Example-based Term Extraction for Minority Languages: A case-study on Ladin[*]

Oliver Streiter, Daniel Zielinski, Isabella Ties and Leonhard Voltmer
EURAC, Viale Druso 1, Bolzano, Italy
http://www.eurac.edu/bistro
{ostreiter;dzielinski;ities;lvoltmer}@eurac.edu

February 25, 2003

Abstract

This paper tackles the problem of Term Extraction (TE) for Minority languages. We show that TE can be realized, even if computerized language resources are sparse. We propose an example-based approach, which draws the knowledge of how a term is formed from a relatively small set of example terms. For the Ladin of Val Ghardena, which we use in our experiments, the example-based approach outperforms simple statistical approaches to TE.

1 Introduction

Electronic linguistic corpora are nowadays crucial for any kind of scientific enterprise related to language. Projects in language documentation, linguistic analysis and software development all require the creation and exploitation of spoken and written language corpora. Among the different techniques for corpus exploitation, the automatic extraction of terms (TE) has become subject to intense studies. The aim of TE is to automatize the first working-steps in the creation of terminological data for specific sub-languages.

TE is a central topic for the terminological work at EURAC. This work aims at the description, preservation and creation of legal and administrative terminology for the German and Ladin speaking minorities in South-Tyrol. The created terminological data have to provide an up-to-data, unambiguous and coherent lexis for concepts of specific subject areas (e.g. law, administration). For minority languages, the creation of terminological data is an important measure in language preservation.

In this paper we focus on TE for small minority languages, taking as example the different idioms of Ladin spoken and written in the Dolomites. In the following section (Section 2) we shortly introduce *TermLad*, a project about Ladin terminology, and illustrate some of the problems in computational terminology one may encounter when dealing with minority languages. In Section 3 we will introduce the basic notions underlying modern approaches to TE. We then explain the unithood-problem and the termhood-problem and present different techniques how to approach these problems (Section 3.2). In Section 3.3 we give an introduction to standard evaluation techniques for TE.
In a further step we critically review these approaches with respect to their applicability to minority languages. In Section 4 we review the few past experiments related to TE for less spoken languages. In Section 4.1 we introduce an example-based approach to TE and provide finally experimental data comparing the proposed approach to those cited in the literature (Section 4.2).

2 Ladin Terminology

In 1989 Ladin has received official status in the Ladin valleys of Badia and Ghardena and in 1993 in Val di Fassa. Since then, legal documents have been written in Ladin. Still, the usage of Ladin is hampered by terminological insecurity. Our survey in the municipalities confirmed that public officials encounter difficulties with the translation of regulations and ordinances from Italian. The main problem is lexical in nature and reflects a lack of elaborated administrative termi-

[*] The research described in this article has been partially supported by the region Trentino/Alto-Adige in the framework of the project *Terminologia Ladina*.

nology. This gap cannot be bridged with the available Ladin dictionaries as these feature few legal terms. Therefore paraphrases and quasi-synonyms are frequently used. The subject fields which most urgently require an elaborated terminology are town planning and building, accountancy, finance and property law.

In 2001 EURAC has launched a project to promote Ladin juridical and administrative terminology. The primary goal of the TermLad project is to describe frequently used administrative terms and to make them publicly available in form of the juridical online data-base BISTRO (http://www.eurac.edu/bistro).

The project started from delicate conditions: The use of Ladin as an official language is very recent and few electronic documents are available. The existing documents are of different quality and unequally distributed across different idioms and subject areas. At first, the EURAC collected existing electronic legal texts with the support of the local administration and Spell, the center for the planning and elaboration of a standard Ladin, and created a fragmentary and unbalanced corpus in the three Ladin idioms Ghardena, Badia and Fassa. In order to extract from these texts the underlying accepted terminology, we investigated the possibility to support terminographers with a TE tool, which is only built on a seed-list of terms and some additional, non-related texts.

3 Term extraction

3.1 Definitions

TE is an operation which takes as input a document and produces as output a list of term candidates ($\{TC\}$). Term candidates are words or phrases which are potential terms of the subject area represented by the input document. TE is still a semi-automatic procedure and requires a manual confirmation or rejection for each TC.

Traditionally, the problem of TE is seen as finding the intersection of possible language units (the unithood problem) and possible terms (the termhood problem) [c.f. CCEBVP01]. The *Unithood Problem* describes the task to select from a set of word combination those combinations which form language units (e.g. *red car* but not *is very*). The *Termhood Problem* on the other hand describes the task to select from a set of word combination those combination fulfill the requirements of a term (e.g. *red pepper* but not *red car*). These problems may be tackled separately. More often than not, the unithood problem is solved first and the output TCs of the filter are checked for their term-status (the termhood-problem). It is possible however, to start with the termhood-problem and check the TCs with respect to their unithood. For reasons of an efficient processing, often the maximum number of filter criteria for unithood and termhood are combined within the first processing step.

