Matching Texts with SUMMA

Horacio Saggion
TALN
Department of Information and Communication Technologies
Universitat Pompeu Fabra
C/Tanger 122
Barcelona - 08018
Spain
horacio.saggion@upf.edu

Résumé. On décrit notre approche au problème de l'appariement de résumés/articles scientifiques proposé par le programme DÉfi Fouille de Textes (DEFT). Nous avons développé un algorithme d'appariement de textes qui utilise des ressources quasiment indépendantes de la langue. L'algorithme crée des representations de documents tout en utilisant le système SUMMA et les compare grâce à une mesure de similarité cosinus qui nous permet de sélectionner le meilleure candidat pour former la paire. Nos résultats indiquent que cette approche est très précise et qu'elle pourrait s'appliquer à d'autres langues.

Abstract. We describe our solution to the abstract-document matching problem proposed in the DÉfi Fouille de Textes (DEFT) evaluation programme. We have developed a text matching algorithm using quasi language independent resources. Our algorithm creates document representations using the SUMMA system and compares representations using a cosine similarity measure selecting the best matching candidate. Results indicate that the solution is highly accurate and could be applied to other languages.

Mots-clés : Système SUMMA, résumé automatique, similarité textuelle.

Keywords: SUMMA System, Text Summarization, Text Similarity.

1 Introduction

The DÉfi Fouille de Textes (DEFT) evaluation programme focuses on natural language processing technologies for text mining problem solving in the French language. Different challenges have been put forward in previous editions of DEFT such as that of opinion mining (Grouin *et al.*, 2009) or text classification (Grouin *et al.*, 2008).

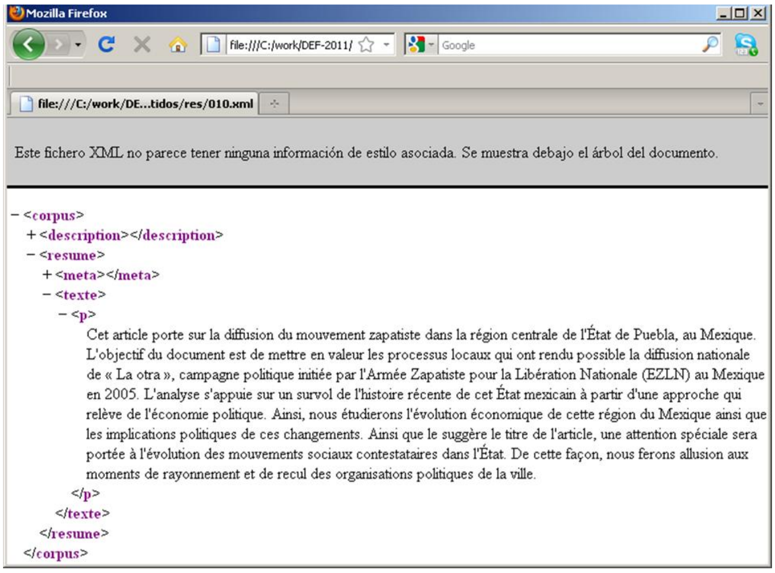
The DEFT 2011 evaluation chapter proposed two different text mining tasks for participating teams : (i) identify the publication year of a given French article and (ii) identify the source document (out of a pool of documents) for a given text abstracts ("abstract document matching" problem).

In our first participation in DEFT, we concentrated on the abstract-document matching problem only. The data set for the abstract-document matching problem is a set of scientific articles and their abstracts. These were published in reviews in the field of humanities. The corpus was transformed into the following three components :

- *ART* : the set of articles with their author abstracts removed ;
- *RES* : the set of author's abstracts ;
- *TXT* : same as ART but where introduction and conclusion sections have been removed from the articles.

Information about article's authors and titles was removed. Examples of matching documents are shown in Figures 1 (abstract), 2 (article), and 3 (article w/o introduction/conclusion).

