Automatic Annotation of Bibliographical References for Descriptive Language Materials

Harald Hammarström*

Max Planck Institute for Evolutionary Anthropology
Department of Linguistics
Deutscher Platz 6
D-04 150 Leipzig
Germany
h.hammarstrom@let.ru.nl

Abstract. The present paper considers the problem of annotating bibliographical references with labels/classes, given training data of references already annotated with labels. The problem is an instance of document categorization where the documents are short and written in a wide variety of languages. The skewed distributions of title words and labels calls for special carefulness when choosing a Machine Learning approach. The present paper describes how to induce Disjunctive Normal Form formulae (DNFs), which have several advantages over Decision Trees. The approach is evaluated on a large real-world collection of bibliographical references.

Keywords: Document Categorization, Supervised Learning, Cross-Lingual Information Retrieval, Decision Trees, Language Documentation.

1 Introduction

LangDoc is a large-scale project to list bibliographical references to descriptive materials to all of the ca 7 000 languages of the world [1]. The present collection contains nearly 160 000 such references.

A linguist, typically a typologist searching/browsing through references, would want the collection *systematically* annotated with metadata, such as the identity of the [target-]language(s) the reference treats, the geographical location country/continent, the content-type of the document the reference refers to (e.g., (full-length) `grammar`, `grammar sketch`, `dictionary`, `phonological description`) and so on.

The present collection of 160 000 references comes from a variety of sources, some of which are already annotated with metadata, and this can be exploited in terms of supervised learning.

For example, a bibliographical reference to a descriptive work may look as follows:

* The author wishes to thank Sebastian Nordhoff, Martin Haspelmath, Guillaume Ségérer, Jouni Filip Maho and Alain Fabre for various kinds of input relevant to the present study.

P. Forner et al. (Eds.): CLEF 2011, LNCS 6941, pp. 62–73, 2011.

Schneider, Joseph. 1962. *Grammatik der Sulka-Sprache (Neubritannien)* (Micro-Bibloteca Anthropos 36). Posieux: Anthropos Institut.

This reference happens to describe a Papuan language called `Sulka [sua]`, it is a `grammar` (rather than a `dictionary`, `grammar sketch` etc.), and is further tagged with `Oceania` (macro-area) and `Papua New Guinea` (country). This example reference is written in German (i.e., the [meta-]language that the publication, and therefore reference, is written in – not the [target-]language that the publication aims to describe).

Now suppose we are given a new bibliographical reference which has no annotation. We would like to automatically annotate it with identity, type and whatever other labels are justified, given the training data consisting of already annotated references. For example, many titles in the training data will contain the word "Grammatik" and be annotated with `grammar`, those few which have the word "Neubritannien" will likely be annotated with `Oceania` and `Papua New Guinea` and so on.

Unfortunately, the problem is not as simple as checking for statistically significant keywords.

2 Problem Statement

The problem at hand can be seen as a special case of a more general Information Extraction problem with the following characteristics.

- There is a set of natural language objects O
- There is a fixed set of categories C
- Each object in O belongs to zero or more categories, i.e., there is a function $Z : O \rightarrow Powerset(C)$
- The task is to find classification function f that mimics Z.

The special case we are considering here is such that:

- Each object in O contains a small amount of text, on the order of 100 words
- The language of objects in O varies across objects, i.e., not all objects are written in the same language
- $|C|$ is large, i.e., there are many categories (in our case 5 471 + 14 + 6 = 5 491 classes, see Table 1)
- $|Z(o)|$ is small for most objects $o \in O$, i.e., most objects belong to very few categories
- Most objects $o \in O$ contain a few tokens that near-uniquely identifies $Z(o)$, i.e., there are some words that are very informative as to category, while the majority of tokens are very little informative. (This characteristic excludes the logical possibility that each token is fairly informative, and that the tokens *together*, on an equal footing, serve to pinpoint category.)