Table 1: A list of abbreviations used in the paper.

TC	Term candidate
TCF	Frequency of TC in a given document
DF	Document Frequency, the number of documents containing TC
IDF	Inverted Document Frequency = $\frac{1}{DF}$
$\{TC\}$	Set of term candidates
$\{T\}$	Term collection == unordered set of terms
$\{T\}_{doc}$	Subset of T belonging to one specific document
$TC \in \{T\}$	TC is a term
$\{TC\} \cap \{T\}_{doc}$	set of correct TCs
$\#\{...\}$	The cardinality of a set
MI	Mutual Information

3.2 Approaches

Approaches to TE can be classified according to the knowledge used as linguistic or statistic approaches. All of these approaches may be combined in hybrid approaches in order to join the strong aspects of complementary approaches. Another, orthogonal, classification describes approaches to TE as intrinsic relative to the TC (e.g. morphological information) or extrinsic relative to the TC. The extrinsic approach may be syntagmatic (e.g. syntactic, contextual information) or paradigmatic (e.g. relations among TCs and $\{T\}$).

3.2.1 Linguistic Approaches

Linguistic approaches make use of morphological, syntactic or semantic information implemented in language-specific programs. Its main aim is to identify language units. For reasons of efficiency and accuracy, assumptions on how terms are formed are

Table 2: Linguistic approaches to TE, an overview.

linguistic approaches			
		methods	publications
intrinsic		POS-tagging,chunking	[BJ99]
		stop-words	[MNA94, MM00]
extrinsic	syntagmatic	full parsing	[Arp95, SVT99, Hei99**?**]
	paradigmatic	term variation	[SW99, Jac99]

weaved into the linguistic analysis. These assumptions may refer to the number of words to be combined, special suffixes or part of speech requirements. Morphological analyzers, part-of-speech taggers and parsers are used for this type of analysis.

A list of stop-words, e.g. words that might not occur in a specific position of a TC (beginning, middle, end) may be used alone as a shallow linguistic criterion or in addition to other criteria.

3.2.2 Statistical Approaches

In this section we briefly explain the basic idea behind using statistics for TE and introduce some of the most frequently employed statistical measures.

Statistical approaches to TE are based on the detection of one or more lexical units in specialized documents with a frequency-derived value higher than a given threshold. They are useful both for extracting single-word and multi-word units. The assumption underlying these approaches is that specialized documents are characterized by the repeated use of certain lexical units or morpho-syntactic constructions. Contrary to common-language documents, mono-referentiality and avoidance of synonyms prevail over linguistic elegance. Statistical information that can be computed, based on these assumptions, are the frequency of a TC and its parts in a given corpus, the frequency of this TC as part of other longer TCs and the length of the TC-string (number of words, number of characters). We discuss only the more elementary measures.

(1) Frequency of occurrence: The easiest way to calculate the importance of lexical units in a corpus is counting. The more frequently a lexical unit appears in the corpus the more likely it is that this unit has a special function or meaning in a given specialized document. In part this might be true, but trying to extract TCs just by frequency would also identify frequently appearing combinations of function words as TCs. Even if used in combination with a filter for certain morpho-syntactic patterns, this approach is not always satisfactory.

The frequency of occurrence can be improved when expressed as TF.IDF. This measure, widely used for the purpose of information retrieval, divides the TCF_x, the frequency of occurrence of TC_x in the document, by the document frequency DF_x, i.e. the number of other documents which contain TC_x. The underlying assumption is that linguistic expressions which characterize a document are frequent within a document but infrequent across different documents. The same assumption is expressed in the weirdness-ratio which uses relative frequencies for TF and IDF [c.f. BMA96]).

$$\texttt{TF.IDF}_x = \frac{TCF_x}{DF_x} \qquad (1)$$

$$\texttt{weirdness ratio}_x = \frac{\frac{TCF_x}{\#\{TC\}}}{\frac{DF_x}{\sum_{d=1}^{d=m} doc_j}} \qquad (2)$$

The main problem with frequency-based approaches is that they may work well for one-word units, but do no scale up for two- or three-word units. If the TE-approach, whatever it may be, is based on a sample of 10.000 words, an extension to two-word units would require a 100.000.000 word sample in order to obtain equally good results. For the treatment of tree-word units, $10.000^3 = 1.000.000.000.000$ words would be required. This problem is known as sparse-data problem and is present in all measures which involve the frequency of the TC as undivided unit, e.g. the joint probability. If calculation of the DF includes the document from which TE starts, which is not correct for hypothesis testing, DF of most two- and tree word units will have the value 1. If the occurrence in the document is not included, DF will be 0 in most cases. However, each of these values is a bad estimate.

