L'Approche MARF à DEFT 2010: A MARF Approach to DEFT 2010

Serguei A. Mokhov
Université Concordia University
Montréal, QC, Canada
mokhov@cse.concordia.ca

Résumé. On présent l'approche MARF aux problèmes de classification de variation diachronique et origine géographique des textes de la presse francophone pour l'atelier DEFT 2010. Cet étude utilise MARF, un framework open-source écrit en Java pour la reconnaissance des formes automatique en générale, incluant le traitement multimédia, l'analyse forensique de fichiers, la reconnaissance des auteurs, des langues parlées (accents), des sexes, de l'âge, et langues naturelles. On présent les officiels et les meilleures résultats obtenus et notre approche, ses difficultés, avantages et désavantages etc. Pour les résultats complets veuillez consulter un autre document relié (Mokhov, 2010a).

Abstract. We present a MARF-based approach to classification problems of the decades and place of origin of various French publications in the DEFT 2010 challenge. This case study of MARF, the open-source Java-based Modular Audio Recognition Framework, is intended to show the complete general pattern recognition pipeline design methodology for machine learning to test and compare multiple algorithms, including supervised and unsupervised, statistical, etc. learning and classification for a spectrum of recognition tasks, applicable not only to audio recognition but to general pattern recognition for various NLP applications, e.g. writer, language identification, and others. We summarize our best results and the results used for the challenge here along with the methodology used to obtain them. For vast, a lot more complete results of this work please refer to the related document (Mokhov, 2010a).

Mots-clés : DEFT2010, MARF, frameworks, comparaison des algorithmes pour TAL.

Keywords: DEFT2010, MARF, frameworks, algorithm comparison for NLP.

1 Introduction

We present an approach to the DÉfi Fouille de Textes (DEFT) 2010 challenge (Forest *et al.*, 2010) by using the MARF framework (The MARF Research and Development Group, 2002 2010; Mokhov *et al.*, 2002 2003; Mokhov, 2008a,b; Mokhov & Debbabi, 2008; Mokhov *et al.*, 2009; Mokhov, 2010b; Mokhov & Vassev, 2009) by combining approaches from the related MARF applications into `DEFT2010App`. The DEFT2010 NLP challenge proposed two tracks in identification within francophone press : Piste 1 for identification of the decade of a publication, and Piste 2 for publications varying across geographic locations, specifically France vs. Quebec from several prominent journals. The corpora were compiled by Cyril Grouin from a variety of sources kindly provided by Gallica, CEDROM-SNi, ELDA, and CNRTL (Grouin, 2010a,c,e,d,b).

To the author's knowledge MARF is a still holding up as a unique framework that attempts various signal processing, etc. and NLP for comparative studies of implementations of algorithms and algorithm combinations. The closest open-source system is probably CMU Sphinx (The Sphinx Group at Carnegie Mellon, 2007 2010), which is a powerful speech-to-text system, but at the same time too complex for comparative scientific experiments MARF's primarily designed for. Plus, MARF's multifaceted approach allowed it to be used outside of the domain of audio and voice processing (Mokhov, 2010c).

The core founding works for this approach are (Mokhov, 2008a,b; Mokhov *et al.*, 2009; Mokhov, 2010c) adapted to the NLP tasks using the same multi-algorithmic approach – providing a framework for algorithm selection and selecting the best available combination of algorithms for a given task (Mokhov, 2010c).

Some MARF's example applications (on which `DEFT2010App` is based), such as text-independent speaker-identification, language (natural and programming) identification, natural language probabilistic parsing, etc. are released along with MARF as open-source and some are discussed in several related publications mentioned earlier (Mokhov *et al.*, 2002 2010; Mokhov & the MARF Research & Development Group, 2003 2010a,2; Mokhov, 2008 2010).

2 Methodology

Using MARF, we show the usefulness of helping researchers to decide the algorithm combinations the best or better suited for each particular task they need to work on (Mokhov, 2008b, 2010c). MARF provides an ability to train the system for each task and then test it on the unseen samples giving back statistics for each algorithm combination (different permutations of algorithms are used in loading, preprocessing, feature extraction, and classification stages) used from the best to the worst, including second-best statistics augmented with statistical estimators and NLP parsing and other modules. The DEFT workshop participants (and later the rest of the world) will get the complete set of non-copyrighted materials and will be able to extend it and contribute to the project, which is open-source, if they choose to. This approach is similar to that described in (Mokhov, 2008b), but applied to corpora instead of voice samples using primarily spectral analysis of texts (spectral analysis was previous applied to texts by others as well, e.g. (Vaillant *et al.*, 2006), and speech and signal processing (Russell & Norvig, 1995; Press, 1993; O'Shaughnessy, 2000; Ifeachor & Jervis, 2002; Bernsee, 1999 2005)).

2.1 Two Pipelines Approach / L'Approche Deux Pipelines

This approach is applied to both tracks, Piste 1 for the decades and Piste 2 for the geographic locations. We detail the methodology further.

2.1.1 Classical MARF Pipeline Approach

The classical MARF pipeline is in Figure 1 (Mokhov, 2008b). It shows variety of algorithms that can be selected at run time via configuration settings to allow any permutation of the algorithms on the recognition path. The pipeline is augmented with the classical statistical NLP components from (Jurafsky & Martin, 2000; Martin, 2003) and others to do the statistical NLP processing are illustrated in a high-level UML class diagram in Figure 2. It's split into the subframeworks covering n-gram language models, natural language parsing, collocations, etc. and the supporting utility modules. The focus here, however, is primarily on the statistical analysis and recognition using machine learning, language models, and signal processing techniques.

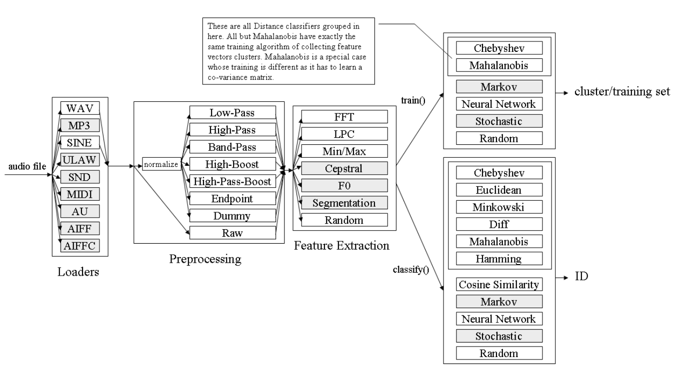

FIGURE 1 – Classical Pattern Recognition Pipeline of MARF

2.1.2 NLP MARF Pipeline Approach

This is another, somewhat distinct branch of experiments than that briefly described in Section 2.1.1 (while still using the same MARF-implementing system). This path of experiments is based on the largest part of the `LangIdentApp` (Mokhov & the MARF Research & Development Group, 2003 2010a) and the corresponding `MARF.NLP` class as well as the statistical estimators framework shown on Figure 2. This approach takes a different set of options and the parameters than the one in Section 2.1.1. We tokenize the input stream as individual characters and the build classical n-gram models with $n = 1, 2, 3$ and the corresponding statistical and smoothing estimators. The language model is then the smoothed using a 1D, 2D, or 3D frequency matrix and any future comparison for classification is that of the matrices learned. Each portion of an article (for both tasks) is used from the training data to compute the language models and serialize them per estimator. Then, we use the same models to test on the testing data. The precision statistics is computed identically to the classical pipeline approach. Thus, we still have a comparison of algorithms in a pipeline, but this is a more traditional statistical NLP pipeline.