Two subtasks were proposed for the abstract-document matching problem : (i) identify for each abstract in *RES* the article in *ART* where the abstract comes from, and (ii) identify for each abstract in *RES* the text in *TXT* where



Mozilla Firefox
file:///C:/work/DEF-2011/
Google
file:///C:/work/DE...tidos/res/010.xml

Este fichero XML no parece tener ninguna información de estilo asociada. Se muestra debajo el árbol del documento.

```
- <corpus>
  + <description></description>
  - <resume>
    + <meta></meta>
    - <texte>
      - <p>
          Cet article porte sur la diffusion du mouvement zapatiste dans la région centrale de l'État de Puebla, au Mexique.
          L'objectif du document est de mettre en valeur les processus locaux qui ont rendu possible la diffusion nationale
          de « La otra », campagne politique initiée par l'Armée Zapatiste pour la Libération Nationale (EZLN) au Mexique
          en 2005. L'analyse s'appuie sur un survol de l'histoire récente de cet État mexicain à partir d'une approche qui
          relève de l'économie politique. Ainsi, nous étudierons l'évolution économique de cette région du Mexique ainsi que
          les implications politiques de ces changements. Ainsi que le suggère le titre de l'article, une attention spéciale sera
          portée à l'évolution des mouvements sociaux contestataires dans l'État. De cette façon, nous ferons allusion aux
          moments de rayonnement et de recul des organisations politiques de la ville.
        </p>
      </texte>
    </resume>
  </corpus>
```

FIG. 1 – Abstract used during training

the abstract comes from. This problem is related to the abstract-descriptors matching problem we have proposed for manual evaluation of automatic text summarization systems (Saggion & Lapalme, 2002).

In DEFT, in addition to providing a single answer per abstract, there was the possibility of providing a set of matching documents, each with an associated confidence score where the confidence scores for a given abstracts have to add up to 1 (i.e., a probability distribution).

We developed our “abstract to text” matching solution in a very short period of time taking advantage of our SUMMA toolkit (Saggion, 2008a) which can be used to compare different document representations and which has been used successfully in previous evaluation campaign such as multi-document summarization (Saggion & Gaizauskas, 2004) or text clustering (Saggion, 2008a) which require document representations to be compared.

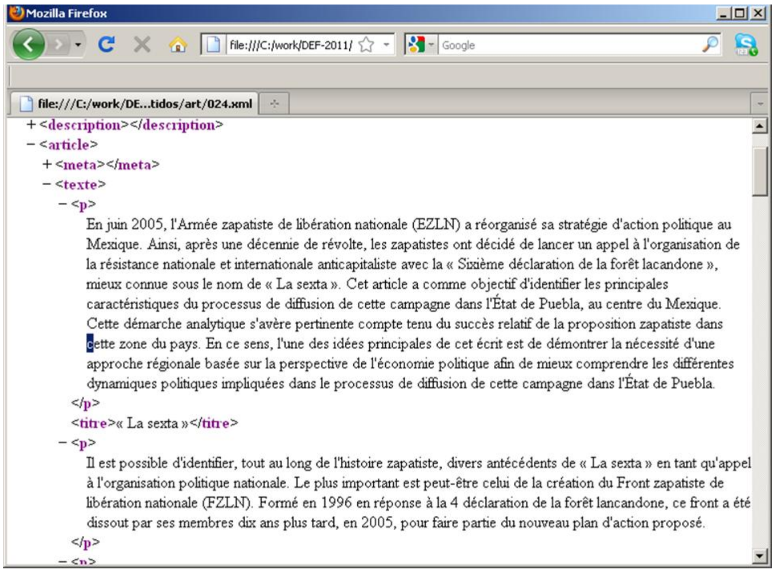
In the rest of this short communication we first describe how we have developed our system using available components and how we propose a solution (Section 2), we then describe the evaluation framework and results we have obtained (Section 3), and finally close the paper with a discussion and outlook (Section 4).

2 Document Processing

For basic document processing we have used functionalities available in the SUMMA toolkit (Saggion, 2008b) and GATE system (Maynard *et al.*, 2002).

