic detection. On the one hand we will present a method of classification, on the other hand we will show a method of fully-automatically defining topic clusters so that we can use unlabeled data as input for a text classification system.

## 2.1 Text Classification

Text classification, as presented here, is based on the maximum entropy technique. Maximum entropy modeling shows to be a promising approach for a large variety of tasks, e.g. language modeling [Rosenfeld 94, Martin et al. 97, Peters & Klakow 99], part-of-speech tagging [Ratnaparkhi 96], context free parsing [Ratnaparkhi 98], and text classification [Nigam et al. 99].

Maximum entropy is a general technique for modeling probability distributions. Deriving the right information from the training data makes the obtained probability distribution converge towards the "optimal", theoretical distribution by using constraining features. Features are functions of an example with a real-value. The generalized iterative scaling algorithm (GIS) estimates an optimal parameter set for the required distribution.

In general, the classification problem can be described as follows. We use two types of units, namely full-form words and character trigrams. For a given document with $N$ units $w_1 ... w_N = w_1^N$ we search for a class $c$ that maximizes the posterior probability $p(c \mid w_1^N)$. When using maximum entropy modeling, this probability is estimated by:

$$p_\Lambda(c \mid w_1^N) = \frac{\exp\left(\sum_i \lambda_i \cdot f_i(w_1^N, c)\right)}{Z(w_1^N)} ,$$

where $\Lambda = \{\lambda_i\}$ describes the parameter vector, $f_i(w_1^N, c)$ is a feature function with a real-value, and $Z(w_1^N)$ is a normalizing constant, called the partition function, which is defined as follows:

$$Z(w_1^N) = \sum_{c'} \exp\left(\sum_i \lambda_i \cdot f_i(w_1^N, c')\right) .$$

The features we use for our classification are relative frequencies of units $\tilde{w}$ in a document $w_1^N$:

$$f_{\tilde{w},c'}(w_1^N, c) = \begin{cases} \dfrac{n(\tilde{w}, w_1^N)}{N} & if \quad c = c' ; \\ 0 & otherwise . \end{cases}$$

$n(\tilde{w}, w_1^N)$ is the count of the unit $\tilde{w}$ in the document $w_1^N$, $N$ is the count of the units in document.

It has to be pointed out that the assignment of a document to a class does not need to be unique, due to the fact that in practice a document may belong to more than one topic. Therefore, we use a *N*-best list that contains the best *N* classes for the tested document. The system then determines how much classes have to be selected by the following criterion:

$$p(c \mid w_1^N) > \max\left\{ B, \frac{1}{A} \max_{c'} p(c' \mid w_1^N) \right\} .$$

Both parameters *A* and *B* depend on the task and restrict the result set to the most probable classes for a specific document.

In the present system, we only use continuous sequences of units (words or character *n*-grams, where two adjacent *n*-grams have characters in common, if $n \geq 2$) as units $w_j$. Preliminary experiments show that, especially for the classification task, long-range and parsing features should also be considered so that long-range dependencies and embedded structures can help the classification decision, as in language modeling [Martin et al. 99, Sawaf et al. 00].

## 2.2 Document Clustering

For document clustering we use a criterion based on the mutual information criterion as described in [Melamed 97] and other publications. The criterion is defined as follows:

$$MI(w_1^N, c) := \sum_{i=1}^{N} n(w_n, c) \cdot \log \frac{n(w_i, c) \cdot n}{n(w_i) \cdot n(c)} .$$

Here, $n(w_i, c)$ is the count of a unit $w_i$ in the class $c$, $n(w_n)$ and $n(c)$ the count of a unit $w_i$ and of a class $c$ in the whole corpus, respectively. $n$ denotes the overall count of all units in the corpus.

To extract the most important phrase for a class, we can use the following criterion:

$$\hat{w} = \arg\max_{w_i} \left\{ n(w_i, c) \cdot \log \frac{n(w_i, c) \cdot n}{n(w_i) \cdot n(c)} \right\} .$$

The algorithm for clustering documents in a predefined number of classes is similar to the algorithm introduced in [Kneser & Ney 91], where words are assigned to word classes. The algorithm starts with a random assignment of documents to the classes, and every time an improvement is gained, a document is shifted from one class to another. The algorithm terminates when no further improvement is gained by shifting any of the documents.

With this algorithm, a optimal solution is not necessarily found. But in practice experiments show that the quality of classes of more complex algorithms like hill-climbing or simulated-annealing is not significantly better and thus is not worth the effort.