Table 3: Statistic approaches to TE, an overview.

statistic approaches			
		methods	publications
intrinsic		mutual information	[CH89]
		likelihood ratio	[HFH01]
extrinsic	syntagmatic	nc-value	[MA99, MA00, PL01, DHJ96]
		entropy	[MM00]
	paradigmatic	c-value	[Nak01]
	to document	weirdness	[BMA96]

Table 4: The contingency table of observed frequencies $O_{11} \ldots O_{22}$ for the word pair (A,B).

	$w_2 = B$	$w_2 \neq B$	$\sum$
$w_1 = A$	O_{11}	O_{12}	R_1
$w_1 \neq A$	O_{21}	O_{22}	R_2
$\sum$	C_1	C_2	N

Table 5: The contingency table of estimated frequencies $E_{11} \ldots E_{22}$ under independency assumption.

	$w_2 = B$	$w_2 \neq B$
$w_1 = A$	$E_{11} = \frac{R_1 * C_1}{N}$	$E_{12} = \frac{R_1 * C_2}{N}$
$w_1 \neq A$	$E_{21} = \frac{R_2 * C_1}{N}$	$E_{22} = \frac{R_2 * C_2}{N}$

The frequency of a TC is one possible variant of association measures. Association measures are used to rate the correlation of word pairs. These measures can be derived from the *contingency table* of the word pair (A,B). The contingency table contains the observed frequencies of (A,B), (A,notB), (notA,B) and (notA,notB), marked here as $O_{11} \ldots O_{22}$. If the occurrences of (A,B), (A,notB) ... are independent, their expected frequencies are estimated from the product of the marginal sums. These are stored as $E_{11} \ldots E_{22}$. Lexical association measures are formulas that relate the observed frequencies O to the expected frequency E under the assumption that A and B are independent. From the fact that the frequency is defined as O_{11}, we see that the frequency includes the cell of the contingency table most difficult to estimate (for $n > 1$) and none of the other cells which allow to relativize the observed frequency.

The Mutual Information as association measure is defined as:

$$\mathtt{MI} = \frac{O_{11}}{E_{11}} \tag{3}$$

Mutual Information (MI) is used frequently in corpus linguistics to calculate the association between two lexical units, even if this measure doesn't work very well for low-frequency events. The MI can be equally defined as the probability of the joint occurrence of w_1 and w_2, divided by the product of the probabilities of the singular occurrences. If two words occur once side by side in a one hundred words corpus, they get a MI of $\sim log_2(100)$. On the other hand, if they co-occur twice, they get a MI of $\sim log_2(50)$. This shows that the probability of the joint singular occurrence has been rudely overestimated.

Another problem with frequency-based measures is that they may rank TCs correctly if they contain the same number of words, but rank TCs of more words too low or too high. Actually, many measures are simply not defined for measuring the association between more than two words. If we would be forced to create a MI-measure for 3-word units, this could be defined as, starting from a three-dimensional contingency table:

$$\mathtt{MI}_3 = \frac{O_{111}}{E_{111}} \tag{4}$$

A singular 3-word expression in a 100 word corpus would consequently receive the MI of $\sim$ 10.000! Actually, most measures are simply not defined for measuring the association between more than two words.

Other, more appropriate measures are e.g. the χ^2-measure, the *t-score* and the likelihood ratios: The

χ^2-measure for dependence doesn't assume normally distributed probabilities. χ^2 is defined as:

$$\chi^2 = \sum_{ij} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \quad (5)$$

Frequencies should be 5 or higher in order to apply the chi^2-measures. The t-score and the log-likelihood ratio are better suited for low-frequency data and are defined as:

$$\mathtt{t - score} = \frac{O_{11} - E_{11}}{\sqrt{O_{11}}} \quad (6)$$

$$\mathtt{log - likelihood} = 2 \sum_{ij} O_{ij} * log_2(\frac{O_{ij}}{E_{ij}}) \quad (7)$$

Note, that the likelihood-ratio is more appropriate for sparse data than the χ^2-measure. It is however not defined if w_i or w_j appears only in the pair (w_i,w_j) [Dai94]. [1]

To sum up, all frequency-based techniques assign a numerical value to sets of words. These numerical values allow to extract and rank TCs and to exclude TCs below a certain threshold. The unit-hood problem is not properly addressed for two reasons. First, the measure is applicable only to n-word sequences with a fixed n, e.g. $n = 2$. Secondly, word associations do not respect phrase boundaries, ie. they may identify parts of a phrase or associations as *look at*, where *at* belongs to the following PP.