2.1 SUMMA Functionalities in DEFT

SUMMA operates on GATE document representations and computes feature values for different text units. For example, for text summarization it computes different features for sentences to measure sentence relevance (i.e., sentence position, sentence-title similarity, sentence-document centroid similarity, etc.) ; these are implemented components (e.g., processing resources) which can be integrated in stand-alone applications. Features and annotations computed in SUMMA are added to the document representations and in this way passed from one module to the next one. The main SUMMA components used in DEFT were : (i) a module to create corpus statistics, (ii) a


Mozilla Firefox
file:///C:/work/DEF-2011/
Google
file:///C:/work/DE...tidos/art/024.xml

```
+ <description></description>
− <article>
  + <meta></meta>
  − <texte>
    − <p>
        En juin 2005, l'Armée zapatiste de libération nationale (EZLN) a réorganisé sa stratégie d'action politique au
        Mexique. Ainsi, après une décennie de révolte, les zapatistes ont décidé de lancer un appel à l'organisation de
        la résistance nationale et internationale anticapitaliste avec la « Sixième déclaration de la forêt lacandone »,
        mieux connue sous le nom de « La sexta ». Cet article a comme objectif d'identifier les principales
        caractéristiques du processus de diffusion de cette campagne dans l'État de Puebla, au centre du Mexique.
        Cette démarche analytique s'avère pertinente compte tenu du succès relatif de la proposition zapatiste dans
        cette zone du pays. En ce sens, l'une des idées principales de cet écrit est de démontrer la nécessité d'une
        approche régionale basée sur la perspective de l'économie politique afin de mieux comprendre les différentes
        dynamiques politiques impliquées dans le processus de diffusion de cette campagne dans l'État de Puebla.
      </p>
      <titre>« La sexta »</titre>
    − <p>
        Il est possible d'identifier, tout au long de l'histoire zapatiste, divers antécédents de « La sexta » en tant qu'appel
        à l'organisation politique nationale. Le plus important est peut-être celui de la création du Front zapatiste de
        libération nationale (FZLN). Formé en 1996 en réponse à la 4 déclaration de la forêt lancandone, ce front a été
        dissout par ses membres dix ans plus tard, en 2005, pour faire partie du nouveau plan d'action proposé.
      </p>
    − <p>
```

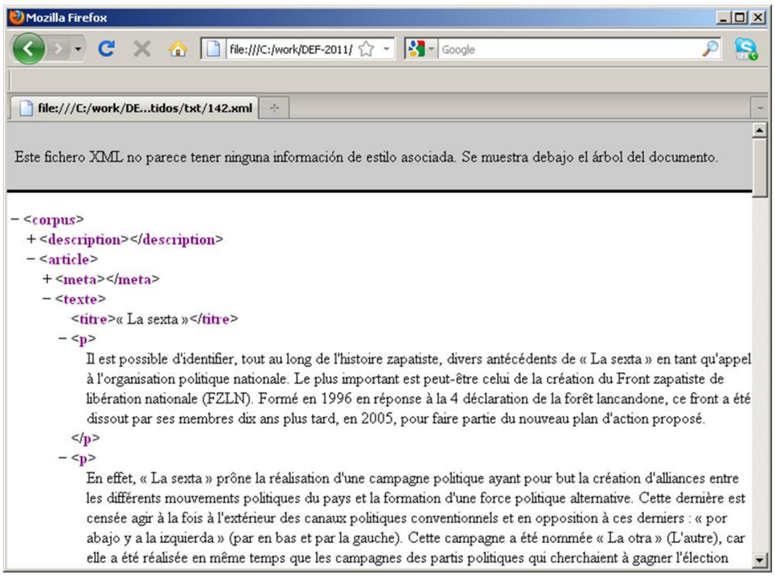
FIG. 2 – Article used during training

module to compute word document statistics, (iii) a flexible module to compute document vectors, and (iv) a module to compute document-document similarity measures. The DEFT algorithm is very simple and uses SUMMA as a Java library.

2.2 Processing Steps

Each document collection (ART, RES, TXT) was processed as illustrated in Figure 4 :

- First the collection of documents is transformed into XML format (for DEFT this was a simple document renaming procedure necessary for subsequent processes) and saved to data stores for more efficient processing since GATE needs considerable amount of memory for the documents ;
- Second, a basic document processing step takes place to identify different types of words in the document. In order to identify words in the documents we used a default tokenizer available in the GATE system. We have tried to use the French language tools from GATE (e.g. named recognition, etc.), but we finally gave up because these are very limited, containing mostly English resources which are of little help in the analysis of French language ;
- Third, an inverted document frequency table is created based on the set of documents to be processed (i.e., there is a table computed for each document collection). The table is created with functionalities available in the SUMMA summarization toolkit. Given a corpus of tokenized documents, an inverted document frequency table is created and stored to disk. The value of inverted document frequency for term t is $idf(t) = log(N+1/M_t+1)$ where M_t is the number of documents containing t and N is the number of documents in the collection ;
- Finally, a vector representation for each document in the collection is created. We also use SUMMA which implements the vector space model for this purpose (Salton, 1988). The tool computes token statistics including term frequency – the number of times each term occurs in the document (tf). Each vector contains for each term occurring in the text fragment, the value tf(t)*idf(t) (term frequency * inverted document frequency for term t). The vector creation component in SUMMA is flexible enough to allow specification of types of tokens to exclude, or a list of stop words to ignore, etc. For the experiments reported here we have not included punctuations, symbols, or numerical tokens in the vector representations. Figure 5 presents a document analysed with the tools and shows two vectors computed for the document : a tf*idf vector and a normalized tf*idf vector where the normalized values are obtained dividing the tf*idf of each term by the vector's Euclidean norm.


