Application-based Evaluation of Automatic Terminology Extraction

Marija Brkic Bakaric[1], Nikola Babic[2], Maja Matetic[3]
Department of Informatics
University of Rijeka
Rijeka, Croatia

***Abstract*—The aim of this paper is to evaluate performance of several automatic term extraction methods which can be easily utilized by translators themselves. The experiments are conducted on German newspaper articles in the domain of politics on the topic of Brexit. However, they can be easily replicated on any other topic or language as long as it is supported by all three tools used. The paper first provides an extensive introduction into the field of automatic terminology extraction. Next, selected terminology extraction methods are assessed using precision with respect to the gold standard compiled on the same corpus. Moreover, the corpus has been completely annotated to allow for the calculation of recall. The effects of using five cut-off points are examined in order to find an optimal value which should be used in translation practice.**

***Keywords*—Terminology extraction; hybrid methods; evaluation; precision; recall; gold standard; language resources**

I. Introduction

The aim of this paper is to provide an extensive introduction into the field of automatic terminology extraction (ATE) and to evaluate one statistical and two hybrid methods of extraction on German newspaper articles in the domain of politics on the topic of Brexit. The main highlights of the research are the following: (1) extensive introduction into the field of terminology extraction followed by the (2) creation of the gold standard on the Brexit-related German terminology and (3) evaluation of the selected methods on the Brexit-related jargon which shows abundance of creative coinages by (3a) conducting both manual and automatic extraction on the same corpus, and (3b) by annotating the corpus completely in order to allow for the calculation of recall.

The problem of the related work is that it is not directly comparable due to the differences in "corpus selection (e.g. domain, size), evaluation methodology (e.g. human judges, dictionary based, gold standard based), and scope (e.g. entire results, parts of results, top *n* best results)" [1]. With the above said in mind, this study does not propose a novel approach to ATE, but compares performance of well-known extraction methods under the same experimental settings. Since the use of the gold standard supports reproducibility of results and comparison between different methods, this paper opts for that approach. Although there are toolkits such as JATE 2.0 [2] or ATR4S [3] which implement more than ten automatic terminology extraction methods, these toolkits, as well as other related software toolkits are rather limited for several reasons – some of them lack the adequate language support, some cannot be used by the users who need these tools in practice but do not have enough technical expertise, e.g. translators, and lastly, some are proprietary. Moreover, the evaluation of ATE is usually conducted in technical domains such as biology or medicine, as acknowledged by [1]. Therefore, evaluation in less technical domains is missing.

The following section not only presents the related work, but it can serve as an introduction for those who wish to enter the field of terminology extraction. The experimental study is presented in section three, which is subdivided into descriptions of corpus, and manual and automatic extraction tasks. The results and discussion are given in section four. Concluding remarks are provided in the last section.

II. Background

A. Basic Concepts and Definitions

Terminology extraction aims at "structuring terminological knowledge from unstructured texts" [4] and identifying "the core vocabulary of a specialized domain" [5]. Terms can be defined as a "designation of a defined concept in a special language by a linguistic expression" (ISO 1087). Terms are usually nominal constructions, while collocations represent preferred ways of expressing things and thus contain more verbal parts [6]. For an overview of the existing definitions for the concepts "term" and "domain", please refer to [7].

In the Traditional Manual Terminology Extraction (MTE), a terminologist first makes a list of potential term candidates (TCs) which are then discussed with domain experts. The resulting list contains all validated terms [5]. Automatic terminology extraction (ATE) is based on the computational analysis of a textual corpus. The process is carried out by the computer and is thus objective. The fact that ATE is based on objective corpus evidence compensates for possible human errors [8]. On the other hand, humans identify terms not only by form, but also according to extra-linguistic criteria, and the terms detected on the basis of semantics also have to fit domain. The automatic process can therefore only assist humans who must be engaged during the final verification or filtering stage [9]. As long as humans are needed at least in the verification stage, the process of ATE will be considered semi-automatic [10]. However, since MTE is error-prone, labor intensive, time-consuming, and subjective, ATE is useful even if used only as a "preliminary identification" of TCs [5].

This research was supported by the Erasmus+ grant number 19-203-060377 – KA2-HE-01/19 and the University of Rijeka grant number uniri-drustv-18-122.