Another statistical approach aims at the identification of boundaries of TCs. If the boundaries are defined as the first and last word of a TC and the words preceeding and following them, this approach is suitable for TCs of variable length, without requiring larger corpora for the identification of longer TCs. This approach is equally applicable to low frequency TCs. If, however, boundaries are defined as the entire TC and the words preceeding and following it, the problem of sparse data reappears. The boundary is classically gauged via the entropy, but any association measure could be used to locate a boundary there where low associations are found.

3.3 Evaluation

Although much of the usefulness of TE depends on the way how TE programs are integrated into the terminographer's working environment, approaches to TE are frequently evaluated in terms of *recall* and *precision* to which we will add here the *ranked recall* as another criterion [c.f. JM02].

The *recall* describes the capacity to identify all terms contained in a document. It is defined as the number of correctly identified TCs divided by the number of terms in the text:

$$\mathtt{recall} = \frac{\#\{\{TC\} \cap \{T_{doc}\}\}}{\#\{T_{doc}\}} \quad (8)$$

With a recall of 80%, 20% of the terms remain undetected. The recall is often not measured as it is difficult to determine the exact set of terms in a text.
The *precision* describes the accuracy with which words and phrases are classified as terms. If the terminographer has to discard many TCs, the precision is low. The precision is defined as the number of correctly identified TCs divided by the number of all proposed TCs.

$$\mathtt{precision} = \frac{\#\{\{TC\} \cap \{T\}_{doc}\}}{\#\{TC\}} \quad (9)$$

With a precision of 80%, 20% of the TCs are not terms.
As high values of recall often imply low precision scores and vice versa, recall and precision are frequently combined into the harmonic mean as:

$$\mathtt{mean} = \frac{2 * \#\{\{TC\} \cap \{T\}_{doc}\}}{\#\{T_{doc}\} + \#\{TC\}} \quad (10)$$

As TE may produce for a medium-sized text many thousands TCs, it is important to rank them. TCs at the top of the list should be the most probable terms, and TCs at the bottom the less probable terms. The better the TE, the more accurately terms are separated from non-terms. Accordingly, we use the *ranked recall* as a further evaluation criterion. If we define r_i as the rank of the ith $TC|TC \in \{\{TC\} \cap \{T\}_{doc}\}$ then the ranked recall is defined as:

$$\mathtt{ranked\ recall} = \frac{\sum_i^n i}{\sum_i^n r_i} \quad (11)$$

.

In a list of 3 TCs with the second and third $TC \in \{T\}_{doc}$, the ranked recall is $\frac{1+2}{2+3} = 0.6$.

[1] Most word association measures are implemented in the Perl-module *N-gram Statistics Package* which can be freely downloaded from CPAN.

4 TE for Minority Languages

TE can be regarded as one technique that is crucial for the support of minority languages. Nevertheless, very little research has been done in this area. The resources that are usually used for TE rarely exist to the same degree for minority languages. Expensive creation of lingware by trained linguists is often infeasible: Formally trained linguists in the minority language are hard to find and it may take years to build such resources.

An alternative approach is the transfer of a system developed for language A to language B. However, the differences among languages also impede a transfer between languages. E.g. Germanic and Slavonic terms need a compound analysis, while Romance language terms require an analytic analysis. The transfer of approaches or systems from well-developed languages to minority languages therefore seems to be limited to languages of the same language family or type.

[BMA96] explore the potential of the weirdness ratio for TE for small languages, taking Norwegian as an example. They use a 10.000 word specialist text and a 100.000 word general language corpus. For this purpose they use a 10.000 word specialist text and a 100.000 word general language corpus. Following the limitations of this measure, only one-word units are extracted and ranked. For languages which almost exclusively use compounding for term formation, this method may be adequate. For analytical languages, this method leaves out too many terms as we demonstrate below. The weirdness-ratio has also been applied in [AD94], in this case to Welsh, with the specialist text and the general language corpus having each the size of 100.000 words.

