# LIF at TAC Multiling: Towards a Truly Language Independent Summarizer

**Firas Hmida**
Luminy CS Dept.
Aix-Marseille Univ., France
firas.hmida@gmail.com

**Benoit Favre**
LIF/CNRS
Aix-Marseille Univ., France
benoit.favre@lif.univ-mrs.fr

## Abstract

This paper presents the LIF system for the TAC'2011 Multilingual pilot track. We followed a language-independent approach to summarization for this task. In particular, we tried to remove the following dependences to language: sentence segmentation, word segmentation, stop-word lists, and word-level relevance assessment. We applied these modifications to an MMR-based system and observed little degradation on English data. The submitted system had a bug that impeded all official results, therefore we propose in this paper an updated set of results with relevant analysis.

## 1 Introduction

The aim of the TAC Multilingual summarization pilot is to evaluate the performance of automatic summarizers on a range of languages. The task can either be tackled by adapting a system to each language or designing a language-independent system that will not need further adaptation for processing new languages. We follow the later approach and try to design a system that removes as many assumptions about the language of the documents being processed.

We have identified several language-dependent factors in existing summarization systems.

- *Sentence segmentation*: most extractive summarizers rely on sentence boundaries, that can be detected using punctuation and abbreviation lists.
- *Tokenization*: rules to split a sequence of words are language independent, an can be non trivial, as for instance in Chinese.
- *Relevance assessment*: relies mostly on comparing word frequency histograms, and involves language-dependent *stop-word lists*.
- *Morphological, syntactic and semantic analysis, information extraction*: high-level processing often requires dictionaries and corpora to train the analysers.
- *Training data*: any supervised summarizer needs training data from the same language as used in the documents to be summarized.

In this work, we focus on the notion of word, the notion of sentence and stop-word lists. We modify a system based on the Maximal Marginal Relevance (MMR) algorithm (Carbonell and Goldstein, 1998) in order to remove or reduce those factors. In particular, sentence segmentation is replaced by a crude heuristic: use the last character of the input documents as splitting point; the need for word tokenization is replaced by using character n-grams to represent the content of sentences (Damashek, 1995); stop-word lists are not needed because we only include character sequences of length n, with n long-enough to cover multiple words relevance is assessed at the n-gram level instead of the word level using the unmodified *cosine/tfidf* framework.

After presenting related work (Section 2), the paper exposes our methodology and in particular our system for TAC'2011 (Section 3) and discusses the

results of the evaluation on both older TAC datasets and the multilingual pilot (Section 4).

## 2 Previous work

The work in extractive summarization has mostly revolved around sentence selection algorithms: non-redundant relevance maximization (Carbonell and Goldstein, 1998; Radev et al., 2004), graph-based ranking (Mihalcea and Tarau, 2004), global inference (McDonald, 2007), probabilistic topic-driven modeling (Chang and Chien, 2009), or event-based scoring (Filatova and Hatzivassiloglou, 2004). These well-known methods have in common that they all assume the availability of a sentence segmentation and that stop words (non-content words) are removed before processing.

Supervised methods for summarization (Kupiec et al., 1995; Shen et al., 2007; Berg-Kirkpatrick et al., 2011), that make most of TAC submissions, directly learn to select sentences. They require training data in the same language as the documents being summarized. In addition, they often take advantage of features extracted from high-level information extracted from the documents, such as syntactic trees (Siddharthan et al., 2004), entities and relations (White et al., 2001) or sentiments (Titov and McDonald, 2008), which rely on supervised systems that also need training data.

However, there has been work on unsupervised natural language processing in a variety of areas useful for summarization: Chung and Gildea (2009) take advantage of bilingual text alignments to perform unsupervised tokenization; Kiss and Strunk (2006) proposed a method for unsupervised sentence segmentation but that method still assumes the knowledge of punctuation marks; Unsupervised part-of-speech tagging and grammar induction are also mature fields (Goldwater and Griffiths, 2007; Cohen et al., 2010); named entities (Cucerzan and Yarowsky, 1999) can be detected in an unsupervised manner. These methods are promising but they generally require large in-domain datasets in the target language and they underperform their supervised counterparts.

In this work, we propose simple methods for removing the need for language-dependent components in summarization systems. We follow the tracks of Mihalcea (2005) who proposed a language-independent method based on graphs using the principle of ranking algorithms. In their work, the importance of a sentence is derived through its influence in the graph of all sentences. The method, however, depends on tokenization, stop-word lists and sentence boundaries. Boudin et al. (2011) extend their approach and and propose an approach using a similarity measure which captures the similarity of sentences in term of morphology. They combine the similarity between two sentences with a measure of the longest common substring (LCS), to represent the similarity between vertices in the graph. They also extend the LCS formula (LCS*) in order to minimize intra-summary redundancy. Their approach is evaluated on corpora in English, Spanish and French.