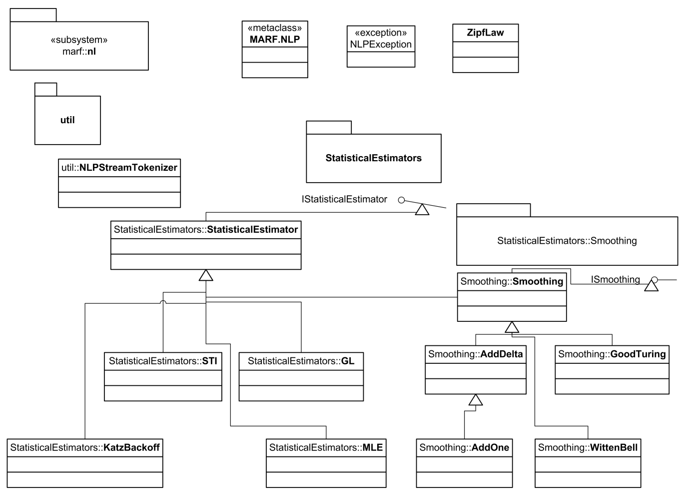

FIGURE 2 – A Partial Set of the NLP Components of MARF

2.2 Variable Tunable Parameters

This section describes the permutations of variable parameters used throughout the experiments to determine the best permutation/combination available.

2.2.1 DEFT'10-Specific Options

These options combine with all the subsequent options to select an algorithm combination to do the actual data processing. What follows is to tell what type of training and testing data to process in the pipelines.

1. `-piste1` – indicates we are dealing with the Piste 1 data, such that the internal data structures are properly configured to handle that ; it corresponds to Section 2.6.
2. `-journal` – a Piste 2 option indicating to use journals as primary classes as opposed the geographical locations ; and then compute the latter from the former (as knowing each journals gives uniquely the place of origin, while the reverse is not true). It corresponds to Section 2.7.3.
3. `-text-only` indicates to process just the article fragments themselves. This is default of loading bodies of the articles.
4. `-title-only` – a Piste 2 option indicating to use only article titles for the tasks instead of the article bodies.
5. `-title-text` – a Piste 2 option indicates to use both the article title and its body as a sample It corresponds to Section 2.7.2.
6. `-ref` – indicates to load and validate against the reference data supplied by the organizers (added later when the data became available)

2.2.2 Sample Loading and Interpretation

At present there are two types of interpretation of the data performed after the initial XML loading : (a) interpreting each character bigram as an amplitude of a waveform and treat the whole input data as an audio wave signal with 2 bytes per amplitude value, mono, WAVE PCI-encoded data (all are the defaults in MARF). The `Sample` objects produced in that way are processed by the traditional MARF pipeline, which is more signal processing and spectral oriented (cf. Figure 2.1.1). (b) interpreting each n-gram ($n = 1, 2, 3$) token from `NLPStreamTokenizer` to compute the language models and use the NLP pipeline of MARF (cf. Section 2.1.2). `NLPStreamTokenizer` is itself configurable to filter different kinds of tokens and characters. Reduction of the data set loaded for training is another variable to explore as sometimes "oversaturation" of the training models with a lot of training data actually lower the recognition performance. There are a lot more experiments possible with the loaders if the interpreted data, which was not done for this submission, but may be performed later in one of the versions of (Mokhov, 2010a).

2.2.3 Preprocessing

Preprocessing algorithms in the classical pipeline take loaded `Sample` objects of arbitrary length and produce arbitrary length "preprocessed" objects of the same type. The resulting objects may become smaller or keep the same length as the original depending on the module used. These are typically normalization and filtering modules. They also can be chained, but the chaining aspect was not really fully explored in this work.

1. `-silence` and `-noise` – suppress "silence" and "noise" from the data (cf. Section 2.7.4).
2. `-norm` – whether to apply normalization of the input data or with `-raw` just to pass it through without any preprocessing.
3. `-low`, `-high`, `-band`, `-bandstop`, `-high`, `-boost`, `-highpassboost` – are a variety of FFT-based filters, such as low-pass, high-pass, band-pass, band-stop, high-frequency amplitude boost, and high-pass with high frequency amplitude boost.

2.2.4 Feature Extraction

All feature extractor modules take a variable-length sample input from preprocessing and produce a fixed-length feature vector x. The resulting feature vectors get stored in various training models such as mean (default) or median clusters, or plain collections of feature vectors.

1. `-fft` – use the Fast Fourier Transform (FFT) with 512 default features (frequencies)
2. `-lpc` – use the Linear Predictive Coding (LPC) algorithm with the default 20 features (poles)
3. `-aggr` – uses aggregation of the FFT and LPC features as a feature vector x
4. `-minmax` – use 50 minimums and 50 maximums by default as features

2.2.5 Classification

Classical mostly spectral pipeline. Classifier modules typically take the fixed-length feature vectors x from the feature extractors and then store them as either mean (default) or median cluster or just as a collection of feature vectors came in during training. This is true for the majority of distance and similarity classifiers. Additional training is done for Mahalanobis distance (Mahalanobis, 1936) (covariance matrix) and the artificial neural network (the network itself). These respresent the language models that are compared against when testing.

1. `-cheb`, the Chebyshev (aka city-block or Manhattan) distance classifier (Abdi, 2007)
2. `-eucl`, the Euclidean distance
3. `-cos` – use cosine similarity measure as a classifier (Garcia, 2006; Khalifé, 2004)
4. `-mink` – use Minkowski distance measure
5. `-diff` – use Diff distance measure (Mokhov, 2008b)
6. `-hamming` – use Hamming distance measure (Hamming, 1950)

Mean cluster vs. median cluster vs. feature set. It has been shown that the choice of cluster or absence thereof may positively impact precision (Mokhov, 2008b) and providing distinct top algorithm combinations from that of mean. Both type of clustering while save processing time and storage space, are deemed to contribute less accurately to precision than just keeping all the collections of all feature vectors as-is. Thus, this set of experiments is to test variability of the precision (and other metrics) to the Piste 1 and Piste 2 tasks. These experiments can be combined with the previously described variations and the mean clustering was the default that was used. The follow up experiments and observations will be reported in the subsequent ongoing study in (Mokhov, 2010a).

NLP statistical estimators pipeline. In this pipeline we script the task in a similar manner where `-char` used in the preprocessing to set the character model for the tokenizer ; `-unigram`, `-bigram`, and `-trigram` act as feature selectors, and finally the smoothing estimators enumerated below act as classifiers. This experiment is still ongoing and its results are expected to appear in (Mokhov, 2010a).

