Categorizing Web Pages as a Preprocessing Step for Information Extraction

Viktor Pekar, Richard Evans, and Ruslan Mitkov
Computational Linguistics Group, HLSS
University of Wolverhampton
Stafford Street, WV1 1SB, UK
{v.pekar, r.j.evans, r.mitkov}@wlv.ac.uk

Abstract

At present, information systems combining crawling and information extraction (IE) technologies acquire a lot of research and industrial interest. In this paper, we present an algorithm that exploits techniques for unsupervised IE pattern acquisition in order to facilitate identification of web pages containing information relevant to the IE task.

1. Introduction

At present, information systems combining crawling and information extraction (IE) technologies acquire a lot of research and industrial interest. These systems monitor the web for relevant pages, extract specific information from them, and place it into a database, possibly integrating data from various sources, so that users can access it via database queries. Within the ESRC-funded BiRD[1] project, our overall goal is to create an automatically updateable information facility for researchers/educators/students of a particular discipline that enables them to have easy access to information about existing resources in the areas of their interest. The present paper is concerned with building such a resource for the domain of NLP and computational linguistics. We would like our system to locate and extract information on:

(1) computational tools and data repositories (software, corpora, grammars, lexicons, evaluation datasets, etc);

(2) forthcoming conferences on NLP and computational linguistics;

(3) job openings for specialists in relevant areas.

Many repositories of this kind already exist for various disciplines, including NLP (e.g., the NLP software registry, ACL NLP/CL Universe), and enjoy a lot of popularity among special interest communities. However most of them are created and maintained manually which is a source of many limitations, such as poor coverage and the speed with which they fall out of date.

Our system is to keep track of emails sent to specialized mailing lists on the subject, such as the Corpora List, identify messages describing relevant resources and extract pre-defined types of information from them. A quick inspection of email list archives reveals that very often valuable resources are not properly described in emails, but only informally mentioned together with corresponding URLs that do contain detailed descriptions. That is why relevant information has to be looked for not only in email messages, but also in web pages that are referred to in them. Moreover, URLs mentioned in emails often do not point to pages containing an actual description of a resource, but to some related page, e.g., the home page of the developer, so that several neighboring pages need to be considered as well. Therefore, the first important step is to identify relevant documents in the vast pool of email messages and web pages, and assign them to three groups corresponding to the three types of NLP resources.

Traditional text categorization methods seem to be an obvious solution to this task. However, simple bag-of-words representations of documents can prove quite ineffective here. The problem is that, on the one hand, the “categories” of documents we are interested in are not formed on the basis of topic similarity between documents, but on the basis of very specific information the documents contain. On the other hand, documents of a particular category are characterized by very different styles of presentation and very different vocabularies. The goal of the present study is to investigate the hypothesis that categorization of the documents can be facilitated by the use of techniques for unsupervised acquisition of IE patterns, whereby those parts of documents that are likely to contain extractable information are used for their categorization.

2. The approach

The information we would like to obtain can be described as a set of pre-defined semantic classes (*entities*, e.g., DATES, VENUE, EMPLOYER, SALARY). For each particular resource, we need to fill in a *template* with instances of these entities. The filled template will afterwards be mapped to a database entry. Assuming that one document describes one resource and thus all instances found in it belong to one template, it is sufficient to locate individual instances in a document, without discovering relations between them. Therefore *extraction patterns* needed for this task may consist of only one slot and specify constraints that the filler phrase needs to meet in order to fill that slot. These constraints may include the presence of a particular trigger phrase, certain morphological,

[1] http://clg.wlv.ac.uk/projects/BiRD/

Extractable Entities	Extraction patterns
NameOfResource	Filler<NP> is a tool/system/package for Filler<NP> allows you to Filler<NP> features
Developer	is developed by Filler<Person>
Institution	is developed at Filler<Organisation>
ApplicationArea	performs Filler<NP>, is a tool/system/package for Filler <NP> a tool that Filler<VP>
NaturalLanguage	available for the following languages: Filler, Filler, Filler is used to …<VP> Filler<adjective> texts
LicenceInformation	is available Filler<adverb> is Filler<adverb> available
SupportedPlatform	supports the following platforms: Filler, Filler, Filler a Filler implementation
RequiredResources	uses/requires the following resources: Filler, Filler, Filler

Figure 1. A template for an NLP resource. Text in angle brackets specifies linguistic and conceptual constraints on the phrase immediately before it.

and on the trigger, as well as cues derived from punctuation and general page layout. Figure 1 describes a template for a computational resource for NLP and, for each entity, gives examples of patterns that extract its instances.

