

Term Extraction for Ladin: An Example-based Approach

Oliver Streiter, Daniel Zielinski, Isabella Ties and Leonhard Voltmer
European Academy, 39100 Bolzano/Bozen, Italy
{ostreiter;dzielinski;ities;lvoltmer}@eurac.edu

Mots-clefs – Keywords

langues minoritaires, extraction terminologique basé sur les exemples
minority languages, example-based term extraction, n-gram similarity

Résumé - Abstract

Cette communication traite le problème de l'extraction de termes pour les langues minoritaires. Nous présentons une méthode basée sur des exemples qui fonctionne même si les ressources linguistiques digitales sont rares. Notre méthode se base sur modèles de termes générés à partir d'un nombre limité de termes d'exemple. Les résultats obtenus pour le Ladin du Val Gherdena sont meilleurs que ceux des approches statistiques simples à l'extraction de termes.

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 Gardena, which we use in our experiments, the example-based approach outperforms simple statistical approaches to TE.

1 Introduction

Minority Languages have few speakers, few native linguists and even fewer computational linguists. If those languages have a writing system, they often may not have strict writing rules. They always lack adequate corpora and financial support. Under such conditions, which are the approaches to be followed? Statistical approaches of corpus linguistics need large amounts of language resources. Rule-based approaches (for tagging and parsing) require expensive skilled workers. Technology transfer from other languages often fails if designed specifically for one language or language family. This is the case for shallow NLP techniques which make implicit assumptions about the language. Therefore, example-based approaches seem to be promising. Required are a relatively small number of specific examples which can be created by any native speaker.

Among the most basic needs of a minority language figures the creation and management of terminology. This is the situation for the different idioms of Ladin spoken and written in the Dolomites (Italy). In 1989 Ladin has received official status in the Ladin valleys of Badia and Gardena and in 1993 in Val di Fassa. Since then, legal documents have been written in Ladin.

Term extraction helps creating terminology from texts. We will see now what this task implies, what solutions the scientific literature proposes and how our example-based approach rates amongst them: In Section 2 we describe modern approaches to TE by reference to the unithood-problem and the termhood-problem. In Section 2.3 we describe evaluation techniques for TE. In Section 3 we review the past experiments in the field and propose in Section 3.1 an example-based approach to TE and related this approach to those cited in the literature (Section 3.2).¹

2 Term extraction

2.1 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. Traditionally, TE is seen as intersection of two problems. The *Unithood Problem* is the task to select language units from a set of word combination (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 combinations which fulfill the requirements of a term (e.g. *red pepper* but not *red car*). More often than not, the unithood problem is solved first and the output TCs are checked for their term-status (the termhood-problem).

2.2 Approaches

Approaches to TE can be classified according to the knowledge used as linguistic or statistic approaches. They may be combined in hybrid approaches in order to join the strong aspects of

¹Abbreviations used in the paper: TC == Term candidate; TCF == Frequency of TC in a given document; DF == Document Frequency, the number of documents with 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; $\#\{\dots\}$ == The cardinality of a set.

Table 1: Approaches to TE, an overview.

Linguistic		methods	publications
intrinsic		POS-tagging, chunking	(Bourigault & Jacquemin, 1999)
		stop-words	(Merkel & Mikael, 2000)
extrinsic	syntagmatic	full parsing	(Arppe, 1995; Soininen <i>et al.</i> , 1999)
		term variation	(Jacquemin, 1999)
Statistical		methods	publications
intrinsic		mutual information	(Church & Hanks, 1989)
		likelihood ratio	(Hong <i>et al.</i> , 2001)
extrinsic	syntagmatic	nc-value	(Maynard & Ananiadou, 1999)
		entropy	(Merkel & Mikael, 2000)
		c-value	(Nakagawa, 2001)
		weirdness	(Brekke <i>et al.</i> , 1996)

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 TC s and $\{T\}$).

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 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 in addition to other criteria.