1. `-add-delta` by implementing the general Add-Delta smoothing we get MLE, ELE, and Add-One "for free" as special cases of Add-Delta :
   - $\delta = 0$ is MLE (maximum likelihood estimate), `-mle`
   - $\delta = 1$ is Add One, `-add-one`
   - $\delta = 0.5$ is ELE, `-add-delta`
2. Witten-Bell and Good Turing smoothing options `-witten-bell` and `-good-turing`. These two estimators were implemented as given in hopes to get a better recognition rate over the Add-Delta family.

2.3 Omitted "Slow" and Other Algorithms of MARF

The following algorithms were not conclusively tested yet for either task and are either in process of execution or being debugged due to slowness, etc. Their respective results are to appear in (Mokhov, 2010a) as they become available. Please consult that article from time to time for the respective updates. The algorithms include the artificial neural network, continuous fraction expansion (CFE) filters (Haridas, 2006) (high-, low-, band-pass, and band-stop), Zipf's Law, and Mahalanobis distance – all combined with all the mentioned algorithms as applicable.

2.4 Zipf's Law

The `ZipfLaw` classifier exists for both classical and NLP pipelines of MARF. In the former it uses the `DiffDistance` class to compute the distance between the two ranked dictionaries of `Double` values. This corresponds to the `-zipf` option in the `DEFT2010App` application. This is a very slow approach by default in the traditional pipeline. In the NLP pipeline, the `ZipfLaw` module is used to collect the most discriminative terms from training and rank them. We keep the top N rank lexemes (n-grams, where lexemes can be a combination of n characters or words producing shorter or longer dictionaries). On classification, we compute a ranked ZipfLaw top N rank set, and compute their distance or similarity as vectors to decide the class. The option `-cheat` corresponds to this. (At the time of this writing we do not have the Zipf's Law results available ; all options are enumerated in (Mokhov, 2010a) and that's where we plan placing the Zipf's law results as well).

2.5 Extra Testing on the Unseen Data

This section proposes an additional testing methodology on the unseen data while the reference and the test data are not available.

2.5.1 Using Piste 1's Data in Piste 2

All of the Piste 1's data are reusable in Piste 2 to partially test Piste 2's setup on an unseen corpora, which is conveniently and freely available in order to improve the testing of the system. This extra testing requires to be able to load both the training data of the Piste 1 and the language models of Piste 2 at the same time and provide the appropriate option and scripting support as well as mapping the relevant Piste 1's journals to France. While only partial testing, it is more honest than testing on the training data. This methodology is further being developed in (Mokhov, 2010a).

2.5.2 Using Francophone Websites

Using the `wget` (Niksic & Free Software Foundation, Inc., 1996 2009) tool in a scripted manner on the French and Quebec web sites in a recursive manner helps creating a corpora of HTML pages. Then, treat either each page as an article of its own (after tags removal) or each textual node paragraph larger than 2000 bytes as an article. The good candidate sites for this type of testing data are various government and press sites openly available. The results of this testing are planned to further to appear in (Mokhov, 2010a).

2.6 Piste 1 : Decades

Decades proved difficult to get correctly. The best configuration on the training and testing data are in Section 3.2. Most of the general methodology described in Section 2.1 and Section 2.2 were used to come up with the best available configuration to date that produces highest macro precision. Top N (usually $N = 50$) of those best configurations were subsequently measured with the testing and reference data later provided by the organizers. The complete result set for Piste 1 trials can be found in (Mokhov, 2010a).

2.7 Piste 2 : Place of Origin

There is a lot more variety in the experiments for the methodology used in Piste 2 as opposed to Piste 1. The major variations are listed here. The results are summarized in Section 3.3.

2.7.1 Small Increase in the Article Text Size

We collect titles as well as texts. For the training purposes one way to experiment is to merge the two effectively increasing a little bit the training sample size for each article fragment. The experiment is to verify if adding titles to texts increases the precision of detection. The results prove that this hypothesis is not quite correct. In fact, the precision has gone worse and the processing time increased in many configurations.

This experiment was conducted both with the location being the leading class, followed by the journal as well as when the journal is the leading class as described in Section 2.7.3.

2.7.2 Titles vs. Texts

For a 300-word article, its title may act as an abstract. Often in NLP, e.g. BioNLP, NLP techniques area applied to the article abstracts rather than to the complete texts. A title in our sample can be considered as an abstract on the same scale of a short article excerpt. Thus, we test the approach to see if titles alone are sufficient to generate enough precision and increasing the performance.

Disadvantages : sometimes articles do not have titles (i.e. `<titre />`, so only guessing at random (an improvement is to use the main body in such cases).

- Training data's articles without titles : 297
- Testing data's articles without titles : 207

Despite the missing titles in some articles and without the implementation of the mentioned improvement, the results produces the higher precision than in the other experiments mentioned thus far.

2.7.3 Journal as a Primary Class

There is a relationship between journals and locations where they were published. The journals uniquely identify the place ; however, the place does not necessarily uniquely identify the journal, but puts a constraint on what the journals may be allowing cross-validation to prevent journal-place mismatches.

In most of the experiments the location was a leading class, with the journal being identified later as follows :

- Assume a common journal per location without actually running a classifier.

– Run the pipeline the second time for the journal class once the location is known and pick the likely journal from that country.
– Assume the journals are the primary class and then deduce the location from them. It turns out while the journal is primary, despite relatively low macro precision for journals ($\approx 40\% - 48\%$), the macro precision for some experiments turned out better for countries and for an equivalent non-journal identification.

2.7.4 Silence and Noise Removal

These are spectral options (`-silence`, `-noise`) from the classical MARF. The experiments show the default options of signal or noise removal (see Section 2.2) made difference some cases. The silence is removed when the preprocessed wave signal's amplitude points are below a certain empirical threshold (0.001). The "noise" is presently removed by applying the low-pass FFT filter by default. The complete result set containing the experiments along with others are in (Mokhov, 2010a).

3 Results / Résultats

Here we present some official results first, and then other related experiments and improvement upon them. These are just a small fraction of all the experiments conducted to date.

3.1 Official / Officiels

Our official results at the submission time were not great ; a part of the reason the testing on all combinations were not completed yet (some are still executing), so the best available configurations at that moment were used, which are *emphasized* in the tables. Those configurations gave highest macro precision on the training data back then, found in Table 1 and Table 5. Those submissions produced the results found in Table 9 and and Table 10. We, however, did not stop the experiments running and conducting some more at the time of this writing, that are being maintained more-or-less in full in (Mokhov, 2010a). We were able to improve the results over the official submission, and what follows are some of the top picks.

TABLE 1 – Piste 1 : Top 10 configurations tested on the training data.

Run #	Guess	Configuration	GOOD	BAD	Precision,%
1	1st	*-piste1 -norm -fft -cos*	1025	2569	*28.52*
2	1st	-piste1 -silence -norm -aggr -cos	1025	2569	28.52
3	1st	-piste1 -silence -norm -fft -cos	1025	2569	28.52
4	1st	-piste1 -norm -aggr -cos	1025	2569	28.52
5	1st	-piste1 -raw -fft -cos	1023	2571	28.46
6	1st	-piste1 -silence -raw -aggr -cos	1023	2571	28.46
7	1st	-piste1 -silence -noise -raw -fft -cos	1023	2571	28.46
8	1st	-piste1 -noise -raw -fft -cos	1023	2571	28.46
9	1st	-piste1 -silence -noise -raw -aggr -cos	1023	2571	28.46
10	1st	-piste1 -noise -raw -aggr -cos	1023	2571	28.46

3.2 Piste 1 : Decades

The resulting tables pertinent to the Piste 1 are in Table 1 and Table 2 per configuration and per class with macro precision tested on the training data ; Table 3 and Table 4 correspond to the same with the testing and reference data, while in Table 9 is the official result at the submission time.

