CONTRIBUTION TO TOPIC IDENTIFICATION BY USING WORD SIMILARITY

Armelle Brun, Kamel Smaili and Jean-Paul Haton

Loria BP 239 54506 Vandoeuvre-Lès-Nancy, France - Tel: (+33)3-83-80-20-97, Fax: (+33)3-83-41-30-79 {brun, smaili, jph}@loria.fr

ABSTRACT

In this paper, a new topic identification method, WSIM, is investigated. It exploits the similarity between words and topics. This measure is a function of the similarity between words, based on the mutual information. The performance of WSIM is compared to the cache model and to the well-known SVM classifier. Their behavior is also studied in terms of recall and precision, according to the training size. Performance of WSIM reaches 82.4% correct topic identification. It outperforms SVM (76.2%) and has a comparable performance with the cache model (82.0%).

1. INTRODUCTION

Statistical language models (SLMs) play a pivotal role in many natural language processing applications as speech recognition, machine translation, information retrieval, handwriting recognition, etc. In speech recognition systems, which we are interested in, the speech recognizer determines an estimate \(\hat{W}\) of the identity of the spoken word sequence from the observed acoustic evidence \(O\). The sequence \(\hat{W}\) is determined by combining two scores: \(p(O|W)\), the probability of observing the acoustic evidence \(O\) when the sequence \(W\) is uttered and \(p(W)\) the probability of \(W = w_1, \ldots, w_n\). The task of the language model is to provide estimates of \(P(W)\).

In the past two decades, statistical \(n\)-gram models have steadily emerged as the preferred way to model language behavior. The dependence of the conditional probability of observing a word \(w_i\) at a position \(i\) is assumed to be restricted to its immediate \(n-1\) predecessor words \(w_{i-n+1} \ldots w_{i-1}\).

Probabilities of \(n\)-grams are estimated using large corpora. However, several \(n\)-grams do not appear in the training corpora, leading to null probabilities. To prevent from these probabilities, smoothing methods are used [1]. The higher \(n\) is, the higher the number of unseen \(n\)-grams. Moreover, the lower \(n\) is, the worse the model. Thus, \(n\) is usually set to 3 (trigram model). The information used to estimate word \(w_i\) is however small. Researches have then been conducted to increase the information used to predict \(w_i\): Kuhn integrates a cache memory [2], probabilities of words present in the history are increased. Similarly, Rosenfeld incorporates triggers of words in language models [3]. Recently, Chelba has presented a model which uses at the same time a tagger, a parser and a \(n\)-gram models [4].

Another way to increase the information taken into account to predict words is to exploit the topic of the text: the language is assumed to vary according to the topic dealt in the text. Then, adapting language models to the topic detected will improve its representation of the text. In this article, an original topic identification method, WSIM, is presented. It is based on the exploitation of the similarity between words and topics.

Section 2 studies relationships between topic identification and text categorization and a state of the art of text categorization is made. Section 3 presents the principles of our new topic identification method, WSIM. Then, section 4 studies its performance, which is compared to two well-known topic identification methods. We finally conclude and present several perspectives.

2. TOPIC IDENTIFICATION

2.1. Text categorization definition

Text categorization is the task of assigning a boolean value to each pair \(< d_i, c_j > \in D \times C\), where \(D\) is a domain of documents and \(C = \{c_1, \ldots, c_{|C|}\}\) is a set of predefined categories. A value True assigned to \(< d_i, c_j >\) indicates a decision to file \(d_i\) under \(c_j\), while a value of False indicates a decision not to file \(d_i\) under \(c_j\).

In the framework of topic identification, the set of categories \(C\) is the set of topics. For a given document \(d_i\), the corresponding topic is searched.

2.2. State of the art

Most of text categorization methods exploit the frequency of words in the texts. Each document is first represented as a vector of term weights \(d_i = w_1, \ldots, w_{|V|}\) where \(V\) is the
vocabulary, and $0 < w_{k_i} < 1$ represents, loosely speaking, how much term $t_k$ contributes to the semantics of document $d_i$.