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 is that documents are characterized by the repeated use of certain lexical units or morpho-syntactic constructions. We discuss only the more elementary measures.

(1) Frequency of occurrence: The more frequently a lexical unit appears in a given document the more likely it is that this unit has a special function or meaning. Yet extracting TCs just by frequency would also render frequently appearing combinations of function words as TCs. Even in combination with a filter for certain morpho-syntactic patterns, this approach is not always satisfactory.

A second assumption is that linguistic expressions which characterize a document are frequent within a document but infrequent across different documents. One measure to capture this idea is TF.IDF (Equation 2 in Annex), widely used in information retrieval. It 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 same assumption is expressed in the weirdness-ratio (Equation 3) which uses relative frequencies for TF and IDF (Brekke *et al.*, 1996).

The main problem with frequency-based approaches is that they may work well for one-word

units, but do not 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.

The frequency of a TC is one type 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) (Tab. 6). The contingency table contains the observed frequencies of (A,B), (A,notB), (notA,B) and (notA,notB), marked here as $O_{11} \dots O_{22}$. 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 $E_{11} \dots E_{22}$. 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. The simple frequency corresponds to O_{11} in this table.

Mutual Information (MI) (Equation 4) measures the association between two units. This measure is used frequently in corpus linguistics, even though it works badly for low-frequency events. The MI can be 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 as in Equation 5. A singular 3-word expression in a 100 word corpus would consequently receive the MI of $\sim 10.000!$

Other, more appropriate measures are e.g. the χ^2 -measure, the *t-score* and the likelihood ratios: The χ^2 -measure for dependence (Equation 6) doesn't assume normally distributed probabilities. Frequencies should be 5 or higher in order to apply the *chi*²-measures. The *t-score* (Equation 7) and the log-likelihood ratio (Equation 8) are better suited for low-frequency data. The latter however is not defined if w_i or w_j appears only in the pair (w_i, w_j) (Daille, 1994).²

To sum up, all frequency-based techniques assign a numerical value to sets of words to 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 preceding and following them, this approach is suitable for TCs of variable length, without requiring larger corpora for the identification of longer TCs. If boundaries are defined as the entire TC and the words preceding and following it, the problem of sparse data reappears. The boundary is classically gauged via

²Most word association measures are implemented in the Perl-module *N-gram Statistics Package* which can be freely downloaded from CPAN.

the entropy, but any association measure could be used to locate a boundary there where low associations are found.

2.3 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*, *precision*, *mean* and the *ranked recall*.

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. With a recall of 80%, 20% of the terms remain undetected.

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. With a precision of 80%, 20% of the TCs are not terms.

High values of recall often imply low precision scores and vice versa. Therefore recall and precision are frequently combined into the harmonic *mean*.

TE may produce for a medium-sized text many thousands TCs and it is important to rank them. We use the *ranked recall* as a further evaluation criterion. If we define r_i as the rank of the i th $TC|TC \in \{\{TC\} \cap \{T\}_{doc}\}$ then the ranked recall is defined in Equation 1: In a list of 3 TCs with the second and third $TC \in \{T\}_{doc}$, the ranked recall is $\frac{1+2}{2+3} = 0.6$.

3 TE for Minority Languages

(Brekke *et al.*, 1996) 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. 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 (Ahmad & Davies, 1994), in this case to Welsh, with the specialist text and the general language corpus having each the size of 100.000 words.

(Daille *et al.*, 2000) report on two experiments with Malagasy, an Austronesian language in Africa. In the first experiment, a statistical language-independent TE approach (ANA (Enguehard & Pantera, 1994)) 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. 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 documents the difficulties of TE with non-European languages. It gives several hints at possible solutions to the question of how linguistic approaches may be put to work even in difficult circumstances.

Table 2: 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

3.1 Example-based 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 to train parsing, is an example-based approach to parsing. Feeding a computer with examples of classified documents 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 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 few training examples.