3.3 Piste 2 : France vs. Quebec

The resulting tables pertinent to the Piste 2 are in Table 10, Table 5, Table 6, Table 7, Table 8. Additional extra results that were obtained after the submissions are in Table 11, which shows consolidated top 4 macro precision results per configuration for the title + text experiment (cf. Section 2.7.1), title-only (cf. Section 2.7.2) ; then repeat of the three experiments with journal being the leading class (cf. Section 2.7.3).

TABLE 2 – Piste 1 : Macro precision on training data per decade across 834 configurations

Run #	Guess	Decade	GOOD	BAD	Precision,%
1	1st	1830	84722	125446	40.31
2	1st	1940	38330	147652	20.61
3	1st	1810	38811	171357	18.47
4	1st	1820	29107	181061	13.85
5	1st	1900	15229	166583	8.38
6	1st	1920	15360	168954	8.33
7	1st	1850	16807	193361	8.00
8	1st	1880	15643	194525	7.44
9	1st	1800	14974	195194	7.12
10	1st	1860	14038	195296	6.71
11	1st	1870	14037	196131	6.68
12	1st	1930	11314	173000	6.14
13	1st	1840	12468	197700	5.93
14	1st	1890	10727	176089	5.74
15	1st	1910	10408	173072	5.67
16	2nd	1830	105137	105031	50.03
17	2nd	1940	53906	132076	28.98
18	2nd	1810	98904	111264	47.06
19	2nd	1820	45219	164949	21.52
20	2nd	1900	28690	153122	15.78
21	2nd	1920	30199	154115	16.38
22	2nd	1850	32345	177823	15.39
23	2nd	1880	36923	173245	17.57
24	2nd	1800	38997	171171	18.56
25	2nd	1860	29063	180271	13.88
26	2nd	1870	28356	181812	13.49
27	2nd	1930	23456	160858	12.73
28	2nd	1840	25489	184679	12.13
29	2nd	1890	21281	165535	11.39
30	2nd	1910	19657	163823	10.71

TABLE 3 – Piste 1 : Testing on the evaluation + reference data top 10 configurations

Run #	Guess	Configuration	GOOD	BAD	Precision,%
1	1st	**-piste1 -ref -silence -bandstop -aggr -cos**	331	2390	**12.16**
2	1st	-piste1 -ref -noise -raw -aggr -eucl	315	2406	11.58
3	1st	-piste1 -ref -silence -raw -aggr -eucl	315	2406	11.58
4	1st	*-piste1 -ref -norm -fft -cos*	315	2406	*11.58*
5	1st	-piste1 -ref -raw -aggr -eucl	315	2406	11.58
6	1st	-piste1 -ref -noise -raw -fft -cos	315	2406	11.58
7	1st	-piste1 -ref -silence -raw -fft -eucl	315	2406	11.58
8	1st	-piste1 -ref -silence -raw -aggr -cos	315	2406	11.58
9	1st	-piste1 -ref -silence -noise -raw -fft -eucl	315	2406	11.58
10	1st	-piste1 -ref -silence -noise -raw -aggr -cos	315	2406	11.58

Table 12 lists the corresponding macro results per class. Following that, the tables Table 13, Table 14, Table 15, Table 16, and Table 17 list the corresponding output from the `3_evalueResultats_t2.pl` evaluation tool provided by the DEFT2010 organizers in the end each corresponding to the top configuration of each experiment.

3.4 Results Summary / Résumé des résultats

- Total more than 8688 configurations tested.
- Apparent highest precision and recall results come from title-only processing for Piste 2 despite some testing and training items missing titles.
- While journal-leading-class experiments give 48% at their best macro precision per configuration they sometimes give better results for the corresponding location, than if the location is the leading class.

TABLE 4 – Piste 1 : Macro precision on testing + reference data per decade across 49 configurations

Run #	Guess	Decade	GOOD	BAD	Precision,%
1	1st	1940	3852	5850	39.70
2	1st	1830	2187	6094	26.41
3	1st	1820	1571	6710	18.97
4	1st	1810	1180	7101	14.25
5	1st	1880	757	7524	9.14
6	1st	1900	864	9083	8.69
7	1st	1920	703	9097	7.17
8	1st	1850	554	7727	6.69
9	1st	1870	417	7864	5.04
10	1st	1840	413	7868	4.99
11	1st	1860	395	7935	4.74
12	1st	1930	442	9358	4.51
13	1st	1890	403	9250	4.17
14	1st	1800	281	8000	3.39
15	1st	1910	290	9559	2.94
16	2nd	1940	4828	4874	49.76
17	2nd	1830	3457	4824	41.75
18	2nd	1820	3068	5213	37.05
19	2nd	1810	1772	6509	21.40
20	2nd	1880	1770	6511	21.37
21	2nd	1900	1924	8023	19.34
22	2nd	1920	1725	8075	17.60
23	2nd	1850	1068	7213	12.90
24	2nd	1870	881	7400	10.64
25	2nd	1840	966	7315	11.67
26	2nd	1860	884	7446	10.61
27	2nd	1930	1150	8650	11.73
28	2nd	1890	806	8847	8.35
29	2nd	1800	978	7303	11.81
30	2nd	1910	554	9295	5.62

TABLE 5 – Piste 2 : Top 38 of 839 results of testing on the training data using text only

Run #	Guess	Configuration	GOOD	BAD	Precision,%
1	1st	**-silence -high -fft -cos**	2203	1516	**59.24**
2	1st	-silence -high -aggr -cos	2202	1517	59.21
3	1st	-high -fft -cos	2169	1550	58.32
4	1st	-high -aggr -cos	2167	1552	58.27
5	1st	-silence -band -aggr -cos	2154	1565	57.92
6	1st	-silence -band -fft -cos	2154	1565	57.92
7	1st	-silence -noise -band -fft -cos	2138	1581	57.49
8	1st	-silence -noise -band -aggr -cos	2134	1585	57.38
9	1st	-silence -high -aggr -eucl	2134	1585	57.38
10	1st	-silence -high -fft -eucl	2133	1586	57.35
11	1st	-high -fft -diff	2129	1590	57.25
12	1st	-high -aggr -diff	2127	1592	57.19
13	1st	-high -aggr -cheb	2126	1593	57.17
14	1st	-high -fft -cheb	2124	1595	57.11
15	1st	-band -aggr -eucl	2124	1595	57.11
16	1st	-band -fft -eucl	2123	1596	57.09
17	1st	-noise -band -aggr -cos	2122	1597	57.06
18	1st	-band -fft -cos	2122	1597	57.06
19	1st	-band -aggr -cos	2122	1597	57.06
20	1st	-noise -band -fft -cos	2121	1598	57.03
21	1st	-high -aggr -eucl	2119	1600	56.98
22	1st	-high -fft -eucl	2118	1601	56.95
23	1st	-silence -band -fft -eucl	2115	1604	56.87
24	1st	-silence -band -aggr -eucl	2113	1606	56.82
25	1st	-silence -band -aggr -diff	2084	1635	56.04
26	1st	-silence -band -aggr -cheb	2084	1635	56.04
27	1st	-silence -band -fft -cheb	2083	1636	56.01
28	1st	-silence -band -fft -diff	2082	1637	55.98
29	1st	-band -fft -cheb	2071	1648	55.69
30	1st	-band -aggr -cheb	2070	1649	55.66
31	1st	-noise -raw -lpc -eucl	2058	1661	55.34
32	1st	-silence -raw -lpc -eucl	2058	1661	55.34
33	1st	-silence -noise -raw -lpc -eucl	2058	1661	55.34
34	1st	-band -fft -diff	2058	1661	55.34
35	1st	-silence -norm -lpc -eucl	2058	1661	55.34
36	1st	-norm -lpc -eucl	2058	1661	55.34
37	1st	-band -aggr -diff	2058	1661	55.34
38	1st	*-raw -lpc -eucl*	2058	1661	*55.34*