Our hypothesis is that phrases constituting IE patterns are the most useful evidence for the categorization of documents. Research in IE has developed a range of techniques to automatically acquire IE patterns, such as wrapper induction techniques (e.g., Kuschmerick et al., 1997). However, the nature of our task, particularly the fact that the documents we have to deal with come from very diverse sources, make it necessary to adopt a pattern acquisition method which does not depend on the domain-specific formatting of documents. Our approach for identifying the most useful evidence for document categorization builds on unsupervised methods of IE pattern acquisition (e.g., Riloff, 1996; Yangarber et al., 2000).

The basic assumption behind these methods is that an extraction pattern is usually a verb-argument structure, where the verb is the trigger and its argument is the filler phrase; useful extraction patterns are those verb-argument structures that are most strongly correlated with domain-specific documents. Riloff's Autoslog-TS (Riloff 1996) obtains primary extraction patterns in this manner, requiring no annotated text but only a set of documents pre-classified as relevant or non-relevant. The primary patterns are later revised by a human expert who also assigns conceptual roles to their slots. Yangarber et al. (2000) exploited this idea to bootstrap a lexicon of extraction patterns and a domain-specific corpus from a general text collection. The approach starts with a small amount of seed patterns and uses them to locate domain-specific stretches of text in the general corpus. Verb-argument structures are then extracted from this text. At each iteration, a small set of new verb-argument structures which have the highest association to relevant documents is added to the IE lexicon and used to bootstrap the search for more relevant text in the general corpus.

For our document categorization task, we discover primary extraction patterns from the distribution of verb-argument tuples across document categories. In addition to parsed text, we also obtain primary IE patterns from tables and bulleted lists appearing on web pages, as well as from semi-structured text where relations between words are indicated by, for example, semicolons, tabs, multiple spaces, etc. Once primary patterns are identified, they are used to form vector representations of documents which are input into a text categorization system.

We conducted a series of experiments comparing the quality of document classification achieved using "bag-of-words" representations of documents versus the representations built of phrases contained in the primary IE patterns.

3. Experimental evaluation

3.1. Collection of the corpus

In our experiments we used a corpus of web pages and email messages, where each document is provided with one of the four category labels: *conferences, jobs, resources* and *trash*. This corpus was compiled as follows. On the ELSNET list archive, email messages are arranged into categories, two of them being job announcements and calls for papers. *Conferences* and *jobs* categories were collected by simply downloading corresponding documents from the ELSNET web site. Since the entire text of job announcements and calls for papers is usually included in an email, these two categories were made up entirely of email messages. Documents describing NLP resources were downloaded by following URLs of the resources specified on the DFKI NLP software registry web site. Only those pages that contained instances of at least three extractable entities were included into the corpus. The *trash* category was prepared by manually classifying the Corpora email messages and web pages referred to in them.

The downloaded messages and web pages were pre-processed as follows. Duplicate documents (resulting from multiple postings of the same message, the same web page being mirrored on different web sites, etc) were identified by computing the cosine similarity between documents of comparable size. From pairs with a similarity score over 0.97 we randomly deleted one document.

Text appearing on every page of an email archive (such as the name of the archive, the names of email threads, etc) as well as text of the cited messages was deleted. Non-English portions of documents were filtered by splitting the documents into paragraphs and comparing the relative frequencies of several most frequent English words in each paragraph with ones estimated from English texts to decide whether the text in the paragraph is English or not.

Since the bodies of email messages appear as plain text, HTML formatting was automatically introduced into them: headings, subheadings, bulleted lists, names of URLs and email addresses were recognized and tagged accordingly. Special mark-up was introduced for automatically recognized attribute-value lists (i.e., successive lines of text, each containing the same separator symbol, such as a semicolon or tab character, and separated from the rest of the text by either blank lines and/or a semicolon at the end of the preceding line). The structure of the corpus documents was standardised and corrected using HTML Tidy[2] and encoding issues dealt with. A vocabulary of NLP terms was extracted from the downloaded ELSNET and Corpora email messages using the TerminologyExtractor software[3]. The terms were then located in the corpus documents and marked up. After this tagging phase, the corpus was converted to valid XML and a DTD was defined.