3 The Present Dataset

As mentioned already, the present collection contains nearly 160 000 references. The (meta-)languages of the references are (in descending order of frequency): English, German, French, Spanish, Portuguese, Russian, Dutch, Italian, Chinese, Indonesian, Japanese, Nepali, Afrikaans, Thai, Hindi, Turkish, Arabic, Georgian, Urdu, Bulgarian, Swedish, Finnish, Danish, Norwegian, Assamese, Swahili, Burmese, Polish and a few other languages which are represented in less than 10 references. Table 1 shows the incidence of various types of annotation already present. The annotation stems from the various sources of the collection (see [1] for more information on the composition and provenance of the catalogue), and some inconsistencies can therefore be expected. There is also some further largely idiosyncratic annotation from subparts of the collection (e.g., country, keywords, shelf-mark, ...) which is excluded from the present study since it overlaps in function with those selected for Table 1 (but could in principle be addressed with the same methods as in the present study).

Table 1. Size of the present database of references and incidence of annotation already in place

Annotation type	# different labels	# annotated references
Macro-area	6	121 296
Content-type	14	15 236
Target-language	5 471	88 978
Total # of annotated references		158 498

There are six possible macro-area labels, but they are not mutually exclusive. For example, a reference to a publication dealing with Africa as well as South America, should be labelled with both. Similarly, there are 5 471 different labels for target-language, and it is logically possible for a publication to refer to any subset of these, though, in practice, most references tend to target only one or a few languages. Type refers to the type of descriptive data, such as `grammar`, `dictionary`, `grammar sketch`, `wordlist`, `ethnographic work`, `texts` and eight others. The typology of type is, for historical reasons, somewhat ad hoc, but nevertheless useful to the target community of searchers. Some types logically exclude each other, e.g., a reference cannot both be a `grammar` and a `grammar sketch`, while others are compatible, e.g., a single book can contain both a `grammar` and a `dictionary`.

As to the size of the task at hand, Table 1 shows that, for example, out of the 158 498 references, 121 296 of them are already annotated (by a human) as to macro-area, but the remainder, 37 202 are in need of macro-area labels. (It is assumed that references which are already annotated with a certain kind of annotation are not in need of *more* annotation of the same kind.)

At first, this problem, i.e., reference annotation by keyword triggers, might seem like a very easy problem – just find title words which are statistically overrepresented with an annotation label in the training data, and then label

Table 2. Two example labels with some potential trigger words and their ability to "select" the respective labels

		`grammar`		
# references	contains	# with `grammar` label	precision	recall
162	"grammatik"	91	0.56	0.068
668	"der"	137	0.21	0.103
84	"grammatik", "der"	48	0.57	0.036
1	"sulka"	1	1.00	0.001
		`Sulka [sua]`		
# references	contains	# with `Sulka [sua]` label	precision	recall
1	"sulka"	1	1.00	0.16
668	"der"	4	0.01	0.67

new instances as such words occur in their titles. However, there are a few reasons why it is not that simple.

- A label may be signalled by more than one word, e.g., "kurzgefaßte grammatik" signals `grammar sketch` rather than `grammar` (not both!).
- It is not given which keyword(s) signal which label(s), e.g., from the example above, is it "Grammatik", "der" or "Grammatik der" (all of them statistically significant[1]) that signals `grammar`?
- Some labels are very common (and thus have frequent trigger words) while other labels are very uncommon (and thus their trigger words are very uncommon). This means that simple frequency thresholds cannot be used to rule out useful trigger words.
- Typically, a small set of trigger words "account" for an annotation label, i.e., no single one of them has a high recall with its label, but together they do.

For example, among 15 236 references annotated for content-type 19 921 distinct word types are present. 3 220 have the label `grammar` and 6 have the label `Sulka [sua]`. Table 2 shows that even a words like "grammatik" and "sulka" have rather low recall for their respective "true" labels, and if we combine precision and recall, there is some serious competition of (combinations with) spurious trigger words such as "der".

We will explore some Machine Learning ideas to come up with a solution tailored to the particularities of this classification problem.

4 Related Work

The approach in the present paper generalizes the method of [2] to annotate bibliographical references with only uncommon labels. We are not aware of any other work specifically targeting the annotation of bibliographical references based on the text of the reference itself.