TABLE 6 – Piste 2 : Macro precision across 839 results per location on training data, text only

Run #	Guess	Location	GOOD	BAD	Precision,%
1	1st	Quebec	932695	737754	55.83
2	1st	France	688020	761772	47.46

TABLE 7 – Piste 2 : Top 15 configurations on the evaluation and reference data, text only

Run #	Guess	Configuration	GOOD	BAD	Precision,%
1	1st	-text-only -ref -noise -band -fft -diff	1392	1090	56.08
2	1st	-text-only -ref -noise -band -aggr -diff	1391	1091	56.04
3	1st	-text-only -ref -noise -band -fft -cheb	1386	1096	55.84
4	1st	-text-only -ref -noise -band -aggr -cheb	1386	1096	55.84
5	1st	-text-only -ref -high -fft -diff	1380	1102	55.60
6	1st	-text-only -ref -band -aggr -cheb	1380	1102	55.60
7	1st	-text-only -ref -band -fft -cheb	1379	1103	55.56
8	1st	-text-only -ref -high -fft -cheb	1378	1104	55.52
9	1st	-text-only -ref -high -aggr -diff	1378	1104	55.52
10	1st	-text-only -ref -high -aggr -cheb	1377	1105	55.48
11	1st	-text-only -ref -band -aggr -diff	1376	1106	55.44
12	1st	-text-only -ref -band -fft -diff	1375	1107	55.40
13	1st	-text-only -ref -silence -band -aggr -cheb	1325	1157	53.38
14	1st	-text-only -ref -silence -band -fft -cheb	1322	1160	53.26
15	1st	-text-only -ref -silence -band -aggr -diff	1317	1165	53.06

TABLE 8 – Piste 2 : Macro precision across all configurations per location on evaluation data, text only

Run #	Guess	Location	GOOD	BAD	Precision,%
1	1st	France	32445	24052	57.43
2	1st	Quebec	31720	33401	48.71

TABLE 9 – Fichier évalué : `equipe_3_tache_1_execution_1_1006060550.xml`

- classe 1800 (attendus = 169, ramenes = 127.00, corrects = 9.00) rappel = 0.053 precision = 0.071 f-mesure = 0.061
- classe 1810 (attendus = 169, ramenes = 249.00, corrects = 32.00) rappel = 0.189 precision = 0.129 f-mesure = 0.153
- classe 1820 (attendus = 169, ramenes = 258.00, corrects = 29.00) rappel = 0.172 precision = 0.112 f-mesure = 0.136
- classe 1830 (attendus = 169, ramenes = 354.00, corrects = 36.00) rappel = 0.213 precision = 0.102 f-mesure = 0.138
- classe 1840 (attendus = 169, ramenes = 85.00, corrects = 10.00) rappel = 0.059 precision = 0.118 f-mesure = 0.079
- classe 1850 (attendus = 169, ramenes = 187.00, corrects = 17.00) rappel = 0.101 precision = 0.091 f-mesure = 0.096
- classe 1860 (attendus = 170, ramenes = 146.00, corrects = 13.00) rappel = 0.076 precision = 0.089 f-mesure = 0.082
- classe 1870 (attendus = 169, ramenes = 102.00, corrects = 10.00) rappel = 0.059 precision = 0.098 f-mesure = 0.074
- classe 1880 (attendus = 169, ramenes = 166.00, corrects = 19.00) rappel = 0.112 precision = 0.114 f-mesure = 0.113
- classe 1890 (attendus = 197, ramenes = 83.00, corrects = 7.00) rappel = 0.036 precision = 0.084 f-mesure = 0.050
- classe 1900 (attendus = 203, ramenes = 156.00, corrects = 15.00) rappel = 0.074 precision = 0.096 f-mesure = 0.084
- classe 1910 (attendus = 201, ramenes = 83.00, corrects = 10.00) rappel = 0.050 precision = 0.120 f-mesure = 0.070
- classe 1920 (attendus = 200, ramenes = 133.00, corrects = 15.00) rappel = 0.075 precision = 0.113 f-mesure = 0.090
- classe 1930 (attendus = 200, ramenes = 98.00, corrects = 10.00) rappel = 0.050 precision = 0.102 f-mesure = 0.067
- classe 1940 (attendus = 198, ramenes = 494.00, corrects = 82.00) rappel = 0.414 precision = 0.166 f-mesure = 0.237
- sur l'ensemble des 15 classes macro rappel = 0.116 macro precision = 0.107 macro F-mesure = 0.111

TABLE 10 – Fichier évalué : `equipe_3_tache_2_execution_2_1006060557.xml`

Evaluation du pays
- classe F (attendus = 1153, ramenes = 1313.00, corrects = 650.00) rappel = 0.564 precision = 0.495 f-mesure = 0.527
- classe Q (attendus = 1329, ramenes = 1169.00, corrects = 666.00) rappel = 0.501 precision = 0.570 f-mesure = 0.533
- sur l'ensemble des 2 classes macro rappel = 0.532 macro precision = 0.532 macro F-mesure = 0.532
Evaluation du journal
- classe D (attendus = 652, ramenes = 0.00, corrects = 0.00) rappel = 0.000 precision = 0.000 f-mesure = 0.000
- classe E (attendus = 553, ramenes = 0.00, corrects = 0.00) rappel = 0.000 precision = 0.000 f-mesure = 0.000
- classe M (attendus = 600, ramenes = 1313.00, corrects = 365.00) rappel = 0.608 precision = 0.278 f-mesure = 0.382
- classe P (attendus = 677, ramenes = 1169.00, corrects = 342.00) rappel = 0.505 precision = 0.293 f-mesure = 0.371
- sur l'ensemble des 4 classes macro rappel = 0.278 macro precision = 0.143 macro F-mesure = 0.189