In terms of ATE, designation of a word or a phrase as a term is not a simple binary decision. ATE result is presented as a continuum, in the form of a list of candidates ranked according to the score [11].

There are two types of terminology extraction from unstructured texts – monolingual terminology extraction which processes and extracts terms from texts in one language, and bilingual or multilingual terminology extraction which extracts and aligns terms from texts in two or more languages [4]. This paper is concerned with the first, while the latter will be subject of our future work.

B. Applications of ATE

Three possible applications of ATE, which are identified in [6] and [9], refer to terminology, translation, and document management or retrieval (e.g. automatic keyword extraction [12]). This research has been conducted from the translation aspect. In terms of translation, the extracted terminology might be used as a preparation for interpreting, for the development and improvement of machine translation engines, for ensuring consistency, particularly if multiple translators work on the same project [13], etc. The practical requirement is usually to find everything the system does not know yet, so for this purpose term extraction is followed by term recognition, i.e. the comparison of the extraction results with some dictionary/term bank resource in order to identify known/unknown terms.

C. Approaches to ATE

According to the authors in [12], ATE methods can be analyzed according to two aspects. The first aspect is "unithood", which is defined as "the degree of strength or stability of syntagmatic combinations and collocations" and refers to the internal coherence of language units [14] or to "the identification of linguistic elements that constitute a multiword unit and refer to one conceptual unit" [5]. The second aspect, termhood, on the other hand, refers to "the degree that a linguistic unit is related to domain-specific concepts" [12] or to the affiliation of a certain lexical unit or group to a terminology of a special purpose domain. In simple words, termhood detection is a method which ranks the extracted units according to the likelihood that they constitute a valid term for the specialized domain considered.

There are two basic approaches to the process of ATE – linguistic and statistical. Depending on the method used for ATE, the corpora might undergo pre-processing like lemmatization, part-of-speech (POS) tagging, chunking or full syntactic parsing [5]. Linguistic approaches are thus heavily language dependent. They use morpho-syntactic patterns, while statistical approaches use terms frequencies as evidence for unithood [5]. Co-occurrence measures for unithood include chi-square, t-score, log-likelihood ratio, mutual information, and the phi coefficient. The termhood can be measured by analyzing contextual usage of TCs, TCs' internal structure, or distributional properties of TCs within the domain and the dispersion over different documents [5]. Most of the contemporary systems are hybrid, which means that they are based on the combination of two approaches [4]. This implies that the classification into statistical and linguistic is deprecated, and that linguistic methods are nowadays regarded as mere filters. Co-occurrence measures are therefore usually calculated for word combinations that have passed the linguistic filter. The filter can be either open-class and thus less restrictive, which results in huge lists abundant with false positives, or closed class, boosting precision at the cost of recall [5]. Beside representative domain specific corpus, contrastive approaches to ATE additionally require a general language corpus. Additionally, they should be coupled with word sense disambiguation, since many terms ('belt', 'fault', etc.) are homonymous between a term reading and a general reading [6]. The methods selected for this research can be roughly categorized into statistical (Rainbow), hybrid (Termsuite), and hybrid contrastive (Sketch Engine).

Majority of ATE methods follow the scheme given in Fig. 1, which is based on the work of [7]. For an extensive list of feature computation methods and their classification, please refer to [7].

D. Evaluation of ATE

ATE evaluation methods can be divided into direct (intrinsic) and indirect (extrinsic) methods, as asserted by [14]. While the first evaluate some intrinsic properties, the latter measure the improvement gained in another system which uses the results of term extraction.