[1] There is a reason why a word such as "der" is overrepresented with a label such as `grammar`. The label grammar is triggered by words like "Grammatik", but because of the rules of the (in this case) German language, "der" is also caused to be in title of (most) references with "Grammatik" in them.

The problem, however, has a clear analogue in Information Retrieval in the following sense. Typically, the task is to find a set of relevant documents given a document collection and a query. On the other hand, if we equate documents with the text of a bibliographical reference, and the set of relevant documents and the set of references with a certain label, then the problem addressed in this paper is to find the query given the document collection and the set of relevant documents. As such the problem has been addressed in terms of word-space models [3], and special focus has been on the special case of sentiment analysis [4]. Such work also includes principled approaches to the multi-lingual situation [5,6], though often relying on existing lexical resources, e.g., dictionaries. Such approaches scale well to large collections, but are otherwise imperfectly suited to the specifics of the problem in the present paper. First, we do not have access to dictionaries for the full range of languages featured in the present collection. Second, the techniques described output large word-probability tables which combine evidence from many words in a long document, whereas in the present collection, every document is very short. Third, most techniques described involve human-tuned seed data or thresholds which make the approaches less attractive to work with.

A number of techniques which have been successful for Text Classification (cf. [7], though somewhat dated, the principles outlined therein remain valid) are less well-suited for the present problem. There are dependencies between words that go against the Naive Bayesian assumption, e.g., "grammar" signals the label `grammar` if and only if not occurring with "short", "sketch" etc. Naive Bayes, along with a number of other statistical approaches, have no way of distinguishing which keyword(s) in a title signals signal which label(s) and end up distributing the evidence over all words in the title (which is not fatal, but unnecessary). Extra work is also required in smoothing techniques for infrequent labels. Other statistical techniques work best after text processing such as stopword removal and/or tf-idf-weighting. In the present context, we do not have access to stopword lists for the full range of languages targeted, and there is also the suspicion that what are stopwords in regular prose may not correspond to stopwords in publication titles (the same suspicion can be raised for other enhancements that tap into linguistic structures [8]). Similarly, in the very multilingual setting, tf-idf weighting is significantly crippled, as what are frequent words within *one* language will have only a fraction of their frequency when diluted in large pool of languages (and there may be distorting interferences across languages).

The traditional principled approach to classification of objects with a set of discrete valued features are decision trees [9], which, in addition, are able to "explain" their predictions. ID3 Decision Trees are well-suited for the present problem except that they may become unnecessarily large and that setting a depth-threshold is required. The reason ID3 Decision Trees may become "unnecessarily large" in the present setting is that they are designed to be built complete, e.g., if the attribute "grammaire" is chosen as a branch, the corresponding negated branch must also be present, and both branches must be filled

in the next round of iterations. In the present problem setting, we envisage the optimal tree to look more like a rake than a tree. Although general-purpose pruning heuristics to decision trees are widely used (as per C4.5 [10]), a solution specially designed to allow rake-like classifiers obviates the need for thresholds and pruning heuristics.

There are thus good theoretical arguments for re-assuming from the 1990s the approach of rule-induction classifiers [11,12,13] for the particular problem setting addressed here.

5 A DNF Approach

As outlined above, our domain knowledge suggests that a label can be inferred if and only if a suitable combination of words is present/absent in a given publication title. More formally:

- A trigger-signature $t = w_1 \wedge \ldots \wedge w_k \wedge \ldots \neg w_{k+1} \wedge \ldots \wedge \neg w_{k'}$ for a label l is a conjunct formula of negated/un-negated terms, such that if a title contains all the un-negated terms but none of the negated terms, then the label l should be inferred.
- Each label l can have one or more trigger-signatures $t_1, \ldots, t_n$

For example, one trigger for the label `grammar` might be $\{grammar, \neg sketch\}$, and the full set of triggers for `grammar` might contain $\{grammar, \neg sketch\}$, $\{grammaire\}$, $\{complete, description\}$, $\{phonologie, morphologie, syntax\}$ and so on. Since titles are short (less than 20 words or so), we envisage triggers to be short.