## 3 Experimental Results

This section provides some preliminary experiments on maximum entropy text classification and mutual information document clustering on an Arabic corpus.

### 3.1 Evaluation criteria

For text classification and clustering, we use standard evaluation criteria if possible, to give a possibility for comparing the methods and tasks with other works.

### 3.1.1 Text Classification

The evaluation of the experiments for text classification is done by using the following quality measures on document level :

$$\text{Precision : } P = \frac{\# correct\ classes\ found}{\# classes\ found} ;$$

$$\text{Recall : } R = \frac{\# correct\ classes\ found}{\# correct\ classes} .$$

Both quality measures in combination define the so called *F*-measure :

$$F = \frac{2 \cdot P \cdot R}{P + R} .$$

### 3.1.2 Text Clustering

Document clustering is difficult to evaluate. Therefore, we gave the results to three native speakers to evaluate the experiments. The evaluators gave each cluster one of three possible scores. The three evaluation scores were as follow :

*good cluster* (score: 3): the documents in this cluster have a joint topic, the extracted keywords are of major relevance to the topic;

*moderate quality cluster* (score: 2): in this cluster the documents are of mixed topics, but the extracted keywords define a relevant topic;

*poor cluster* (score: 1): the documents in this cluster have no common topic and the extracted keywords do not form a topic.

The scores for each cluster are summed up and divided by the total number of clusters in the test set.

### 3.2 Corpus

The corpus on which we perform our experiments is the Arabic NEWSWIRE corpus of LDC. This is a news corpus in dialect-free Arabic. The domains in the articles cover politics, economy, culture and sports. We focused our experiments on the volume of the year 1994 without using any preprocessing of the text. The corpus is in SGML format with Unicode UTF-8 encoding.

The NEWSWIRE corpus is voweled only in ambiguous cases. For our experiments, we ignored the diacritics and similar markings.

In Table 1 some corpus statistics on Arabic NEWSWIRE are represented. For comparison reasons, the Reuters-21578 corpus is shown in the same table. As can be seen, the corpus/vocabulary size ratio differs very much. In our context, a word means full-form word.

The Reuters corpus is similar to the Arabic NEWSWIRE corpus in terms of number of documents and running words. the difference in vocabulary size is evident, though.

To measure the difficulty of the corpus for a statistical method for language processing, the bigram and trigram perplexity is also given. The perplexities are measured on both word level and on character level.

As expected, perplexities for both corpora, namely Reuters-21578 and Arabic NEWSWIRE, is very high, both on word as on character level. The bigram perplexity of Arabic NEWSWIRE is about seven times higher than Reuters-21578 corpus, trigram perplexity of NEWSWIRE is even about forty times higher than Reuters. On character level, the perplexities for Arabic are lower by a factor of about thirty compared to the word level. The underlying language models use interpolation with absolute discounting as smoothing method [Sawaf et al. 00].

Table 1: Corpus Characteristics: Comparison Arabic NEWSWIRE 1994 and Reuters-21578

| | NEWSWIRE | Reuters-21578 |
|---|---|---|
| vocabulary size | 230K | 80K |
| number of documents | 33K | 39K |
| number of words | 7M | 6M |
| number of characters | 40M | 40M |
| bi-/trigram word perplexity | 528/495 | 71/12 |
| bi-/trigram character perplexity | 15.3/4.9 | |

## 3.3 Experiments

For the classification task, we use 80% of the corpus for training the statistical models and about 20% for testing, where the selection is done randomly. For the clustering task, we use the full corpus.

### 3.3.1 Clustering Experiments

Table 2 illustrates the preliminary results for the document clustering task. Experiments were performed for several number of clusters.

Some of the results are shown in Table 3. For the definition of the cluster, the extraction of the most relevant phrases is essential, here we present the most relevant words.