TABLE 11 – Consolidated extra results on evaluation+reference data, top 4 each

Run #	Guess	Configuration	GOOD	BAD	Precision,%
1	1st	-title-text -ref -band -fft -cheb	1364	1118	54.96
2	1st	-title-text -ref -band -aggr -cheb	1363	1119	54.92
3	1st	-title-text -ref -band -aggr -diff	1362	1120	54.88
4	1st	-title-text -ref -band -fft -diff	1361	1121	54.83
Run #	Guess	Configuration	GOOD	BAD	Precision,%
1	1st	-title-only -ref -silence -noise -norm -aggr -eucl	1714	768	69.06
2	1st	-title-only -ref -silence -noise -norm -fft -eucl	1714	768	69.06
3	1st	-title-only -ref -low -aggr -eucl	1714	768	69.06
4	1st	-title-only -ref -noise -norm -aggr -eucl	1714	768	69.06
Run #	Guess	Configuration	GOOD	BAD	Precision,%
1	1st	-text-only -journal -ref -silence -noise -endp -fft -mink	1027	1455	41.38
2	1st	-text-only -journal -ref -noise -endp -aggr -mink	1027	1455	41.38
3	1st	-text-only -journal -ref -noise -endp -fft -mink	1027	1455	41.38
4	1st	-text-only -journal -ref -silence -noise -endp -aggr -mink	1027	1455	41.38
Run #	Guess	Configuration	GOOD	BAD	Precision,%
1	1st	-title-text -journal -ref -silence -noise -endp -aggr -eucl	1030	1452	41.50
2	1st	-title-text -journal -ref -silence -noise -endp -fft -eucl	1030	1452	41.50
3	1st	-title-text -journal -ref -silence -endp -aggr -diff	1029	1453	41.46
4	1st	-title-text -journal -ref -noise -endp -aggr -eucl	1029	1453	41.46
Run #	Guess	Configuration	GOOD	BAD	Precision,%
1	1st	-title-only -journal -ref -silence -bandstop -fft -diff	1210	1272	48.75
2	1st	-title-only -journal -ref -silence -bandstop -aggr -diff	1209	1273	48.71
3	1st	-title-only -journal -ref -bandstop -aggr -eucl	1201	1281	48.39
4	1st	-title-only -journal -ref -silence -bandstop -fft -eucl	1201	1281	48.39

4 Conclusion

We presented a MARF approach to DEFT2010 with some we believe interesting experiments and results. While the macro precision and recall leave a lot desired, we also have in indication which approaches to try and refine further. The author believes this is the only such a comprehensive study of multiple algorithm combinations and configurations used. We plan to release `DEFT2010App` as open-source at `marf.sf.net` following the documented description and the follow up analysis in (Mokhov, 2010a).

4.1 Challenges and Disadvantages of the Approach

While MARF offers a great experimentation platform, it also presents some challenges, limitations, and resulting disadvantages of using it. Here are some :

– Too many options and experiments to try (while also an advantage, it does not allow to complete all the planned experiments on time even when running on multiple machines in distributed manner).
– Unexpected UTF8 differences of compiled `.class` files on Linux vs. MacOS Java. The one on MacOS was not UTF8 by default for the strings found in class causing mismatching on the accented words such as "L'Est Républicain" and "Québec" and forcing to redo the experiments.
– Current I/O handling required files on the file system, so each article portion was serialized under its own file – a lot of I/O that hammers the performance.

4.2 Future Work / Travaux Futurs

– Add run-time statistics, recall, and f-measure to the reports.
– Complete testing the "slow" configurations mentioned in Section 2.3.
– Complete the majority of the ongoing experiments listed easrlier (to be reported in (Mokhov, 2010a)).
– MARF internally maintains a number of metrics, other than the macro precision, so we need to be able to output recall, f-measure, run-time, and other metrics in a readable way is another point to work on.
– Explore and exploit the second guess statistics.
– Explore dynamic classifier ensembles (Cavalin *et al.*, 2010).

TABLE 12 – Consolidated extra results per experiment per class on evaluation+reference data

Run #	text-title	Location	GOOD	BAD	Precision,%
1	1st	Quebec	34013	29861	53.25
2	1st	France	27657	27746	49.92
Run #	title-only	Location	GOOD	BAD	Precision,%
1	1st	Quebec	48669	16452	74.74
2	1st	France	33818	22679	59.86
Run #	text-only-journal	Journal	GOOD	BAD	Precision,%
1	1st	Le Monde	22885	6515	77.84
2	1st	L'Est Republicain	18359	8738	67.75
3	1st	Le Devoir	3989	27959	12.49
4	1st	La Presse	2861	30312	8.62
5	2nd	Le Monde	23410	5990	79.63
6	2nd	L'Est Republicain	19072	8025	70.38
7	2nd	Le Devoir	20804	11144	65.12
8	2nd	La Presse	18752	14421	56.53
Run #	text-title-journal	Journal	GOOD	BAD	Precision,%
1	1st	Le Monde	22786	6614	77.50
2	1st	L'Est Republicain	18411	8686	67.94
3	1st	Le Devoir	3880	28068	12.14
4	1st	La Presse	2974	30199	8.97
5	2nd	Le Monde	23318	6082	79.31
6	2nd	L'Est Republicain	19166	7931	70.73
7	2nd	Le Devoir	20680	11268	64.73
8	2nd	La Presse	19042	14131	57.40
Run #	title-only-journal	Journal	GOOD	BAD	Precision,%
1	1st	Le Monde	22887	6513	77.85
2	1st	L'Est Republicain	18418	8679	67.97
3	1st	Le Devoir	8686	23262	27.19
4	1st	La Presse	8324	24849	25.09
5	2nd	Le Monde	23530	5870	80.03
6	2nd	L'Est Republicain	19485	7612	71.91
7	2nd	Le Devoir	24466	7482	76.58
8	2nd	La Presse	24455	8718	73.72

TABLE 13 – Fichier évalué : `equipe_3_tache_2_execution_3-testing.sh-title-text-ref-band-fft-cheb.prexml.xml`

Evaluation du pays
- classe F (attendus = 1153, ramenes = 1521.00, corrects = 778.00) rappel = 0.675 precision = 0.512 f-mesure = 0.582
- classe Q (attendus = 1329, ramenes = 961.00, corrects = 586.00) rappel = 0.441 precision = 0.610 f-mesure = 0.512
- sur l'ensemble des 2 classes macro rappel = 0.558 macro precision = 0.561 macro F-mesure = 0.559
Evaluation du journal
- classe D (attendus = 652, ramenes = 0.00, corrects = 0.00) rappel = 0.000 precision = 0.000 f-mesure = 0.000
- classe E (attendus = 553, ramenes = 0.00, corrects = 0.00) rappel = 0.000 precision = 0.000 f-mesure = 0.000
- classe M (attendus = 600, ramenes = 1521.00, corrects = 372.00) rappel = 0.620 precision = 0.245 f-mesure = 0.351
- classe P (attendus = 677, ramenes = 961.00, corrects = 311.00) rappel = 0.459 precision = 0.324 f-mesure = 0.380
- sur l'ensemble des 4 classes macro rappel = 0.270 macro precision = 0.142 macro F-mesure = 0.186

TABLE 14 – Fichier évalué : `equipe_3_tache_2_execution_3-testing.sh-title-only-journal-ref-silence-noise-norm-aggr-eucl.prexml.xml`