In other words, a classifier (one for each label) can be described as a boolean formula in DNF, where each disjunct corresponds to a trigger. Moreover, each disjunct can be expected to be relatively short.

Thus, all we need to do is to search for a formula in DNF form which can be expected to have only short disjuncts and which is preferably short (in its number of disjuncts). Thus, a simple algorithm is to start from an empty formula and build it larger as accuracy increases with respect to a label in the training data. One can build a formula larger either:

i by adding a negated/un-negated term to one of its disjuncts (replacing that disjunct[2]), or
ii by adding a new disjunct, inhabited by a negated/un-negated literal.

Since we are interested in both high precision and high recall, a natural way to measure accuracy is f-score.

The following notation will be used:

- $d_i \subseteq \Sigma^*$ be a document, i.e., a set of strings
- $D = \{d_1, \ldots, d_n\}$ be a set of documents

[2] To keep an updated and un-updated disjunct is superflous since $A = A \vee (A \wedge B)$.

- $W_D = \bigcup d_i$ be the set of terms of a set of documents
- $L_D(l) = \{i | d_i \text{ has label } l\}$ be the subset of documents with label l
- $c = \bigvee t_j$ be a DNF boolean formula
- $c_D = \{i | c \text{ is true for } d_i\}$ be the subset of documents whose terms satisfy a boolean formula c
- $Precision_D(c, l) = |c_D \cap L_D(l)|/|c_D|$
- $Recall_D(c, l) = |c_D \cap L_D(l)|/|L_D(l)|$

The training algorithm can be described as follows:

1. Start with a label l, a document collection D and an empty formula c
2. Form sets of candicate formulae

$$C' = \{c \vee w | w \in W_D\} \cup \{c \vee \neg w | w \in W_D\}$$
$$C'' = \{ins(w, t_j, c) \vee t_j | w \in W_D, t_j \text{ of } c\} \cup$$
$$\{ins(\neg w, t_j, c) \vee t_j | w \in W_D, t_j \text{ of } c\}$$

   where $ins(x, t_j, c)$ means "replace t_j with $t_j \wedge x$ in the formula c", e.g., $ins(c, t_2, (a \wedge \neg b) \vee (a)) = (a \wedge \neg b) \vee (a \wedge c)$.
3. Compute $c' = argmax_{c' \in C' \cup C''} \text{f-score}_D(c', l)$
4. If c' equals c finish, otherwise set c to c' and iterate from step 2

6 Experiments

6.1 Experiment Design

Since the labels are largely independent, we trained one DNF for each label. To classify an unseen reference, we test it with all DNFs in parallel, and label it accordingly.

As noted already, the labels fall into three classes: Target language, content-type and macro-area. For each class, we randomly selected 1000 previously annotated references to use as a test set. These were set apart from the beginning and were never accessed during development.

With the intended search audience in mind, we believe that precision is more important than recall, especially since there are catch-all labels based on geography that make up for some loss of recall. Consequently, all experiments were run to optimize the $F_{0.5}$-score of a DNF, where precision is twice as important as recall [14].

All titles are in roman script or have transcriptions. All title words were lowercased and all diacritics and accents were removed.

6.2 Results

Overall results, in numbers, grouped by label class, are shown in Table 3. The overall f-score, precision and recall figures are based on numbers of labels (rather than numbers of references, since many references have more than one label of the same class).