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 term-base, 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, dictionary entries may suffice. These entries may be reworked manually in order to improve TE.

Some non-example-based filters are used for experimentation. The first filter concerns *function words* (f-words). These are identified automatically and used to exclude TCs with function words as first or last word (Merkel *et al.*, 1994). The 100 words of the background corpus 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.

Example-based filters are (1) affix term-patterns (2) upper-case/lower-case (graphic) term patterns and length filters. These can best be explained with an example. The Latin term *tofla de comune* is transformed into the pattern **a—de—*e*, 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 *tofla de comune* generates the graphic pattern *l—l—l*.

Words are extracted as TCs if they fit to one affix and one graphic pattern, even if coming from different examples. The sequences of words *ciasa de comune*, *cuntlameda de comune* etc would then be recognized as TC.

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.

Table 3: Combination of simple methods for TE.

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

Terms which are too small or too long are therefore filtered out.

3.2 Experiments

The experiments are conducted using a text of 994 words, written in the Latin variant that is spoken in Val Gardena. The text describes the by-laws of the community and contains, according to a manual examination, 113 terms.

In Table 7 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.

Table 2 compares the results of the different TE methods. The first row, 'no method', represents the base line. 19019 TCs are extracted with the perfect recall of 1 and the precision of 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. The example-based patterns method on the other hand is more selective than any other and retains a good recall (0.85).

Table 3 shows the effect on the results when combining different methods. We start with the example-based method and try to improve its precision. The combination with the function word filter reduces the recall only by 1% but enhances the precision by 4%.

Table 4 gives the results for the weirdness-ratio method, which only extracts single-word terms ($n = 1$). 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. Table 5 shows the results of the TE with Mutual Information with $n = 2$. The recall of Mutual Information with only around 10% is quite low. 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: The best precision value, 0.255, is yet not better than the precision with unrestricted n .

Figure 1, shows 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.

Table 4: Extraction of 1-word TCs with the weirdness ratio.

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

Table 5: Extraction of 2-word TCs with Mutual Information.

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

4 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 input texts can be very short. 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 and with linguistic tools like stemmers.

The TE tools is freely available under <http://dev.eurac.edu:8080/perl/all.tar.gz>. A graphical interface is provided with Bistro <http://dev.eurac.edu:8080>. Currently we exploit within the Project Logos Gaias (<http://logos-gaias.themenplattform.com>) the integration of the example-base TE tools into GYMN@ZILLA (Streiter *et al.*, 2003), for the purpose of classifying, indexing and pedagogic elaboration of documents.

References

- AHMAD K. & DAVIES A. (1994). 'weirdness' in special-language text: Welsh radioactive chemicals texts as an exemplar. *Journal of the International Institute for Terminology Research*, 5(2), 22–52.
- ARPE A. (1995). Term extraction from unrestricted text. Helsinki. Short Paper presented at the 10th Nordic Conference of Computational Linguistics (NoDaLiDa).

Term Extraction for Latin: An Example-based Approach

BOURIGAULT D. & JACQUEMIN C. (1999). Term extraction + term clustering. An integrated platform for computer-aided-terminology. In *Proceedings of EACL*, p. 15–22. Bergen.

D. BOURIGAULT, C. JACQUEMIN & M.-C. L'HOMME, Eds. (2001). *Recent Advances in Computational Terminology*, Natural Language Processing, John Benjamins, Amsterdam.

BREKKE M., MYKING J. & AHMAD K. (1996). Terminology management and lesser-used living languages: A critique of the corpus-based approach. In (*Sandrini, 1996*), p. 179–189.

CHURCH K. W. & HANKS P. (1989). Word association norms, mutual information and lexicography. In *27th Annual Meeting of the ACL*, p. 76–83, Vancouver.

DAILLE B. (1994). Combined approach for terminology extraction: lexical statistics and linguistic filtering. Université Paris VII.