In order to remove the problem of tokenization, instead of considering LCS, we look at character n-grams which were first proposed for comparing texts in term of topics by Damashek (1995). We also remove the need for stop-word lists by only considering long n-grams, likely to span multiple words, and therefore reducing the effect of stop-words. The topic of automatic stop-word removal was also tackled by Makrehchi and Kamel (2008) and Darling and Song (2011).

## 3 Method

This section details our system for the TAC 2011 Multilingual summarization pilot.

### 3.1 Maximal Marginal Relevance

Maximal Marginal Relevance (MMR) is a greedy method to extract iteratively the most relevant sentences relative to a query to generate a summary, while minimizing redundancy. At each iteration, the sentence added to the selection maximizes the similarity to the query while minimizing the similarity to sentences already selected. The algorithm stops when the 250 word length constraint is met. For query, we use the centroid of the cluster of documents being summarized. The MMR algorithm is summarized in Algorithm 1.

The $cosine$ similarity used for assessing relevance and redundancy in MMR is defined as the fol-

lowing formula:

$$cosine(a,b) = \frac{\sum_i a_i b_i}{||a|| \times ||b||} \quad (1)$$

Here, $a$ and $b$ are vector representations of the sentences using $tf.idf$ weights defined in the following sections.

**Data**: $S_0 = sentences, i = 0$
**Result**: $M_0 = \emptyset$
**while** $\sum_{s \in M_i} length(s) < 250$ **do**
  $\alpha(s) = cosine(s, q)$
  $\beta(s) = max_{r \in M_{i-1}} cosine(s, r)$
  $\hat{s} = argmax_{s \in S_i} [\lambda\alpha(s) - (1-\lambda)\beta(s)]$
  $M_i = M_{i-1} \cup \hat{s}$
  $S_i = S_{i-1} \setminus \hat{s}$
  $i = i + 1$
**end**

**Algorithm 1:** The MMR algorithm. Inputs a set of sentences $S_0$ and outputs a set of selected sentences $M_i$ that respects the length constraint. $cosine(\cdot)$ is the cosine similarity and $\lambda$ is a trade-off parameter between relevance and redundancy that has to be set on a held-out dataset.

### 3.2 Sentence segmentation

The problem of automatic detection of sentence arises because of the ambiguity of certain punctuation marks. The characters that represent punctuation marks are language dependent, and therefore, we propose to use a crude heuristic: consider the last character of text to be summarized as the punctuation mark that indicates the limit between sentences.

### 3.3 Tokenization

In order to reduce the dependence to language, we ignore the notion of word and we use n-grams of characters as tokens to represent sentences. In the following we call *term* n-grams of all characters in a text.

### 3.4 Importance of words

The discriminative power of terms in the context of a similarity measure between sentences depends on their relative importance to all documents. We used the metric proposed in *tfidf* (Salton and Buckley, 1988) to determine the importance of a term in a set of documents without use of external resources. In particular, we use the following $tf.idf$ weighting scheme:

$$tf(w_i, s) = \frac{|w_i|_s}{\sum_k |w_k|_s} \quad (2)$$

$$idf(w_i) = \frac{|S|}{|s : w_i \in s|} \quad (3)$$

where $|\cdot|_s$ is the number of times a word is seen in sentence $s$, and $|S|$ is the number of sentences to be summarized.

## 4 Evaluation

When developing the system, we performed a first experiment on data from the TAC 2008 and TAC 2009 evaluation campaigns. The corpus consists of 48 topics for TAC 2008 and 44 topics for TAC 2009. For each topic, a 100-word summary has to be generated. The corpus is a collection of news articles in English from different sources: AFP, NYT, APW, LTW, and Xinhua. The $\lambda$ parameter and the size of n-grams are adjusted on TAC 2008 and evaluation is performed on TAC 2009. No language-dependent processing is applied to any of the input documents.

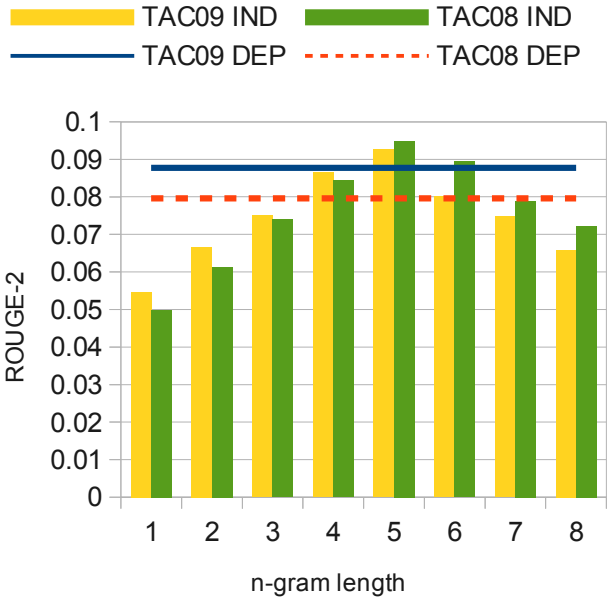


Figure 1: Effect of n-gram length on the TAC 2008 and 2009 datasets. The language-independent system (IND) is compared to the language-dependent (DEP) topline.