In order to use the documents in text categorization experiments, all words constituting a term were joined, all web and email addresses were substituted by the strings *web_address* and *email_address*, all digits were substituted by the *digit* string, after which tags were stripped off. From the resulting plain text files we discarded those smaller than 2 Kbytes as they typically contain no extractable information and would only create noise during categorization. After the removal of small files, the resulting corpus contained 100 documents in *conferences*, 99 in *jobs*, 75 in *resources* and 166 in the *trash* categories.

3.2. Primary IE patterns

To obtain primary IE patterns, the original XML files corresponding to the selected plain text files were located. First, in the contents of each file, grammatical text was extracted by removing small tables, bulleted lists of words and phrases, and attribute-value lists. The remaining text was parsed using Connexor's FDG Parser (Tapanainen & Jarvinen, 1997), from which verb-argument and noun-PP tuples were extracted. In order to recognize syntactic variants of the same pattern, passive verbs in the extracted tuples were changed to the active form, and the dependency relation to its argument changed accordingly. The parsed text yielded on average 159 tuples per corpus document.

Primary IE patterns were extracted from ungrammatical text, i.e., tables, bulletted lists, and attribute-value lists, as follows. Since tables are often used in order to conveniently arrange text on web pages, only small tables (those containing less than 1000 characters), possibly appearing within larger ones, were considered, as others are unlikely to contain extractable information. Text in larger tables was assumed to contain sentences and hence was processed by a parser. In each table, trigger phrases indicating extractable information were looked for in cells of the first column and the first row. The text in these cells was recognized as a trigger phrase if it appeared in bold and/or in capital letters. The text in the rest of the cells was taken to be potential fillers signalled by the trigger phrases. The sentence immediately before each bulleted list was taken to be the trigger and each list instance the filler signalled by that trigger. Attribute-value lists were decomposed and text before the separator was taken to be the trigger whereas text after it was taken to be the filler phrase.

Table 1 describes the document categories used in the experiments. As can be seen, the *resources* and the *trash* categories are characterized by a noticeably larger number of unique terms contained in their documents in comparison to *conferences* and *jobs*, although the size of the documents is comparable across categories. One can thus expect that it will be more difficult to correctly classify *resources* and *trash* documents. We also note that the number of unique terms in the representations prepared from primary IE patterns is around half that of those prepared from the entire text of the documents ("the baseline").

	Baseline		IE patterns	
	#unique terms	#words per doc	#unique terms	#words per doc
Conf	8951	1237.74	4077	561.55
Jobs	10395	419.87	5028	203.59
Res	14747	1197.89	7645	575.49
Trash	39754	3402.03	18066	984.22

Table1. Characteristics of the document categories.

3.3. Experiments

In our experiments we compared the quality of document categorization resulting from training the classifier on representations derived from the entire text of the documents versus those containing only the primary IE patterns found in them. Two different classifier-induction algorithms were used. The first one is Probabilistic TFIDF, a probabilistic version of the Rocchio classifier (Joachims, 1996). The second one is the multinomial Naïve Bayes classifer (McCallum & Nigam, 1998). In the experiments we used the implementations of the algorithms found in the Rainbow tool kit[4].

During the experiments the entire corpus was randomly divided into training and test parts in a proportion of 9 to 1. The effectiveness of the categorization was measured in terms of precision and recall, which were then used to compute the F_β measure

[2] http://tidy.sourceforge.net

[3] http://www.chamblon.com/terminologyextractor.htm

[4] http://www-2.cs.cmu.edu/~mccallum/bow/rainbow/

(β=1). In addition, the effectiveness was measured within each category to derive macroaveraged figures and for individual test documents to obtain microaveraged figures. Each method of document representation was tested on 10 test/training splits, and the reported effectiveness figures are averaged over the 10 runs.