Several approaches have been studied to determine the class of a document. We present in this section four well-known approaches. Probabilistic approaches compute the probability of each class given the test document. The topic unigram model [5] and the naive bayes classifier [6] are well-known probabilistic classifiers. In the decision tree approach [7], each node represents a term and branches departing from them are labelled by tests on the weight that the term has. In neural networks, the term weights in the test document, and leaves are labelled by tests on the weight that the term has [7], each node represents a term and branches departing from known probabilistic classifiers. In the decision tree approach [7], each node represents a term and branches departing from them are labelled by tests on the weight that the term has. In neural networks, the term weights in the test document, and leaves are labelled by tests on the weight that the term has [7], each node represents a term and branches departing from known probabilistic classifiers. In the decision tree approach [7], each node represents a term and branches departing from them are labelled by tests on the weight that the term has [7].

In this article, we present an original topic identification method, vocabulary, and $0 < w_{k_i} < 1$ represents, loosely speaking, how much term $t_k$ contributes to the semantics of document $d_i$.

Several approaches have been studied to determine the class of a document. We present in this section four well-known approaches. Probabilistic approaches compute the probability of each class given the test document. The topic unigram model [5] and the naive bayes classifier [6] are well-known probabilistic classifiers. In the decision tree approach [7], each node represents a term and branches departing from them are labelled by tests on the weight that the term has. In neural networks, the term weights in the test document, and leaves are labelled by tests on the weight that the term has [7], each node represents a term and branches departing from known probabilistic classifiers. In the decision tree approach [7], each node represents a term and branches departing from them are labelled by tests on the weight that the term has [7].

In neural networks, the term weights $w_{k_j}$ are loaded into the input units; the activation of these units is propagated forward through the network, and the value of the output units determines the categorization decision, see [8] for an example. Support Vector Machines (SVMs) attempt to find the $V$-dimensional space that separates the positive from the negative training examples by the widest possible margin. SVMs have been first used in text categorization by Joachims [9]. In [10], a comparative study of five approaches (a neural network, a probabilistic approach, a SVM and an kNN and a LLSF) has been presented. SVMs and kNN have led to the best results.

Contrary to the previous methods, binary approaches represent documents as binary vectors, where term weights account for the presence or absence of terms in the documents [11].

3. WSIM TOPIC IDENTIFICATION METHOD

This article presents an original topic identification method, based on the similarity between words and topics. The similarity between word $w_i$ and topic $T_j$ is based on the similarity between $w_i$ and each word characteristic of topic $T_j$.

3.1. Similarity between two words

In [12], a similarity measure between two words $x$ and $y$ is presented. It is based on the comparison of the behavior of $x$ and $y$ with other words. Precisely, two words are similar if their mutual information with other words are close. The similarity is evaluated as follows:

$$\text{Similarity}(x, y) = \frac{1}{2V} \sum_{i=1}^{[V]} \min\{I(z_i, x), I(z_i, y)\} + \max\{I(z_i, x), I(z_i, y)\}$$ \hspace{1cm} (1)

where $V$ is the vocabulary and $I(z_i, x)$ represents the mutual information between words $z_i$ and $x$. We have adapted this measure to perform topic identification. In this case, we are interested in a topic-dependent similarity. Thus, we propose to compute similarity between $x$ and $y$, in the topic $T_j$ as:

$$\text{Similarity}_{ij}(x, y) = \frac{1}{2l_j} \sum_{i=1}^{l_j} \min\{I_j(z_i, x), I_j(z_i, y)\} + \max\{I_j(z_i, x), I_j(z_i, y)\}$$ \hspace{1cm} (2)

where $l_j$ is the number of words in the vocabulary of topic $T_j$ and $I_j(z_i, x)$ is evaluated on the training corpus of $T_j$.