[DEJ^{+}00] report on two experiments with Malagasy, an Austronesian language in Africa. In the first experiment, a statistical language-independent TE approach (ANA [EP94]) has been tested on a corpus of 25.000 words. The system has a good precision (about 75%) but a low recall: only about 240 TCs have been extracted. A second experiment tested a hybrid linguistic and statistical approach. In a second experiment a hybrid, linguistic and statistical, approach has been tested which required the prior creation of a dictionary and the training of POS-tagger. This required the creation of a dictionary and the training of POS-tagger, before TE could start. With 819 TCs, the number of the extracted TCs is higher than in the purely statistical approach. Precision rates, however, are not reported. This work excellently documents the difficulties of TE with non-European languages. Possible solutions to the question of how linguistic approaches may be put to work even in difficult circumstances are hinted at. It gives several hints at possible solutions to the question of how linguistic approaches may be put to work even in difficult circumstances.

4.1 Example-based TE

In an attempt to escape from the limitations of the above outlined TE approaches, we investigated the possibility of an example-based approach to TE. *Example-based approaches* in NLP are characterized by the fact that the training material is of the same type as the system's output. Feeding a computer with parse trees in order to train parsing, is an example-based approach to parsing. Feeding a computer with examples of classified documents in order to classify documents is an example-based approach to document classification.

The advantage of example-based approaches over rule-based approaches is that no abstract rules are required. As for the acquisition of the data this means that examples can be extracted automatically or created manually by enumerating positive examples. As for the representation, no complex formalisms or models are required to express the linguistic knowledge. Exceptions and regular phenomena can be listed side by side. The advantage of example-based approaches over statistical approaches is that the system can start even from a single training example.

Tackling TE with an example-based approach requires only a few example terms, e.g. for English *red pepper, information society.* These examples can be traditionally elaborated terms in an existing termbase, thus reflecting the properties of terms. In this case, the termhood and unithood problem may be treated conjointly at the same time. TE with an example set of only nominal phrases will produce nominal phrases and TE with an example set containing also verbal phrases will extract also those. If no terms are available, even dictionary entries may suffice. These entries may be reworked by deleting all word combinations which do not have the structure of terms. The TE can be piloted by the list of example terms, bad results can be amended at any time through the deletion or editing of example terms.

Table 6: Examples drawn from different resources: Termbanks or Dictionaries.

Method	#{TC}	recall	precision	mean
termbank 1225	299	0.7321	0.284	
dictionary	322	0.75	0.269	0.396
mixture	390	0.839	0.248	0.386

Before going into the details of this approach we have to mention some non-example-based filters we use for experimentation. The first filter concerns *function words* (f-words). These are identified automatically and used to exclude TCs with function words in the position of the first or last word [c.f MNA94]. The 100 words with the highest DF are assumed to be function words. A second filter, called *punctuation* assumes that punctuation marks are not part of a TC. All other measures try to establish a similarity between the list of example terms and a TC. These are

the affix pattern,

the graphic pattern,

the length.

The example terms are used to generate two types of term patterns: (1) affix term-patterns and (2) upper-case/lower-case (graphic) term patterns. Both can best be explained with an example. The term *House of Lords* is transformed into the pattern *e—of—*s, by reducing non-function words to their last character. The upper-case/lower-case term pattern creates a *c* for capitalized words, an *l* for lower-case words and an x otherwise. The term *House of Lords* generates the graphic pattern *c—l—c*.

Words are extracted as TCs if they fit to one affix and one graphic pattern, even if coming from different examples. The sequence *purpose of words* would require the patterns *e—of—*s and *l—l—l* to be known in order to pass this filter.

The length of the example terms can be used to calculate a 'good' length of TCs. We defined this good length as the mean (m) of the length of the example terms ±3 standard deviations. Terms which are too small or too long are therefore filtered out.

4.2 Experiments

The experiments are conducted using a text of 994 words, written in the Ladin variant that is spoken in Val Ghardena. The text describes the by-laws of the community and contains, according to a manual examination, 113 terms. We tried to reproduce this perfect manual term extraction with different approaches and their combination. The result of the automatic TE is evaluated in terms of *recall*, *precision*, *mean* and *ranked recall*. We should keep in mind that the best extraction is the one with the highest *mean* while the best ranking is the one with the highest *ranked recall*.

The TE-tool used for our experiments is web-based tool [2] that can be freely used for the Ladin idioms of Ghardena, Badia as well as for Italian, German, English and French. The tool can be easily extended to other languages by adding a term list or word list and some general language texts of the respective language.

In Table 6 different types of training material are compared with respect to the effects on the TE quality. The training material can be a list of related or unrelated terms, a list of dictionary entries or a mixture of both. The results show that, if terms are used as examples, we get a high precision and a low recall. Using dictionary entries as examples enhances the recall and reduces precision. For the experiments to follow, the mixed method will be used.