Figure 1 presents the average rating by ROUGE-2, of the language-independent MMR systems,

where n-gram size is variated. In addition, a regular language-dependent MMR system is provided as a topline. The evaluation process consists on comparing (using ROUGE), the summaries generated by each system one by one, with the human-written references and deduce an average value. Note that the system running with 5-grams is the most successful among the proposed systems. This implies that the terms containing five characters are more significant (informative) than other sizes of n-grams. In particular, shorter n-grams might not be specific enough to represent meaning from sentences while longer n-grams might be too specific and the overlap between sentences be empty. In fact, this result was to be expected since five is close to the average word length in English.

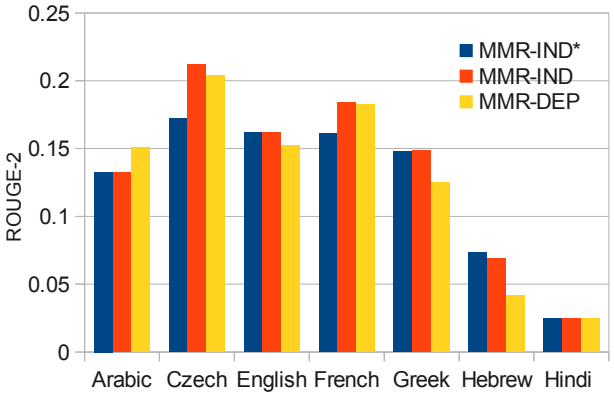


Figure 2: ROUGE-2 results on TAC 2011 by language. MMR-IND* is the buggy submission, MMR-IND is the corrected system and MMR-DEP is the topline.

It can also be observed that the topline performs at lower ROUGE-2 scores than the language-independent system for a well chosen n. This can be explained by the better generalization capabilities of using character n-grams than words. Also, n-grams are less subject to the effect of stop-words and they tend to give credit multiple times to long sequences of characters that appear in many sentences, which single words do not.

Unfortunately, our official submission contained a bug that enabled material from one topic to be output in a summary of a different topic. This bug dramatically reduced the perceived value of our summaries during human evaluation, which lead our system (ID4) to get consistently bad scores. Nevertheless, we removed that bug after the evaluation and show updated ROUGE-2 results even though we did not have the resources to rerun a human evaluation. Table 2 shows the results from the system with the bug (MMR-IND*), the corrected system (MMR-IND) and the language-dependent topline (MMR-DEP). This figure shows that language-dependence is not necessary as the language independent system performs at the same level of performance. It also shows that the bug mostly affected our Czech and French submissions. Interestingly, for some languages like Greek and Hebrew, MMR-DEP is significantly worse than MMR-IND while for Arabic it is better. It is also interesting to see how ROUGE results can vary from language to language even though all the reference summaries were produced by humans in the same conditions. This reveals that some effect can be expected from the nature of lanugage itself when comparing texts using word bigrams.

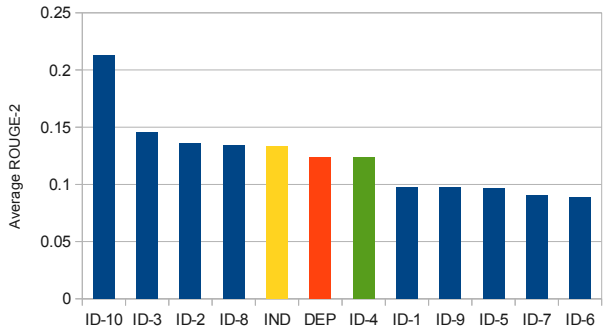


Figure 3: Average system-level ROUGE-2 on TAC'11. ID4 is the buggy submission, IND is the corrected system and DEP is the language-dependent system.

Figure 3 compares ROUGE-2 results for our system against other participants, averaged over languages, counting zero when a system did not include a submission for that language (effectively penalizing non-multilingual systems). It shows that our submission was average compared to the systems, worse than ID10, the human topline and better than ID9, the baseline of the evaluation. Correcting the bug did bring a small improvement in term of ROUGE even though we would have expected a much larger difference given the degradation seen on human evaluation. We believe that this reflects the fact that ROUGE does not assess summaries in term of perceived quality but rather in term of word overlap with human-written summaries.