Evaluation du pays
- classe F (attendus = 1153, ramenes = 1537.00, corrects = 874.00) rappel = 0.758 precision = 0.569 f-mesure = 0.650
- classe Q (attendus = 1329, ramenes = 945.00, corrects = 666.00) rappel = 0.501 precision = 0.705 f-mesure = 0.586
- sur l'ensemble des 2 classes macro rappel = 0.630 macro precision = 0.637 macro F-mesure = 0.633
Evaluation du journal
- classe D (attendus = 652, ramenes = 511.00, corrects = 178.00) rappel = 0.273 precision = 0.348 f-mesure = 0.306
- classe E (attendus = 553, ramenes = 746.00, corrects = 376.00) rappel = 0.680 precision = 0.504 f-mesure = 0.579
- classe M (attendus = 600, ramenes = 791.00, corrects = 472.00) rappel = 0.787 precision = 0.597 f-mesure = 0.679
- classe P (attendus = 677, ramenes = 434.00, corrects = 163.00) rappel = 0.241 precision = 0.376 f-mesure = 0.293
- sur l'ensemble des 4 classes macro rappel = 0.495 macro precision = 0.456 macro F-mesure = 0.475

TABLE 15 – Fichier évalué : `equipe_3_tache_2_execution_3-testing.sh-text-only-journal-ref-silence-noise-endp-fft-mink.prexml.xml`

Evaluation du pays
- classe F (attendus = 1153, ramenes = 2075.00, corrects = 1004.00) rappel = 0.871 precision = 0.484 f-mesure = 0.622
- classe Q (attendus = 1329, ramenes = 407.00, corrects = 258.00) rappel = 0.194 precision = 0.634 f-mesure = 0.297
- sur l'ensemble des 2 classes macro rappel = 0.532 macro precision = 0.559 macro F-mesure = 0.545
Evaluation du journal
- classe D (attendus = 652, ramenes = 318.00, corrects = 108.00) rappel = 0.166 precision = 0.340 f-mesure = 0.223
- classe E (attendus = 553, ramenes = 995.00, corrects = 410.00) rappel = 0.741 precision = 0.412 f-mesure = 0.530
- classe M (attendus = 600, ramenes = 1080.00, corrects = 476.00) rappel = 0.793 precision = 0.441 f-mesure = 0.567
- classe P (attendus = 677, ramenes = 89.00, corrects = 33.00) rappel = 0.049 precision = 0.371 f-mesure = 0.086
- sur l'ensemble des 4 classes macro rappel = 0.437 macro precision = 0.391 macro F-mesure = 0.413

TABLE 16 – Fichier évalué : `equipe_3_tache_2_execution_3-testing.sh-title-text-journal-ref-silence-noise-endp-aggr-eucl.prexml.xml`

Evaluation du pays
- classe F (attendus = 1153, ramenes = 2089.00, corrects = 1010.00) rappel = 0.876 precision = 0.483 f-mesure = 0.623
- classe Q (attendus = 1329, ramenes = 393.00, corrects = 250.00) rappel = 0.188 precision = 0.636 f-mesure = 0.290
- sur l'ensemble des 2 classes macro rappel = 0.532 macro precision = 0.560 macro F-mesure = 0.546
Evaluation du journal
- classe D (attendus = 652, ramenes = 279.00, corrects = 96.00) rappel = 0.147 precision = 0.344 f-mesure = 0.206
- classe E (attendus = 553, ramenes = 996.00, corrects = 414.00) rappel = 0.749 precision = 0.416 f-mesure = 0.535
- classe M (attendus = 600, ramenes = 1093.00, corrects = 480.00) rappel = 0.800 precision = 0.439 f-mesure = 0.567
- classe P (attendus = 677, ramenes = 114.00, corrects = 40.00) rappel = 0.059 precision = 0.351 f-mesure = 0.101
- sur l'ensemble des 4 classes macro rappel = 0.439 macro precision = 0.387 macro F-mesure = 0.412

TABLE 17 – Fichier évalué : `equipe_3_tache_2_execution_3-testing.sh-title-only-ref-silence-noise-norm-aggr-eucl.prexml.xml`

Evaluation du pays
- classe F (attendus = 1153, ramenes = 985.00, corrects = 685.00) rappel = 0.594 precision = 0.695 f-mesure = 0.641
- classe Q (attendus = 1329, ramenes = 1497.00, corrects = 1029.00) rappel = 0.774 precision = 0.687 f-mesure = 0.728
- sur l'ensemble des 2 classes macro rappel = 0.684 macro precision = 0.691 macro F-mesure = 0.688
Evaluation du journal
- classe D (attendus = 652, ramenes = 0.00, corrects = 0.00) rappel = 0.000 precision = 0.000 f-mesure = 0.000
- classe E (attendus = 553, ramenes = 0.00, corrects = 0.00) rappel = 0.000 precision = 0.000 f-mesure = 0.000
- classe M (attendus = 600, ramenes = 985.00, corrects = 468.00) rappel = 0.780 precision = 0.475 f-mesure = 0.591
- classe P (attendus = 677, ramenes = 1497.00, corrects = 514.00) rappel = 0.759 precision = 0.343 f-mesure = 0.473
- sur l'ensemble des 4 classes macro rappel = 0.385 macro precision = 0.205 macro F-mesure = 0.267

Acknowledgments / Remerciements

Nous tenons à remercier chaleureusement tous ceux et celles qui ont contribué de près ou de loin à la réalisation de ce travail, ils se reconnaîtront : Comité d'organisation : Dominic Forest, Cyril Grouin, Lyne Da Sylva ; Les distributeurs de corpora : Cedrom-SNi, ELDA, Gallica, CNRTL ; Co-créateurs origineaux du MARF : Stephen Sinclair, Ian Clément, Dimitrios Nicolacopoulos et les contributeurs du MARF R&D Group ; Drs. Leila Kosseim, Sabine Bergler, Ching Y. Suen, Joey Paquet, Mourad Debbabi ; Michelle Khalifé ; CRSNG ; La Faculté de Génie et Informatique, Université Concordia.

Références

ABDI H. (2007). Distance. In N. J. SALKIND, Ed., *Encyclopedia of Measurement and Statistics*, Thousand Oaks (CA) : Sage.

BERNSEE S. M. (1999–2005). The DFT “à pied” : Mastering the Fourier transform in one day. [online]. http://www.dspdimension.com/data/html/dftapied.html.

CAVALIN P. R., SABOURIN R. & SUEN C. Y. (2010). Dynamic selection of ensembles of classifiers using contextual information. In *Multiple Classifier Systems*, LNCS 5997, p. 145–154.

FOREST D., GROUIN C., SYLVA L. D. & DEFT (2010). Campagne DÉfi Fouille de Textes (DEFT) 2010. [online], http://www.groupes.polymtl.ca/taln2010/deft.php.

GARCIA E. (2006). Cosine similarity and term weight tutorial. [online]. http://www.miislita.com/information-retrieval-tutorial/cosine-similarity-tutorial.html.

GROUIN C. (2010a). Corpus d’archives de presse (1800–1944). [online]. Source : Gallica, BNF, http://gallica.bnf.fr/.