In Table 7 the results of different TE methods are compared. Table 7 compares the results of the different TE methods. The first row, 'no method', represents the base line, it extracts 19019 possible TCs and reaches the perfect recall of 1 with the worst possible precision for this text, 0.0056. Assuming that terms never feature punctuation marks, only 8023 TCs are extracted with perfect recall. The following two methods exclude all TCs with function words in first or last position or with extreme length. This filter is not sufficiently specific though, because there are still 20 TCs for one term actually contained in the input document. The example-based patterns method on the other hand is more selective than any other and retains a good recall (0.85).

Table 8 shows the effect on the results when combining different methods. This so-called *bagging* of-

[2] http://www.eurac.edu/bistro

Table 7: Different simple methods for TE, tested individually. The ranking of TCs is done via TF.

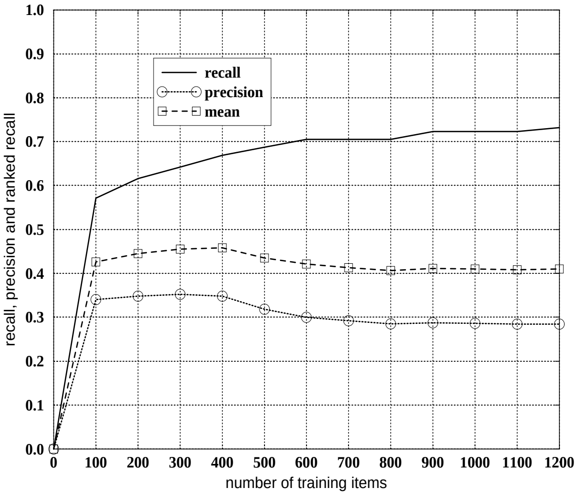
Method	#{TC}	recall	precision	mean	ranked recall
no method	19019	1	0.0056	0.011	0.011
punctuation	8023	1	0.0134	0.026	0.0179
f-words	6289	0.946	0.016	0.033	0.030
length	2419	0.9375	0.044	0.084	0.055
pattern	489	0.848	0.202	0.326	0.388

ten, but not necessarily, improves the quality of the TE. The example-based pattern method has a recall of nearly 85%, so that we start with this method, then trying to improve its precision. The combination with the function word filter is the strictest one. It reduces the recall only by 1% but enhances the precision by 4%.

Table 9 gives the results for the weirdness-ratio method, which only extracts single-word terms ($n = 1$). Table 10 shows the results of the TE with Mutual Information where the number of words is fixed to $n = 2$. With our pattern-based approach such limitations do not exist. The recall column of table 7 shows clearly that only a subset of terms is extracted. With a recall of 54%, 46% of the terms remain undetected. This might still be a good method for compounding languages or for very fundamental terms. The best precision value, 0.255, is yet not better than the precision with unrestricted n. The recall of Mutual Information with only around 10% is quite low, so that an unrestricted n is also better than setting $n = 2$. The results clearly show that free-length approaches are to be preferred over those with a fixed n, because a fixed n drastically reduces the recall without necessarily improving the precision.

Figure 1, Figure 2 and Figure 3 show the learning curve for example-based TE with the examples coming from (a) a term-list, (b) a word list and (c) a mixture of both. The results show a quick rise in recall. While the recall rises continually, the precision drops after a few hundred examples. Apparently, with more training data, no more good term-models are learned, and those term-models which cause noise accumulate. The automatic identification and exclusion of inappropriate term-models is one possible direction for our future research on example-based TE.

Figure 1: The learning curve for Example-based TE. The examples are 1200 terms drawn from a termbank.



5 Conclusions

In this paper we have shown that example-based term extraction offers a feasible approach to TE for minority languages which only needs few or little resources. A few examples, drawn from dictionaries or other terminological data are sufficient to create term-models which cover most terms to be extracted. The texts to be processed can be very short as we have shown. While other approaches, especially statistical approaches require large texts, we could extract about 100 terms from a relatively small text of only 1.000 words.

The proposed example-based approach replaces an in-depth linguistic analysis of the input document. Due to this shallowness, the approach is prone to errors resulting from surface similarities of terms and non-terms. In the same way as linguistic approaches, the example-based approach can be combined with sophisticated statistical ratings when large corpora are available. It can also be combined with linguistic tools like stemmers if those are available.

Table 8: Combination of simple methods for TE. *Example-based term patterns* + *function words* win.

Method	#{TC}	recall	precision	mean
pattern	489	0.848	0.202	0.326
pattern + punctuation	489	0.848	0.202	0.326
pattern + length	477	0.839	0.203	0.328
pattern + f-words	390	0.839	0.248	0.386

Table 9: Extraction of 1-word TCs with the weirdness ratio. This approach is not satisfying and can only be marginally improved.