3.2. Similarity between words and topics

Let $V_j = v_{j1}, v_{j2}, \ldots, v_{j|T_j|}$ the vector representing topic $T_j$, where each element denotes the similarity between a word and $T_j$. The similarity between $x$ and $T_j$ is evaluated as the mean of the similarities between $x$ and each word of the vocabulary of $T_j$, times the probability of $x$ in topic $T_j$:

$$v_{jx} = \text{Sim}(x, T_j) = \frac{\sum_{k=1}^{l_j} \text{Sim}_{ij}(x, v_{k})}{\sum_{k=1}^{l_j} \text{Sim}_{ij}(x, v_{k})} \hspace{1cm} (3)$$

3.3. Determining the topic of a document

In the training phase, a vector $V_j$ has been built for each topic $T_j$ $(j \in 1..J)$. In the test phase, the score of each topic given the test document $d = w_1, w_2, \ldots, w_N$ is evaluated as follows:

$$WSIM(T_j \mid d) = \sum_{i=1}^{N} v_{jw_i} \sum_{k=1}^{l_j} \delta_{ij}$$ \hspace{1cm} (4)

with $\delta_{ij} = \begin{cases} 1 & \text{if } w_i \in T_j \\ 0 & \text{else} \end{cases}$ and $\sum_{j=1}^{J} |\delta|_j$ is the number of words of $T_j$ in $d$, and $\varphi_j$ is a topic weighting coefficient with $\sum_{j=1}^{J} \varphi_j = 1$. $\varphi_j$ are evaluated by cross-validation, on an optimization corpus. The topic of $d$ is the one maximizing (4).

4. EVALUATION

4.1. Data

Topic detection experiments are evaluated on 4 years of the French newspaper Le Monde, 1987-1991 (more than 80M words). This corpus, made up of articles, is divided into 7 topics. Since several topics can be present in one single article, we have decided to work at the paragraph level, to avoid changes in topic. The test data is thus made up of 834 paragraphs, resulting from a random sampling, which have been manually labeled with one topic. Table 1 presents the number of paragraphs in training and test corpora, for each topic.
4.2. Vocabularies

Topic vocabularies have been chosen by using the mutual information measure between words and topics (similarly to the processing of WSIM). Mutual Information is evaluated as follows:

\[ I(x, T_j) = P(x, T_j) \log \frac{P(x, T_j)}{P(x)P(T_j)} \]

with \( P(x, T_j) \) is the probability that word \( x \) appears in a document of topic \( T_j \), \( P(x) \) is the a priori probability of word \( x \) and \( P(T_j) \) is the a priori probability of topic \( T_j \). A high mutual information between \( x \) and \( T_j \) means a strong relationship between them. Then, topic vocabularies will be made up of words with the highest mutual information. We have chosen to keep the same number of topic words per topic.

4.3. Results

We compare the performance of WSIM with two other methods. First, the cache model, which has been proved in [13] to be the best model on Le Monde corpus. Its performance is also compared with the SVM classifier. SVM has been proved to be the best model in [10]. We chose to test the SVM proposed by Joachims [14], with a linear kernel. These three methods do not use the same vocabulary, we chose for each one, the vocabulary which provides the highest performance. The cache model uses a vocabulary of 4,000 words for each topic. These vocabularies are made up of the most frequent words in the topic-training corpora. The SVM uses a vocabulary of 5,500 words, made up of the concatenation of the topic-vocabularies. For each topic, the 1,000 words with the highest topical mutual information are conserved. The WSIM vocabularies are also made up of the words with the highest topical mutual information. We have chosen 2,000 words per topic, leading to a general vocabulary of 13,000 words.

Performance of the three methods are presented in table 2. We can remark that the WSIM method performs slightly better than the cache model (0.4%). Moreover, Cache and WSIM outperform the SVM classifier.