DAILLE B., ENGUEHARD C., JACQUIN C., RAHARINIRINA R. L., RALALAOHERIVONY B. S. & LEHMANN C. (2000). Traitement automatique de la terminologie en langue malgache. In K. C. ET AL., Ed., *Ressources et évaluation en ingénierie des langues, Actualités scientifiques- Universités Francophones*, p. 225–242. De Boek and Larcier s.a.

ENGUEHARD C. & PANTERA L. (1994). Automatic natural acquisition of a terminology. *Journal of Quantitative Linguistics*, 2(1), 27–32.

HONG M., FISSAHA S. & HALLER J. (2001). Hybrid filtering for extraction of term candidates from German technical texts. In *Proceedings of Terminologie et Intelligence Artificielle, TIA'2001*, Nancy.

JACQUEMIN C. (1999). Syntagmatic and paradigmatic representation of term variation. In *ACL'99*, p. 341–348.

MAYNARD D. & ANANIADOU S. (1999). Identifying contextual information for multi-word term extraction. In (*Sandrini, 1999*), p. 212–222.

MERKEL M. & MIKAEL A. (2000). Knowledge-lite extraction of multi-word units with language filters and entropy thresholds. In *Proceedings of RIAO*, volume 1, p. 737–746. Paris: Collège de France.

MERKEL M., NILSSON B. & AHRENBER L. (1994). A phrase-retrieval system based on recurrence. In *Proceedings of the Second Annual Workshop on Very Large Corpora (WVLC-2)*, p. 43–56, Kyoto.

NAKAGAWA H. (2001). Experimental evaluation of ranking and selection methods in term extraction. In (*Bourigault et al., 2001*), p. 303–325.

P. SANDRINI, Ed. (1996). *Proceedings of Terminology and Knowledge Engineering (TKE'96)*, Innsbruck. TermNet.

P. SANDRINI, Ed. (1999). *Proceedings of Terminology and Knowledge Engineering (TKE'99)*, Vienna. TermNet.

SOININEN P., VOUTILAINEN A. & TAPANAINEN P. (1999). An experiment in automatic term extraction. In (*Sandrini, 1999*), p. 234–241.

STREITER O., KNAPP J. & VOLTMER L. (2003). Gymn@zilla: A browser-like repository for open learning resources. In *ED-Media, World Conference on Educational Multimedia, Hypermedia & Telecommunications*, Honolulu, Hawaii.

Table 6: Contingency table of observed frequencies $O_{11} \dots O_{22}$ for the word pair (A,B) (top) and for estimated frequencies $E_{11} \dots E_{22}$ under independency assumption (bottom).

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

Table 7: Examples drawn from different resources: Termbanks or Dictionaries.

Method	$\#\{TC\}$	recall	precision	mean
termbank	299	0.7321	0.284	0.41
dictionary	322	0.75	0.269	0.396
mixture	390	0.839	0.248	0.386

Equations:

$$\text{ranked recall} = \frac{\sum_i^n i}{\sum_i^n r_i} \quad (1)$$

$$\text{TF.IDF}_x = \frac{TCF_x}{DF_x} \quad (2)$$

$$\text{weirdness ratio}_x = \frac{\frac{TCF_x}{\#\{TC\}}}{\sum_{d=1}^{d=m} doc_j} \quad (3)$$

$$\text{MI} = \frac{O_{11}}{E_{11}} \quad (4)$$

$$\text{MI}_3 = \frac{O_{111}}{E_{111}} \quad (5)$$

$$\chi^2 = \sum_{ij} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \quad (6)$$

$$\text{t-score} = \frac{O_{11} - E_{11}}{\sqrt{O_{11}}} \quad (7)$$

$$\text{log-likelihood} = 2 \sum_{ij} O_{ij} * \log_2\left(\frac{O_{ij}}{E_{ij}}\right) \quad (8)$$

Figure 1: Learning curves with 1200 example terms (left), 5000 dictionary entries (middle) and a mixture of both (right).