Table 3. Overall accuracy of the DNF approach, grouped by label class

Label Class	# labels	# training refs	$\lvert W_D \rvert$	Overall $F_{0.5}$	Overall Precision	Overall Recall
Macro-area	6	121 296	117 213	0.60	0.57	0.76
Content-type	14	15 236	19 921	0.57	0.59	0.51
Target-language	5 471	88 978	83 828	0.80	0.85	0.66
	5 491			0.70	0.70	0.69

Table 4. Accuracy for macro-area labels

Label	# in training data	$F_{0.5}$	Precision	Recall
`Australia`	6 988	0.76	0.91	0.46
`Eurasia`	10 579	0.50	0.61	0.28
`North America`	2 311	0.45	0.50	0.31
`Africa`	69 734	0.57	0.52	0.95
`Oceania` [except Australia]	3 681	0.46	0.51	0.33
`South America`	29 100	0.62	0.61	0.66
	122 393	0.60	0.57	0.76

It is instructive to look closer at the results for macro-area labels in particular, shown in Table 4

The DNF for `Australia` was extracted straightforwardly as

$$aboriginal \vee australian \vee australia \vee warlpiri \vee queensland \vee aborigines \vee arnhem \vee pitjantjatjara \vee torres \vee nyungan \vee (wales \wedge new \wedge south) \vee arrernte \vee yolngu \vee dyirbal \vee kriol \vee kimberley \vee york$$

i.e., some geographical names and some names of languages/families prominently present in Australia. The DNF for `Eurasia` is similar

$$thai \vee tibetan \vee jazyka \vee burmese \vee viet \vee tai \vee yu \vee vietnamese \vee khmer \vee slovar \vee chinese \vee iazyke \vee siamese \vee nepal \vee hmong \vee miao \vee hindi \vee tibeto \vee phasa \vee india \vee tieng \vee slov \vee japanese \vee thailand \vee grammatika \vee burman$$

except that here we also have a 'grammatika', 'slovar' and 'iazyke', i.e., which are Russian words, indirectly indicating Eurasia only since the vast majority of most Russian works target Eurasian languages. The DNF for `North America` has a list of specific language/family names but only one country name "Mexico". The DNF for `Africa` is different

$$(\neg america \wedge \neg o \wedge \neg american \wedge \neg story \wedge \neg lengua \wedge \neg australia \wedge \neg indians \wedge \neg do \wedge \neg australian \wedge \neg new \wedge \neg thai \wedge \neg grammar \wedge \neg i \wedge \neg review \wedge \neg el \wedge \neg aboriginal \wedge \neg e \wedge \neg los \wedge \neg del \wedge \neg y) \vee africa \vee swahili \vee bantu \vee hausa \vee congo \ldots$$

containing a large trigger of negated literals. This is presumably because the label `Africa` is numerically dominant. This trigger also accounts for the unusually high recall figure. The DNF for South America contains a large number of common

Table 5. Accuracy for content-type labels

Label	# in training data	$F_{0.5}$	Precision	Recall
`handbook/overview`	4 549	0.60	0.61	0.58
`grammar`	3 216	0.63	0.65	0.55
`comparative-historical`	2 992	0.50	0.49	0.59
`grammar sketch`	2 519	0.38	0.42	0.26
`ethnographic work`	1 886	0.59	0.59	0.60
`wordlist`	1 807	0.48	0.57	0.29
`dictionary`	926	0.78	0.83	0.64
`study of a specific feature`	626	0.43	0.46	0.34
`bibliographic`	550	0.72	0.76	0.61
`very small amount of information`	541	0.67	0.68	0.63
`sociolinguistic`	493	0.54	0.55	0.51
`phonology`	347	0.68	0.75	0.51
`text`	149	0.94	0.90	0.90
`dialectology`	124	0.81	0.80	0.88
	20 725	0.57	0.59	0.51

Spanish and Portuguese words, a reflection of the fact that words with Spanish and Portuguese are concentrated to South America.

Results for content-type labels, shown in Table 5 are similar, despite the much smaller size of the training set. All DNFs extracted are short (less than 50 disjuncts) and contain little of surprise. For example, the DNF for `grammar` is

$$grammar \vee grammaire \vee jazyk \vee gramatica \vee grammatik \vee grammatika \vee description \vee course \vee parlons \vee syntax \vee manuel \vee jazyka \vee spraakkunst \vee (phonologie \wedge morphologie) \vee grammatica \vee descriptive \vee manual \vee arte \vee dialekt \vee handbook$$

and the DNF for `phonology` contains the trigger signature $\ldots \vee (phonologie \wedge \neg morphologie) \vee \ldots$, precisely as expected.