As presented in [10], it would be interesting to study the methods performance according to the size of training data. Their behavior is studied for each topic, in terms of precision, recall and \( F_1 \), where:

\[ \text{Recall}_T = \frac{\text{Nb texts correctly labelled } T}{\text{Nb texts of topic } T} \]
\[ \text{Precision}_T = \frac{\text{Nb texts correctly labelled } T}{\text{Nb texts labelled } T} \]
\[ F_{1,T} = \frac{2 \cdot \text{Recall}_T \cdot \text{Precision}_T}{\text{Recall}_T + \text{Precision}_T} \]

Performance of the three methods is presented in Table 3. We can first notice that the topic History, is correctly recognized by none of the methods. We could think this is due to its training corpus size, but Sports topic, which also has a lower training corpus size is as well recognized as other topics. Topic History suffers from a lack of own topic-vocabulary, leading to a higher difficulty to detect the topic. WSIM, which provides the highest general performance has the highest \( F_1 \) performance only for two topics. However, in the other cases, its performance is close to the best one. Moreover, the performance of WSIM in terms of recall is much homogeneous than the two other methods, it seems to not depend on the size of the training corpus.

On History, Politics, Science and Sports topics, which are the 4 topics with the lowest training size, the SVM performs the best in terms of precision, reaching a value of 100 for topic Sports. These results confirm the ones presented in [10].

5. CONCLUSION AND PERSPECTIVES

We have investigated a new topic identification method, WSIM. The originality of this method lies in the use of the similarity between words and topics, based on the mutual information between words. We compared its performance to two well-known topic identification methods: a cache model and a linear SVM. WSIM compares favorably with the SVM, and reaches performance similar to the cache. The SVM we used has a linear kernel. It is obvious that our data is not linearly separable. We will test a SVM with a kernel that fits data considered. Moreover, topic-vocabularies have
Table 3. Recall, precision and $F_1$ for each method and each topic

<table>
<thead>
<tr>
<th>Topic</th>
<th>Cache Rap</th>
<th>Cache Prec</th>
<th>Cache $F_1$</th>
<th>WSIM Rap</th>
<th>WSIM Prec</th>
<th>WSIM $F_1$</th>
<th>SVM Rap</th>
<th>SVM Prec</th>
<th>SVM $F_1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Culture</td>
<td>84.7</td>
<td>90.4</td>
<td>87.4</td>
<td>85.3</td>
<td>87.1</td>
<td>86.2</td>
<td>94.2</td>
<td>74.6</td>
<td>83.3</td>
</tr>
<tr>
<td>Economy</td>
<td>74.6</td>
<td>91.0</td>
<td>82</td>
<td>78.8</td>
<td>84.6</td>
<td>81.6</td>
<td>77.8</td>
<td>81.5</td>
<td>79.6</td>
</tr>
<tr>
<td>Foreign news</td>
<td>86.3</td>
<td>73.9</td>
<td>79.6</td>
<td>85.3</td>
<td>79.8</td>
<td>82.5</td>
<td>92.2</td>
<td>59.1</td>
<td>72.0</td>
</tr>
<tr>
<td>History</td>
<td>16.6</td>
<td>14.3</td>
<td>14.4</td>
<td>8.3</td>
<td>33.3</td>
<td>13.3</td>
<td>14.3</td>
<td>66.6</td>
<td>23.5</td>
</tr>
<tr>
<td>Politics</td>
<td>85.1</td>
<td>75.1</td>
<td>79.8</td>
<td>86.2</td>
<td>83.0</td>
<td>84.6</td>
<td>64.3</td>
<td>84.2</td>
<td>72.9</td>
</tr>
<tr>
<td>Science</td>
<td>88.1</td>
<td>82.7</td>
<td>85.4</td>
<td>83.5</td>
<td>79.8</td>
<td>81.6</td>
<td>63.3</td>
<td>85.2</td>
<td>72.6</td>
</tr>
<tr>
<td>Sports</td>
<td>75</td>
<td>72</td>
<td>73.4</td>
<td>83.3</td>
<td>62.5</td>
<td>71.4</td>
<td>45.8</td>
<td>100</td>
<td>62.8</td>
</tr>
</tbody>
</table>

been chosen to contain an equal number of words. It should be interesting to test performance when different vocabulary sizes per topic are used.

Moreover, as suggested previously, topic History does not have any own topic-vocabulary. It seems to be characterized by dates, proper nouns, etc. Identifying History should pass by the use of semantic classes.

6. REFERENCES