## 5 Conclusion and outlook

We have presented in this paper our submission to the TAC 2011 Multilingual summarization pilot track. In order to design truly language-independent summarization systems, a number of subtasks must be language-independent. In our system, we proposed to use a simple heuristic for sentence segmentation, then use character n-grams to remove the need for tokenization and stop-word lists. While a bug impaired our system at the official manual evaluation, it showed promising results in automatic evaluations, suggesting to introduce more unsupervised linguistic processing submodules (such as unsupervised parsing), and try to modify high performance summarization algorithms that lead to the best systems in the regular TAC tracks in order to remove their dependency to language. We hope that the multilingual summarization pilot will bring focus on this hard task and foster great research in the area.

## References


T. Berg-Kirkpatrick, D. Gillick, and D. Klein. Jointly learning to extract and compress. In *Proc. of ACL*, 2011.

Florian Boudin, Stéphane Huet, and Juan-Manuel Torres-Moreno. A graph-based approach to cross-language multi-document summarization. In *Conference on Intelligent Text Processing and Computational Linguistics (CICLing)*, 2011.

J. Carbonell and J. Goldstein. The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 335–336. ACM, 1998.

Y.L. Chang and J.T. Chien. Latent dirichlet learning for document summarization. In *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*, pages 1689–1692. IEEE, 2009.

T. Chung and D. Gildea. Unsupervised tokenization for machine translation. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2-Volume 2*, pages 718–726. Association for Computational Linguistics, 2009.

S.B. Cohen, D.M. Blei, and N.A. Smith. Variational inference for adaptor grammars. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 564–572. Association for Computational Linguistics, 2010.

S. Cucerzan and D. Yarowsky. Language independent named entity recognition combining morphological and contextual evidence. In *Proceedings of the 1999 Joint SIGDAT Conference on EMNLP and VLC*, pages 90–99, 1999.

M. Damashek. Gauging similarity with n-grams: Language-independent categorization of text. *Science*, 267(5199):843, 1995.

W.M. Darling and F. Song. Probabilistic document modeling for syntax removal in text summarization. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2*, pages 642–647. Association for Computational Linguistics, 2011.

E. Filatova and V. Hatzivassiloglou. Event-based extractive summarization. In *Proceedings of ACL Workshop on Summarization*, volume 111, 2004.

S. Goldwater and T. Griffiths. A fully bayesian approach to unsupervised part-of-speech tagging. In *ANNUAL MEETING-ASSOCIATION FOR COMPUTATIONAL LINGUISTICS*, volume 45, page 744, 2007.

T. Kiss and J. Strunk. Unsupervised multilingual sentence boundary detection. *Computational Linguistics*, 32(4):485–525, 2006.

J. Kupiec, J. Pedersen, and F. Chen. A trainable document summarizer. In *Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 68–73. ACM, 1995.

M. Makrehchi and M. Kamel. Automatic extraction of domain-specific stopwords from labeled documents. *Advances in information retrieval*, pages 222–233, 2008.

R. McDonald. A study of global inference algorithms in multi-document summarization. *Advances in Information Retrieval*, pages 557–564, 2007.

R. Mihalcea. Language independent extractive summarization. In *Proceedings of the ACL 2005 on Interactive poster and demonstration sessions*, pages 49–52. Association for Computational Linguistics, 2005.

R. Mihalcea and P. Tarau. Textrank: Bringing order into texts. In *Proceedings of EMNLP*, volume 4, pages 404–411. Barcelona: ACL, 2004.

D.R. Radev, H. Jing, M. Sty, and D. Tam. Centroid-based summarization of multiple documents. *Information Processing & Management*, 40(6):919–938, 2004.

G. Salton and C. Buckley. Term-weighting approaches in automatic text retrieval* 1. *Information processing & management*, 24(5):513–523, 1988.

D. Shen, J.T. Sun, H. Li, Q. Yang, and Z. Chen. Document summarization using conditional random fields. In *Proceedings of IJCAI*, volume 7, pages 2862–2867, 2007.

A. Siddharthan, A. Nenkova, and K. McKeown. Syntactic simplification for improving content selection in multi-document summarization. In *Proceedings of the 20th international conference on Computational Linguistics*, pages 896–es. Association for Computational Linguistics, 2004.

I. Titov and R. McDonald. A joint model of text and aspect ratings for sentiment summarization. *Urbana*, 51:308–316, 2008.

M. White, T. Korelsky, C. Cardie, V. Ng, D. Pierce, and K. Wagstaff. Multidocument summarization via information extraction. In *Proceedings of the first international conference on Human language technology research*, pages 1–7. Association for Computational Linguistics, 2001.