Method	#{TC}	recall	precision	mean	ranked recall
weirdness ratio	345	0.544	0.188	0.280	0.363
weirdness ratio + pattern	312	0.544	0.210	0.303	0.415
weirdness ratio + length	316	0.544	0.205	0.298	0.400
weirdness ratio + length + pattern	302	0.544	0.215	0.308	0.416
weirdness ratio + f-words	281	0.544	0.225	0.318	0.367
weirdness ratio + f-words + pattern	250	0.544	0.254	0.346	0.404
weirdness ratio + f-words + pattern + length	249	0.544	0.255	0.347	0.404

Table 10: Extraction of 2-word TCs with Mutual Information. Limiting n to 2 leaves out many terms. An approach not to be followed.

Method	#{TC}	recall	precision	mean	ranked recall
MI (2 word terms)	807	0.098	0.013	0.024	0.007
MI + pattern	160	0.099	0.063	0.074	0.064
MI + pattern + f-words	69	0.098	0.144	0.110	0.144

Figure 2: The learning curve for Example-based TE. The examples embrace 5000 dictionary entries.

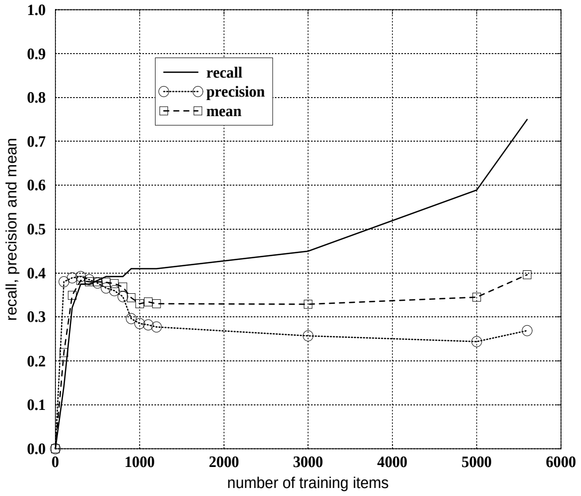
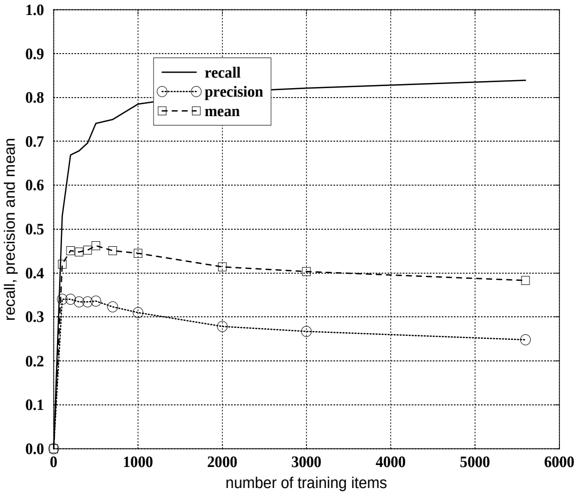


Figure 3: The learning curve for Example-based TE. The examples are mixed terms and dictionary entries.



The approach is applicable to many languages, however, it is not language-independent. Different patterns of compounding, inflection and derivation will certainly influence the quality of the approach. It might work better for Romance languages than for Germanic or Slavonic languages. But even within language groups there may be differences. The approach might work better for Bulgarian than for Russian and better for Italian than for French. By no means however, this implies a failure of the example-based approach, as other kinds of easy-to-calculate similarities between terms and TCs might be used [c.f. ELR$^+$96]. Currently we envisage to apply the approach to other Romance minority languages. The so-extended TE-tools will continue to be free to use under `http://www.eurac.edu/bistro`.

References

[AD94] Khurshid Ahmad and A.E Davies. ’weirdness’ in special-language text: Welsh radioactive chemicals texts as an exemplar. *Journal of the International Institute for Terminology Research*, 5(2):22–52, 1994.

[Arp95] Antti Arppe. Term extraction from unrestricted text. Helsinki, May 30-31 1995. Short Paper presented at the 10th Nordic Conference of Computational Linguistics (NoDaLiDa).

[BJ99] Didier Bourigault and Christian Jacquemin. Term extraction + term clustering. An integrated platform for computer-aided-terminology. In *Proceedings of EACL*, pages 15–22. Bergen, 1999.