Inspection shows that a lot of errors come from the fact that the labels `grammar` and `grammar sketch` are not quite distinguishable by title words alone.

For the labels in the target-language class, their frequency in the training data ranges from 1 440 of Hausa [hau] to 1 025 labels, e.g., Wurrugu [wur], with frequency 1. In the test set of 1000 references, there were an additional 211 target-language label types [221 label tokens] which do not appear even once in the training data. It is impossible to find such labels given the training data, so we also present recall figures which are adjusted upwards. Most labels in the target-language class do not appear in the 1000 item test set, wherefore we also show the precision and recall figures on the training data (as some kind of indication of the power of DNFs for these labels). Table 6 shows these results on the basis of all label tokens together – the labels are far too many to inspect on an individual basis.

Target-language labels are easy to capture with DNFs, because of the typical title contains (at least one) near-unique identifier. Top, median and bottom frequency examples are shown in Table 7.

Table 6. Overall accuracy for target-language labels

Data Set	# label tokens	$F_{0.5}$	Precision	Recall
Test Data	1 292	0.80	0.85	0.66
Test Data Adjusted	1 071	0.83	0.85	0.79
Training data	118 065	0.87	0.93	0.73

Table 7. Example DNFs for target-languages labels on the training data set

Label	# label tokens	$F_{0.5}$	Precision	Recall	DNF
Hausa [hau]	1 440	0.94	0.99	0.80	$hausa \vee haoussa \vee haussa \vee hausaland \vee hawsa \vee xausa$
Nisenan [nsz]	6	0.96	1.00	0.83	$nisenan$
Wurrugu [wur]	1	1.00	1.00	1.00	$wurrugu$

Errors in precision and recall come mainly from cases where one publication treats several languages, but the title does not list them, e.g., *Languages of the Eastern Caprivi*.

All DNFs for target-language labels are short (less than 15 disjuncts). In a fair amount of cases, a trigger signature contains a seemingly superfluous word, e.g., $the \wedge mayi$, transparently because adding one word to *mayi* makes the title unique in the training set. Presumably, there are several words which suffice equally well, but in reality, a specific one is appropriate. A future tweak could target this pattern. Similarly, there are cases of hapax words, e.g., "syllabics", which are not language names, yet since show up with only one target-language label, they are indistinguishable from a true language name in the present approach. Since such words are rare, little or no classification errors can be expected to result from such spurious language identifiers.

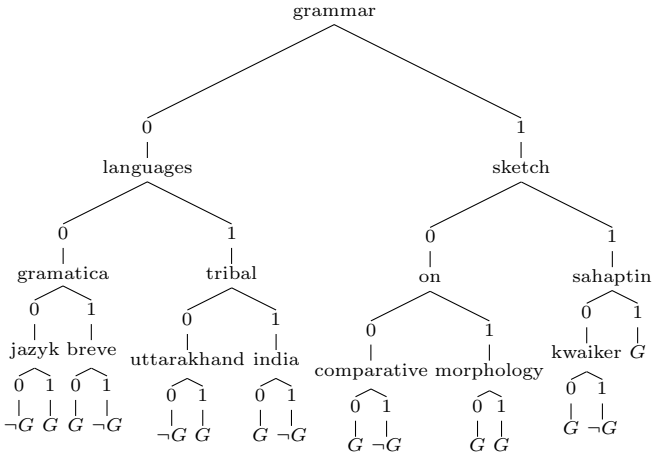
6.3 Discussion

The accuracy in the results obtained will certainly be useful in that it will save a lot of human annotation time. But since automatic annotation is imperfect and incomplete, a human will still need to browse the results and correct errors.