Mozilla Firefox
file:///C:/work/DEF-2011/
Google
file:///C:/work/DE...tidos/txt/142.xml

Este fichero XML no parece tener ninguna información de estilo asociada. Se muestra debajo el árbol del documento.

```
− <corpus>
  + <description></description>
  − <article>
    + <meta></meta>
    − <texte>
        <titre>« La sexta »</titre>
      − <p>
          Il est possible d'identifier, tout au long de l'histoire zapatiste, divers antécédents de « La sexta » en tant qu'appel
          à l'organisation politique nationale. Le plus important est peut-être celui de la création du Front zapatiste de
          libération nationale (FZLN). Formé en 1996 en réponse à la 4 déclaration de la forêt lancandone, ce front a été
          dissout par ses membres dix ans plus tard, en 2005, pour faire partie du nouveau plan d'action proposé.
        </p>
      − <p>
          En effet, « La sexta » prône la réalisation d'une campagne politique ayant pour but la création d'alliances entre
          les différents mouvements politiques du pays et la formation d'une force politique alternative. Cette dernière est
          censée agir à la fois à l'extérieur des canaux politiques conventionnels et en opposition à ces derniers : « por
          abajo y a la izquierda » (par en bas et par la gauche). Cette campagne a été nommée « La otra » (L'autre), car
          elle a été réalisée en même temps que les campagnes des partis politiques qui cherchaient à gagner l'élection
```

FIG. 3 – Article w/o introduction/conclusion used during training

2.3 Text Matching

Given two document collections $A = \{A_1, ..., A_n\}$ and $D = \{D_1, ..., D_n\}$ we carry out pair-wise comparison of documents in A with documents in D and create a similarity matrix (Figure 4) $M_{i,j}$ for $1 \leq i, j \leq n$ containing the similarity between document i (e.g., an abstract) and document j (e.g., a full article or article without introduction/conclusion). The similarity metric we use in this work to compare the documents is the cosine of the angle between two vectors. This metric gives value one for identical vectors and zero for vectors which are orthogonal (no related). The exact formula we use is as follows :

$$\text{cosine}(d_1, d_2) = \frac{\sum_{i=1}^{n} w_{i,d_1} * w_{i,d_2}}{\sqrt{\sum_{i=1}^{n} (w_{i,d_1})^2} * \sqrt{\sum_{i=1}^{n} (w_{i,d_2})^2}}$$

Here d_1 and d_2 are document vectors and w_{i,d_k} is the weight of term i in document d_k (i.e., the tf*idf values). For the experiment reported here we have used the normalized vectors.

System ID	Score
1 ; 5 ; 8 ; 11	1.000
6	0.995
2	0.980
3 ; 4	0.975
10	0.970
7	0.965
9	0.909

TAB. 1 – Results for systems matching abstracts (RES) with articles (ART)

2.4 Selecting a Candidate Summary

For each abstract A$_i$ in A we select a document D$_j$ in D such that $M_{i,j} \geq M_{i,k} \forall k$. For the response without confidence scores, we returned a set of pairs (Abstract$_i$,Document$_j$), for the response with confidence scores,