[BJL01] Didier Bourigault, Christian Jacquemin, and Marie-Claude L’Homme, editors. *Recent Advances in Computational Terminology*, Natural Language Processing, John Benjamins, Amsterdam, 2001.

[BMA96] Magnar Brekke, Johan Myking, and Khurshid Ahmad. Terminology management and lesser-used living languages: A critique of the corpus-based approach. In *[San96]*, pages 179–189. 1996.

[CCEBVP01] M. Teresa Cabré Castellví, Rosa Estopà Bagot, and Jordi Vivaldi Palatresi. Automatic term detection: A review of current systems. In *[BJL01]*. 2001.

[CH89] Kenneth W. Church and Patrick Hanks. Word association norms, mutual information and lexicography. In *27th Annual Meeting of the ACL*, pages 76–83, Vancouver, 1989.

[Dai94] Béatrice Daille. Combined approach for terminology extraction: lexical statistics and linguistic filtering. Université Paris VII, March 1994.

[DEJ+00] Béatrice Daille, Chantal Enguehard, Christine Jacquin, Rabaovololona Lucie Raharinirina, Baholisoa Simone Ralalaoherivony, and Christian Lehmann. Traitement automatique de la terminologie en langue malgache. In Karim Chibout et al., editor, *Ressources et évaluation en ingénerie des langues, Acutalités scientifique-Universités Francophones*, pages 225–242. De Boek and Larcier s.a, 2000.

[DHJ96] Béatrice Daille, Benoî̂t Habert, and Christian Jacquemin. Empirical observations of term variations and principles for their desciption. *Terminology*, 3(2):197–285, 1996.

[ELR+96] F.ç. Ekmekçioglu, M.F. Lynch, A.M. Robertson, T.M.T. Sembok, and P. Willett. Comparison of n-gram matching and stemming for term conflation in English, Malay, and Turkish texts. *Text Technology*, 6:1–14, 1996.

[EP94] Chantal Enguehard and Laurent Pantera. Automatic natural acquisition of a terminology. *Journal of Quantitative Linguistics*, 2(1):27–32, 1994.

[Hei99] Ulrich Heid. Extracting terminologically revelant collocations from german technical texts. In *[San99]*, pages 241–255. 1999.

[HFH01] Munpyo Hong, Sisay Fissaha, and Johann Haller. Hybrid filtering for extraction of term candidates from German technical texts. In *Proceedings of Terminologie et Intelligence Arteficielle, TIA'2001*, Nancy, May 3-4 2001.

[Jac99] Christian Jacquemin. Syntagmatic and paradigmatic representation of term variation. In *ACL'99*, pages 341–348, 1999.

[JM02] Peter Jackson and Isabelle Moulinier. *Natural Language Processing for Online Applications.* Natural Language Processing. John Benjamins, Amsterdam, 2002.

[MA99] Diana Maynard and Sophia Ananiadou. Identifying contextual information for multi-word term extraction. In *[San99]*, pages 212–222. 1999.

[MA00] Diana Maynard and Sophia Ananiadou. Identifying terms by their family and friends. In *COLING'2000*, pages 530–536. Saarbrücken, 2000.

[MM00] Magnus Merkel and Andersson Mikael. Knowledge-lite extraction of multi-word units with language filters and entropy thresholds. In *Proceedings of RIAO*, volume 1, pages 737–746. Collège de France, Paris, April 12-14 2000.

[MNA94] Magnus Merkel, Bernt Nilsson, and Lars Ahrenber. A phrase-retrieval system based on recurrence. In *Proceedings of the Second Annual Workshop on Very Large Corpora (WVLC-2)*, pages 43–56, Kyoto, 1994.

[Nak01] Hiroshi Nakagawa. Experimental evaluation of ranking and selection methods in term extraction. In *[BJL01]*, pages 303–325. 2001.

[PL01] Patrick Pantel and Dekang Lin. A statistical corpus-based term extractor. In Eleni Stroulia and Stan Matwin, editors, *AI 2001, Lecture Notes in Artificial Intelligence*, pages 36–46. Springer, Ottawa, Canada, 2001.

[San96] Peter Sandrini, editor. *Proceedings of Terminology and Knowledge Engineering (TKE'96)*, Innsbruck, 1996. TermNet.

[San99] Peter Sandrini, editor. *Proceedings of Terminology and Knowledge Engineering (TKE'99)*, Vienna, 1999. TermNet.

[SVT99] Pirjo Soininen, Atro Voutilainen, and Pasi Tapanainen. An experiment in automatic term extraction. In *[San99]*, pages 234–241. 1999.

[SW99] Antje Schmidt-Wigger. Term checking through term variation. In *[San99]*, pages 570–581, 1999.