4. Results

The results of the experiments are presented in Tables 2 and 3. The first four rows describe the F_β measures registered within the four document categories, the last two rows describe macro- and microaveraged F_β measures. Figures indicating effectiveness higher than that achieved when using standard document representations (the baseline) are displayed in bold. On the results from PrTFIDF, we see that using primary IE patterns to represent the documents improves on the baseline both in terms of micro- and macroaveraging as well as within individual categories, except in the case of *conferences*. The improvement is statistically significant according to the one-tailed independent group t-test for *jobs*, *trash*, and for microaveraging at α = 0.01. Results from Naïve Bayes show improvement only for macroaveraging, while the effectiveness is lower for microaveraging. However, neither difference is statistically significant. Within individual categories, we observe improvement for *resources* and *trash*. For *resources*, the improvement is significant at α = 0.05.

	Baseline		IE patterns	
	F_β	st.dev	F_β	st.dev
Conf	0.837209	0.067855	0.829069	0.060387
Jobs	0.679559	0.069277	**0.806521**	0.077266
Res	0.761547	0.080084	**0.814232**	0.098387
Trash	0.322443	0.191126	**0.529293**	0.174406
Mac.	0.650190	0.227784	**0.744779**	0.143962
Mic.	0.662791	0.050536	**0.748837**	0.071721

Table 2. The categorization effectiveness achieved using PrTFIDF.

	Baseline		IE patterns	
	F_β	st.dev	F_β	st.dev
Conf	0,890648	0,064295	0,866292	0,063569
Jobs	0,941963	0,03994	0,930557	0,045879
Res	0,635338	0,144265	**0,750675**	0,13196
Trash	0,680668	0,239216	**0,763163**	0,074875
Mac.	0,787154	0,151728	**0,827672**	0,085959
Mic.	0,837209	0,049028	0,827907	0,059240

Table 3. The categorization effectiveness achieved using Naïve Bayes.

Overall, these results indicate that the use of primary IE patterns to represent the content of the documents is preferable when the documents need to be assigned to categories relevant to the IE task. We first found that the quality of categorizations did not deteriorate when using this method to represent documents while the term space was reduced to around half the size of the baseline method. Secondly, in certain cases we found a statistically significant improvement in categorization accuracy, in particular for the *resources* and the *trash* categories, which, unlike *conferences* and *jobs*, are web pages coming from extremely diverse sources and thus their documents are often represented by very diverse vocabularies. The use of primary IE patterns thus seems to emphasize the commonality of the documents belonging to these categories.

5. Conclusion

In this paper we have presented a new method to represent documents for their categorization with respect to an IE task, and have presented preliminary results for its evaluation. These results indicate that the method allows the creation of efficient document representations, and, for some document types, the method enhances the categorization effectiveness. Improvement was paricularly noted in categories consisting of those documents which are characterized by greatly differing vocabularies.

As future work, we envision work on amelioration of the kind of primary patterns used in this study. In particular we are going to examine different ways to select most relevant patterns on the basis of their distribution across document categories and perform the generalization of primary patterns by means of a domain ontology. Furthermore, we are going to investigate how the distribution of primary IE patterns across document categories can be used in order to speedily construct a dictionary of IE patterns.

Acknowledgements

The study was carried out within the research project “An Automatic System to Build Resource Databases for Researchers” funded by the UK Economic and Social Research Council.

References

Joachims, T. 1997. A probabilistic analysis of the Rocchio algorithm with TFIDF for text categorization. In Proceedings of ICML-97, 14th International Conference on Machine Learning (pp.143-151).

Kushmerick, N., Weld, D., and Doorenbos, R. 1997. Wrapper induction for information extraction. Proceedings of the 15th International Conference on Artificial Intelligence (pp.729-735).

McCallum, A. & Nigam, K.1998. A Comparison of Event Models for Naïve Bayes Text Classification. In Proceedings of the AAAI-98 Workshop on Learning for Text Categorization.

Riloff, E. 1996. Automatically Generating Extraction Patterns from Untagged Text. In Proceedings of AAAI-96. The AAAI Press/MIT Press (pp. 1044-1049).

Tapanainen, P. & Jarvinen, T. 1997. A non-projective dependency parser. In Proceedings of 5th Conference on Applied Natural Language Processing (pp.64-71).

Yangarber, R., Grishman, R., Tapanainen, P., Huttunen, S. 2000. Automatic Acquisition of Domain Knowledge for Information Extraction. In Proceedings of COLING-2000, 18th International Conference on Computational LInguistics (pp. 940-946).