GROUIN C. (2010b). Corpus presse francophone (1999, 2002, 2003) de La Presse, la campagne DEFT2010. [online]. Source : Cedrom-SNi, http ://www.cedrom-sni.com/.

GROUIN C. (2010c). Corpus presse francophone (1999, 2002, 2003) de Le Devoir, la campagne DEFT2010. [online]. Source : Cedrom-SNi, http://www.cedrom-sni.com/.

GROUIN C. (2010d). Corpus presse francophone (1999, 2002, 2003) de Le Monde, la campagne DEFT2010. [online]. Source : ELDA, http ://www.elda.org/.

GROUIN C. (2010e). Corpus presse francophone (1999, 2002, 2003) de L’Est Républicain, la campagne DEFT2010. [online]. Source : CNRTL, ATILF (CNRS), http://www.cnrtl.fr/corpus/estrepublicain/.

HAMMING R. W. (1950). Error detecting and error correcting codes. *Bell System Technical Journal*, **26**(2), 147–160. See also http://en.wikipedia.org/wiki/Hamming_distance.

HARIDAS S. (2006). Generation of 2-D digital filters with variable magnitude characteristics starting from a particular type of 2-variable continued fraction expansion. Master’s thesis, Department of Electrical and Computer Engineering, Concordia University, Montreal, Canada.

IFEACHOR E. C. & JERVIS B. W. (2002). *Speech Communications*. New Jersey, USA : Prentice Hall.

JURAFSKY D. S. & MARTIN J. H. (2000). *Speech and Language Processing*. Upper Saddle River, New Jersey 07458 : Prentice-Hall, Inc., Pearson Higher Education. ISBN 0-13-095069-6.

KHALIFÉ M. (2004). Examining orthogonal concepts-based micro-classifiers and their correlations with noun-phrase coreference chains. Master’s thesis, Department of Computer Science and Software Engineering, Concordia University, Montreal, Canada.

MAHALANOBIS P. C. (1936). On the generalised distance in statistics. In *Proceedings of the National Institute of Science of India 12*, p. 49–55. Online at http://en.wikipedia.org/wiki/Mahalanobis_distance.

MARTIN J. H. (2003). The CYK probabilistic parsing algorithm. [online] ; a book insert. http://www.cs.colorado.edu/~martin/SLP/New_Pages/pg455.pdf.

MOKHOV S., CLEMENT I., SINCLAIR S. & NICOLACOPOULOS D. (2002–2003). Modular Audio Recognition Framework. Department of Computer Science and Software Engineering, Concordia University, Montreal, Canada. Project report, http://marf.sf.net, last viewed April 2010.

MOKHOV S. A. (2008a). Choosing best algorithm combinations for speech processing tasks in machine learning using MARF. In S. BERGLER, Ed., *Proceedings of the 21st Canadian AI’08*, p. 216–221, Windsor, Ontario, Canada : Springer-Verlag, Berlin Heidelberg. LNAI 5032.

MOKHOV S. A. (2008b). Study of best algorithm combinations for speech processing tasks in machine learning using median vs. mean clusters in MARF. In B. C. DESAI, Ed., *Proceedings of C3S2E’08*, p. 29–43, Montreal, Quebec, Canada : ACM. ISBN 978-1-60558-101-9.

MOKHOV S. A. (2008–2010). WriterIdentApp – Writer Identification Application. Unpublished.

MOKHOV S. A. (2010a). Complete complimentary results report of the MARF's NLP approach to the DEFT 2010 competition. [online]. `http://arxiv.org/abs/1006.3787`.

MOKHOV S. A. (2010b). Cryptolysis : A Framework for Verification of Optimization Heuristics for the Automated Cryptanalysis of Classical Ciphers and Natural Language Word Segmentation. In *Proceedings of SERA 2010*, p. 295–302 : IEEE Computer Society.

MOKHOV S. A. (2010c). Evolution of MARF and its NLP framework. In *Proceedings of C3S2E'10*, p. 118–122 : ACM.

MOKHOV S. A. & DEBBABI M. (2008). File type analysis using signal processing techniques and machine learning vs. `file` unix utility for forensic analysis. In O. GOEBEL, S. FRINGS, D. GUENTHER, J. NEDON & D. SCHADT, Eds., *Proceedings of the IT Incident Management and IT Forensics (IMF'08)*, p. 73–85, Mannheim, Germany : GI. LNI140.

MOKHOV S. A., SINCLAIR S., CLEMENT I., NICOLACOPOULOS D. & THE MARF RESEARCH & DEVELOPMENT GROUP (2002–2010). SpeakerIdentApp – Text-Independent Speaker Identification Application. Published electronically within the MARF project, `http://marf.sf.net`. Last viewed February 2010.

MOKHOV S. A., SONG M. & SUEN C. Y. (2009). Writer identification using inexpensive signal processing techniques. In T. SOBH & K. ELLEITHY, Eds., *Innovations in Computing Sciences and Software Engineering ; Proceedings of CISSE'09*, p. 437–441 : Springer. ISBN : 978-90-481-9111-6, online at : `http://arxiv.org/abs/0912.5502`.

MOKHOV S. A. & THE MARF RESEARCH & DEVELOPMENT GROUP (2003–2010a). LangIdentApp – Language Identification Application. Published electronically within the MARF project, `http://marf.sf.net`. Last viewed February 2010.

MOKHOV S. A. & THE MARF RESEARCH & DEVELOPMENT GROUP (2003–2010b). ProbabilisticParsingApp – Probabilistic NLP Parsing Application. Published electronically within the MARF project, `http://marf.sf.net`. Last viewed February 2010.

MOKHOV S. A. & VASSEV E. (2009). Leveraging MARF for the simulation of the securing maritime borders intelligent systems challenge. In *Proceedings of the Huntsville Simulation Conference (HSC'09)* : SCS. To appear.

NIKSIC H. & FREE SOFTWARE FOUNDATION, INC. (1996–2009). `wget` – the non-interactive network downloader. [online]. `http://www.gnu.org/software/wget/manual/wget.html`, last viewed June 2010.

O'SHAUGHNESSY D. (2000). *Speech Communications*. New Jersey, USA : IEEE.

PRESS W. H. (1993). *Numerical Recipes in C*. Cambridge, UK : Cambridge University Press, second edition.

S. J. RUSSELL & P. NORVIG, Eds. (1995). *Artificial Intelligence : A Modern Approach*. New Jersey, USA : Prentice Hall. ISBN 0-13-103805-2.

THE MARF RESEARCH AND DEVELOPMENT GROUP (2002–2010). The Modular Audio Recognition Framework and its Applications. [online]. `http://marf.sf.net` and `http://arxiv.org/abs/0905.1235`, last viewed April 2010.

THE SPHINX GROUP AT CARNEGIE MELLON (2007–2010). The CMU Sphinx group open source speech recognition engines. [online]. `http://cmusphinx.sourceforge.net`.

VAILLANT P., NOCK R. & HENRY C. (2006). Analyse spectrale des textes : détection automatique des frontières de langue et de discours. In *Verbum ex machina : Actes de la 13eme conference annuelle sur le Traitement Automatique des Langues Naturelles (TALN 2006)*, p. 619–629. Online at `http://arxiv.org/abs/0810.1212`.