Table 2: Clustering Evaluation

| Type of units | Number of cluster | Evaluator | | | Average |
|---|---|---|---|---|---|
| | | A | B | C | |
| Full-form words | 10 | 2.3 | 2.2 | 2.4 | 2.3 |
| | 50 | 2.0 | 2.2 | 2.2 | 2.1 |
| Character trigrams | 10 | 2.2 | 2.1 | 2.2 | 2.2 |
| | 50 | 2.1 | 2.1 | 2.3 | 2.2 |

Table 3: Extracted Keyphrases for Clustering˜into Ten Cluster using MI criterion

| Cluster | Keyphrases (translated) | Keyphrases |
|---|---|---|
| CL1 | Arafat, Israel, in Gaza, Palestinian, government, self liberation, Rabin, foreign | عرفات, اسرائيل, فىغزة, الفلسطينية, الحكم, التحرير الذاتي, رابين, الخارجية |
| CL2 | Adan, Northern Yemen, Salih, Right Party, Kabul, killed, fights | عدن, اليمن الشمالية, صالح, الحزب اليمني, كابول, البيض, قتلوا, المعارك˜ |
| CL3 | in, cup, voted, better, minute, turn, Al-Itihad, decision | في, كأس, المنتخب, المباراة, افضل, الدقيقة, الدور, الاتحاد, الحكمة ˜ |
| CL4 | <NUM>, from, Dollar, in, Iran, Teheran, Arabic countries, million, Iranian | <NUM>, من, دولار, فى, ايران, طهران, الدول العربية, مليون, الايرانية˜ |
| CL5 | United Nation, National Security Council, Iraq, for the Nations, Ghali, Kuwait, Punishment | الامم المتحدة, مجلس الامن, العراق, للامم, غالي, الكويت, العقوبات, وقف |
| CL6 | In, that, from, Russian, Carter, Bosnia, Rwanda, president, forces, Chechnya | فى, ان, من, الروسية, كارتر, البوسنة, رواندا, الرئيس, القوات, الشيشان˜ |
| CL7 | Lebanon, Police, Hisbollah, Hamas, movement, in, killing, mosque, Gaza | لبنان, الشرطة, حزب الله, حماس, حركة, فى, قتل, الحرم, غزة˜ |
| CL8 | <NUM>, and cursed, championship, then, -, stage, soccer | <NUM>, فسب, بطولة, ف, -, المرحلة, كرة القدم˜ |
| CL9 | In, film, Sudan forces, Police, Egypt, Dollar, explosion, kidnapping, French | في, الفيلم, السلطات السودانية, الشرطة, مصر, الدولار, الانفجار, اعتقل, الفرنسية˜ |
| CL10 | To, peace, Jordan, King Hussein, in, Al-Ahil, Jordanian help, Arabic | الى, السلام, الاردن, الملك حسين, في, العاهل, التعاون الاردني, العربي˜ |

As we carried out clustering experiments for both word and character level, both versions of the experiments are presented. In Table 3 the keyphrases we show are the results of the clustering method on word level.

The experiments show the quality of the clustering system. It is interesting to see that clustering on character trigram level perform only as good as on word level, for ten clusters even slightly worse. The extracted keyphrases in Table 3 show that a human transcriber would use a similar assignment.

In general, the less clusters the system have, the more errors occur in the form that a cluster can have more than one topic. In the example in Table 3, in CL2 two different topics are joined, namely politics in Afghanistan and in Northern Yemen. Intuitively this can be explained as follows: in 1994 in both these areas, there was a civil war.

### 3.3.2 Classification Experiments

Table 4 shows text classification accuracy results in terms of average precision, recall and f-measure. For the test, a human transcriber defined 34 classes where a document can also be assigned to more than one class. The system was trained and tested on these predefined classes (A-34). Additionally, two different human transcribers defined a less complex class model consisting of ten distinct classes. Experiments are also illustrated in Table 4 (B-10 and C-10).

For the experiments we used a cut-off of one, i.e. for the parameter estimation, events are relevant if they have been observed at least once in the training corpus.

The table is divided in such a way that the different class models from the different annotators can be distinguished by the entry in the first column. The second column shows the number of iterations of the GIS algorithm.

The difference of the two models in quality is justified in that way that annotator C uses a more “consistent” class model, i.e. a class model that a machine can learn better when using pure linguistic information that can be derived from the text of the document.

Table 4: Accuracy for Maximum Entropy Text Classification

| Model | Iterations | Recall | Precision | F |
|---|---|---|---|---|
| A-34 | 5 | 45.6 | 25.6 | 32.8 |
| | 10 | 52.9 | 28.2 | 36.8 |
| | 25 | 57.8 | 28.7 | 38.4 |
| | 50 | 58.7 | 28.2 | 38.1 |
| B-10 | 5 | 84.2 | 30.8 | 45.1 |
| | 10 | 84.2 | 40.0 | 54.2 |
| | 25 | 84.2 | 44.4 | 58.1 |
| | 50 | 73.7 | 42.4 | 53.8 |
| | 100 | 73.7 | 48.3 | 58.4 |
| | 250 | 73.7 | 48.3 | 58.4 |
| C-10 | 5 | 89.5 | 31.5 | 46.6 |
| | 10 | 89.5 | 36.2 | 51.5 |
| | 25 | 84.2 | 39.0 | 53.3 |
| | 50 | 84.2 | 42.1 | 56.1 |
| | 100 | 84.2 | 45.7 | 59.2 |
| | 250 | 84.2 | 50.0 | 62.7 |

## 4 Conclusion

We showed that statistical methods for document clustering and text classification are very promising approaches for Arabic, even without any morphological analysis.