The performance is significantly better than ID3 Decision Trees [9] whose performance on this problem (with one tree per label, as with DNFs) yields much larger trees for the same f-scores, and require threshold (tree-height) settings for training to stop. For example, ID3, on the same training data, produces the following tree for the label `grammar`:

which has only $F_{0.5} \approx 0.42$ (precision 0.55, recall 0.22) on the same test set. The similar-sized DNF (shown above) has $F_{0.5} \approx 0.63$ (precision 0.65, recall 0.55). Cutting the tree deeper only has marginally higher scores. This is likely

due to the fact that the tree is designed to be dense, and so has to create a number of inefficient branches, whereas the DNF can mimic a sparse tree, which better fits the problem. A Naive Bayesian classifier for the label `grammar` achieves only $F_{0.5} \approx 0.35$ (precision 0.34, recall 0.45) on the same test set.



The algorithm for finding DNFs is subject to falling into local minima – indeed it is the only obstacle to overfitting. However, Since the output DNFs correspond well to intuitions, we have not investigated to what extent there are globally more accurate DNFs than the ones found by the algorithm.

Training DNFs is worst-case quadratic in $|W_D|$. Given the search space with a large W_D in the present case, this is rather slow. It is likely that intelligent filtering of W_D may significantly reduce it, but since training speed is not an issue, this has not been explored.

A drawback of the present approach is that non-boolean attributes cannot be elegantly integrated. For example, it seems likely that the number of pages (of the work that a reference points to) is highly relevant for the difficult decision between the labels `grammar` and `grammar sketch`. In our current reference collection, page numbers are not systematically present, so we are unable to check this matter thoroughly anyway.

The output formulae are readily interpretable to a human, thus the classifier annotating a new reference can "explain" its result.

7 Conclusion

We have presented a principled approach to supervised document categorization on very short documents written in a variety of languages. The present approach has advantages in elegance over alternative machine learning methods and can cope equally with common and uncommon categories, i.e., with sparse amounts of training data. The approach is thoroughly evaluated on a collection of bibliographic references and will be used in practice.

References

1. Hammarström, H., Nordhoff, S.: Langdoc: Bibliographic infrastructure for linguistic typology. Oslo Studies in Language, 14 (in press, 2011)
2. Hammarström, H.: Automatic annotation of bibliographical references with target language. In: Proceedings of MMIES-2: Workshop on Multi-source, Multilingual Information Extraction and Summarization, ACL, pp. 57–64 (2008)
3. Sahlgren, M.: The Word-Space Model: Using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces. PhD thesis, Stockholm University, Stockholm (2006)
4. Huang, X., Croft, W.B.: A unified relevance model for opinion retrieval. In: CIKM 2009: Proceeding of the 18th ACM Conference on Information and Knowledge Management, pp. 947–956. ACM, New York (2009)
5. Lavrenko, V., Choquette, M., Croft, W.B.: Cross-lingual relevance models. In: SIGIR 2002: Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 175–182. ACM, New York (2002)
6. Zhang, D., Mei, Q., Zhai, C.: Cross-lingual latent topic extraction. In: ACL 2010: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, pp. 1128–1137. Association for Computational Linguistics, Morristown (2010)
7. Sebastiani, F.: Machine learning in automated text categorization. ACM Computing Surveys 34, 1–47 (2002)
8. Al Zamil, M.G.H., Can, A.B.: Rolex-sp: Rules of lexical syntactic patterns for free text categorization. Knowledge-Based Systems 24(1), 58–65 (2011)
9. Quinlan, J.R.: Induction of decision trees. Machine Learning 1(1), 81–106 (1986)
10. Quinlan, J.R.: C4.5: programs for machine learning. Morgan Kaufmann, San Francisco (1993)
11. Cohen, W.W.: Fast effective rule induction. In: Proceedings of the Twelfth International Conference on Machine Learning, pp. 115–123. Morgan Kaufmann, San Francisco (1995)
12. Clark, P., Niblett, T.: The cn2 induction algorithm. Machine Learning 3, 261–283 (1989)
13. Sever, H., Gorur, A., Tolun, M.R.: Text Categorization with ILA. In: Yazıcı, A., Şener, C. (eds.) ISCIS 2003. LNCS, vol. 2869, pp. 300–307. Springer, Heidelberg (2003)
14. van Rijsbergen, C.J.: Information Retrieval, 2nd edn. Butterworths, London (1979)