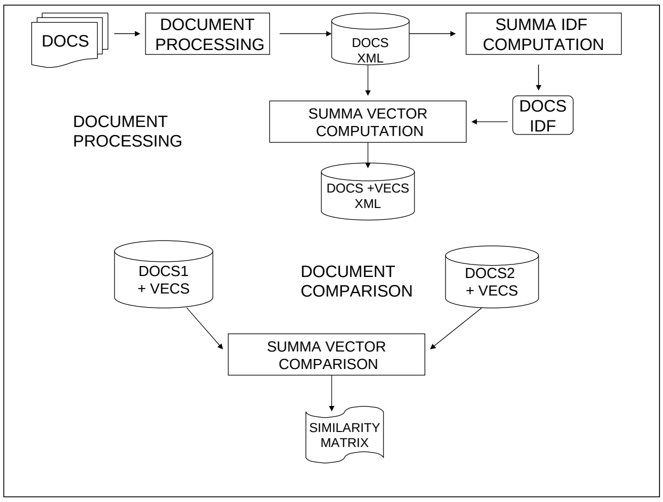


FIG. 4 – Document processing for DEFT 2011

System ID	Score
5 ; 8	1.000
11	0.995
6	0.990
4	0.964
1 ; 3	0.959
2	0.954
7	0.919
10	0.904
9	0.873

TAB. 2 – Results for systems matching abstracts (RES) with articles without introduction or conclusion (TXT)

we returned the best (in terms of similarity to the document) 5 triples (Abstract$_i$,Document$_{j_k}$,Confidence$_{i,j_k}$) for $k = 1, ...5$) for each abstract ; where Confidence$_{i,j_k}$ is the normalized similarity computed as :

$$\text{Confidence}_{i,j_k} = \frac{M_{i,j_k}}{\sum_{l=1,...5} M_{i,j_l}}$$

3 Experimental Results

Each system response (response$_i$ = *(Abstract$_i$,Document$_j$)*) is compared to the true response and evaluated according to the following formula :

$$\text{score}(\text{response}_i) = \begin{cases} 1 & \text{if the match is correct} \\ 0 & \text{otherwise} \end{cases}$$

The system final score is computed as :

$$\text{score(S)} = \frac{\sum_{i=0}^{n} \text{score}(\text{response}_i)}{n}$$

System ID	Score
3	0.512
10	0.417

TAB. 3 – Results for systems matching abstracts (RES) with articles (ART) with confidence scores

System ID	Score
3	0.402
10	0.368

TAB. 4 – Results for systems matching abstracts (RES) with articles w/o introduction/conclusion (TXT) with confidence scores

where n is the number of responses.

We also provided a set of 5 answers per abstract each with a confidence score as previously described (response$_{i,j_k}$ = *(Abstract$_i$, Document$_j$, Confidence$_{i,j_k}$)*). In this case the scores are computed as :

$$\text{confidence\_score}(\text{response}_{i,j_k}) = \begin{cases} \text{Confidence}_{i,j_k} & \text{if the match is correct} \\ 0 & \text{otherwise} \end{cases}$$

The system final confidence score is computed as :

$$\text{Confidence Score(S)} = \frac{\sum_{i=0}^{n} \text{confidence\_score}(\text{response}_{i,j})}{n}$$

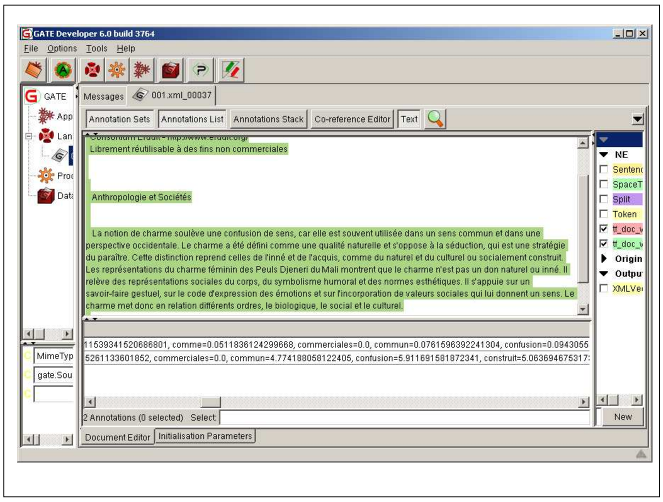
where n is the number of responses.

Tables 1 and 2 show respectively results for all systems for matching abstracts with full articles and abstracts with articles without introductions or conclusions. Systems with similar performance are grouped together. Our system ID is number 3 which obtains a reasonable score in both tasks being placed in the middle of both tables. Since the same strategy was used to select ART and TXT responses and the scores for TXT are lower, it appears that this latter task is slightly more difficult. It is however evident that all systems are able to solve the proposed task.