We carried out experiments on a large Arabic corpus, namely Arabic NEWSWIRE, and used no preprocessing steps. The experiments on document clustering proved to work very well. In most cases the resulting clusters are classes that a human annotator would also define.

For the classification we showed experiments with maximum entropy text classification. Here, we show that even with no morphological analysis, we gain satisfying results.

## 5 Future Work

There are several extensions of the proposed method that should be investigated. For the classification task, the use of features based on *n*-gram characters would be an approach to a simple analysis the Arabic morphology, that is embedded in the classification process. Also gap-*n*-grams should be investigated, so that a more abstract level of morphological analysis can be reached.

Also another clustering criterion than the mutual criterion should be investigated: a weakness of the introduced approach is that it only analyses the frequency of a word or sub-word. The context of this unit is not analyzed explicitly. An alternative approach where the context information can be easily handled is the Likelihood criterion.

Another improvement of the system should be possible by introducing a so-called garbage class. This garbage class is supposed to accept all these documents that cannot be reliably assigned to the regular classes.

For text classification, the feature set used should be enriched by long-range and linguistic features. Also the speed of the training algorithm can be improved by using the improved iterative scaling algorithm (IIS).

### Acknowledgements

The authors would like to thank Jan Kunkis for performing some of the experiments.

## References

[Beesley 96] Ken R. Beesley. *Arabic Finite-State Morphological Analysis and Generation*. In Proceedings of the 16th Intl. Conference on Computational Linguistics, vol. 1, pp. 89-94, Copenhagen, August 1996.

[Kneser & Ney 91] Reinhard Kneser and Hermann Ney. *Forming Word Classes by Statistical Clustering for Statistical Language Modeling*. In Proceedings of the 1st QUALICO Conference, Trier, Germany, September 1991.

[Martin et al. 97] Michael Simons, Hermann Ney, Sven C. Martin. *Distant Bigram Language Modeling Using Maximum Entropy*. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing, Vol. 2, pp. 787-790, Munich, Germany, April 1997.

[Martin et al. 99] Sven Martin, Hermann Ney, Jörg Zaplo. *Smoothing Methods in Maximum Entropy Language Modeling*. In IEEE International Conference on Acoustics, Speech and Signal Processing, Vol. I, pp. 545-548, Phoenix, AR, March 1999.

[Melamed 97] I. Dan Melamed. *A word-to-word model of translational equivalence*. In 35th Conference of the Association for Computational Linguistics (ACL'97), pp. 490-497, Madrid, 1997.

[Nigam et al. 99] Kamal Nigam, John Lafferty, and Andrew McCallum. *Using maximum entropy for text classification*. In IJCAI-99 Workshop on Machine Learning for Information Filtering, pp. 61-67, 1999.

[Peters & Klakow 99] Jochen Peters and Dietrich Klakow. *Compact maximum entropy language models*. In Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding, Keystone, CO, December 1999.

[Ratnaparkhi 96] Adwait Ratnaparkhi, *A maximum entropy model for part-of-speech tagging*. In Proceedings of Conference on Empirical Methods in Natural Language Processing, University of Pennsylvania, 1996.

[Ratnaparkhi 98] Adwait Ratnaparkhi. *A linear observed time statistical parser based on maximum entropy models*. In Proceedings of the Second Conference on Empirical Methods in Natural Language Processing, pp. 1-10, 1997.

[Rosenfeld 94] Ronald Rosenfeld, Adaptive Statistical Language Modeling: *A Maximum Entropy Approach*, Ph.D. thesis, Computer Science Department, Carnegie Mellon University, TR CMU-CS-94-138, April 1994.

[Sawaf et al. 00] Hassan Sawaf, Kai Schütz, Hermann Ney. *On the Use of Grammar Based Language Models for Statistical Machine Translation*. In Proceedings 6th Intl. Workshop on Parsing Technologies, pp. 231-241. Trento, Italy, February 2000.