Where the response with confidence scores is concerned, obtained results are shown on Tables 3 and 4. In addition to our system (number 3) only other team provided answers with confidence scores (number 10). Results are much lower than those obtained for unique answers because of the low confidence given to the best matching solution : the first solutions is almost always the correct one and confidence computation should take into consideration this fact.

4 Discussion

This paper has described our first participation in the DEFT evaluation campaign. We have developed a solution able to identify the document where an abstract comes from with very high accuracy using available and *quasi* "language independent" tools, in fact only the tokenization process we have used is language dependent, no sophisticated resources such as morphological analysers or stemmers have otherwise been used. We therefore argue that our solution could be easily ported to similar tasks in other Romance languages and to English. Because most systems were able to obtain good accuracy we argue that the task has to be made harder by considering document collections which are closer in content and form or by considering the cross-lingual matching problem.



Consortium Erudit - http://www.erudit.org/
Librement réutilisable à des fins non commerciales

Anthropologie et Sociétés

La notion de charme soulève une confusion de sens, car elle est souvent utilisée dans un sens commun et dans une perspective occidentale. Le charme a été défini comme une qualité naturelle et s'oppose à la séduction, qui est une stratégie du paraître. Cette distinction reprend celles de l'inné et de l'acquis, comme du naturel et du culturel ou socialement construit. Les représentations du charme féminin des Peuls Djeneri du Mali montrent que le charme n'est pas un don naturel ou inné. Il relève des représentations sociales du corps, du symbolisme humoral et des normes esthétiques. Il s'appuie sur un savoir-faire gestuel, sur le code d'expression des émotions et sur l'incorporation de valeurs sociales qui lui donnent un sens. Le charme met donc en relation différents ordres, le biologique, le social et le culturel.

```
11539341520686801, comme=0.0511836124299668, commerciales=0.0, commun=0.0761596392241304, confusion=0.0943055
5261133601852, commerciales=0.0, commun=4.774188058122405, confusion=5.911691581872341, construit=5.063694675317:
```

2 Annotations (0 selected) Select:

FIG. 5 – Document with normalized and non-normalized vectors computed.

Acknowledgements

We are grateful to Programa Ramón y Cajal 2009 from Ministerio de Ciencia e Innovación, Spain. We are grateful to Simon Mille who helped us with the French abstract. Experiments have been done in the UPF-CIBER-BBN cluster for High Performance Computing, massive storage and software for Biomedical Applications which is part of the Biomedical Research Services of CIBER-BBN `http://www.ciber-bbn.es` and UPF `http://www.upf.edu`.

Références

GROUIN C., BERTHELIN J.-B., AYARI S. E., HURAULT-PLANTET M. & LOISEAU S. (2008). Présentation de deft'08 (défi fouille de textes). In *Actes de JEP–TALN–RECITAL 2008*. 13 juin 2008.

GROUIN C., HAURAULT-PLANTET M., PAROUBEK P. & BERTHELIN J.-B. (2009). Deft'07 : une campagne d'évaluation en fouille d'opinion. *RNTI*, **E-17**.

MAYNARD D., TABLAN V., CUNNINGHAM H., URSU C., SAGGION H., BONTCHEVA K. & WILKS Y. (2002). Architectural Elements of Language Engineering Robustness. *Journal of Natural Language Engineering – Special Issue on Robust Methods in Analysis of Natural Language Data*, **8**(2/3), 257–274.

SAGGION H. (2008a). Experiments on semantic-based clustering for cross-document coreference. In *Proceedings of the Third Joint International Conference on Natural Language Processing*, p. 149–156, Hyderabad, India : AFNLP AFNLP.

SAGGION H. (2008b). SUMMA : A Robust and Adaptable Summarization Tool. *Traitement Automatique des Langues*, **49**(2), 103–125.

SAGGION H. & GAIZAUSKAS R. (2004). Multi-document summarization by cluster/profile relevance and redundancy removal. In *Proceedings of the Document Understanding Conference 2004* : NIST.

SAGGION H. & LAPALME G. (2002). Generating Indicative-Informative Summaries with SumUM. *Computational Linguistics*.

SALTON G. (1988). *Automatic Text Processing*. Addison-Wesley Publishing Company.