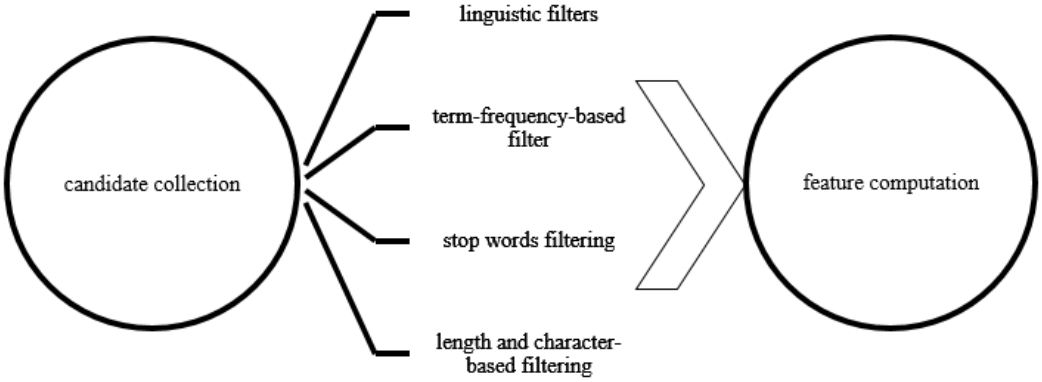


Fig. 1. General ATE Pipeline (based on [7]).

Furthermore, there are two direct approaches to ATE evaluation. The first approach is to have domain experts conduct a manual evaluation. This approach suffers from low inter-annotator agreement. It can be conducted in a strict mode in which all evaluators have to agree on the term, and lenient in which all candidates approved by at least one evaluator are counted [1]. The other approach is to use the "gold standard", also known as Reference Term Lists (RTLs), which implies assessing the quality relative to a list of terms manually compiled by a domain specialist [5]. However, the main difficulty of the approach is that there are no objective rules to distinguish terms from non-terms [15]. RTLs "may be more or less detailed, depending on terminologists' needs and preferences" [16], which significantly affects evaluation. Moreover, RTLs should contain not only reference terms, but also their variants. In relation to the gold standard, the authors in [17] differentiate between two types of true positives – actual true positives (ATP) which relate to all the terms in the gold standard, and recoverable true positives (RTP) which relate to the intersection of the filtered term candidates and the gold standard. In this research, in this paper the direct approach of using the gold standard is taken.

The approach of using the gold standard can be further subdivided, as the authors in [7] suggest, into labelling of a whole dataset, labelling of a subset, and, lastly, adaptation of available resources, which depends on the availability of domain thesauri, vocabularies, keywords or indexes.

The measures which are usually used for ATE evaluation are precision and recall, and, lately, average precision. These are the measures used also in this research. Precision (P) shows the percentage of correct terms out of the list of extracted terms (1). It is sometimes referred to as precision at level K (P@K). Recall (R) shows the percentage of correct terms out of the list of manually extracted terms (2). It is worth noting, as the authors in [7] emphasize, that the recall is usually implicitly evaluated because it is determined by the specified number of recognized terms. Average precision (AvgP), given in (3), is a standard ATE metric. If recall cannot be calculated as no gold standard exists, the union of all the correct terms predicted by methods which are being evaluated can be used for calculating at least a relative recall [18].

$$P@K=\frac{correct \cap retrieved}{K} \quad (1)$$

$$R@K=\frac{correct \cap retrieved}{correct} \quad (2)$$

$$AvgP@K=\frac{1}{K}\sum_{k=1}^{K} P@k \quad (3)$$

A slightly different measure of precision is sometimes used (e.g. [1]), which is taken over from [19]. It averages precision at the i^{th} correct term (with respect to how many candidate terms preceded the i^{th} term) out of the total K correct terms in the output. There is also a version which takes recall into account, as given in [3].

Since precision and recall have many underlying problems (e.g. there is no intuition as to which terms are considered relevant, or the system gives a term in its base form but the reference term bank has it in inflected form, should proper names be included or not, etc.), the work of [6] proposes the evaluation measure which relies on the issue of usability, which is calculated as the ratio between really possible concepts (semantic units) and candidates which will never be used in any application, and are only noise.

The evaluation procedure is additionally burdened by the fact that two actors of different profiles are involved in the manual compilation. A terminologist is an expert on deciding whether an expression is a real term or belongs to the general language, while a domain expert uses a specific expression to refer to a concept in the domain [14]. Since it is not realistic to have a terminologist available for all the task types, two domain specialists are employed for the task. In general, it is difficult to obtain the complete set of terms in a given corpus (i.e. it is easier to ask a specialist about the termhood of a given TC than to ask him or her to compile the complete list of terms),which is why only precision is sometimes calculated [14]. If the gold standard is present, as in this research study, there is a possibility to accept the TCs listed in the standard and to ask human evaluators to judge the remaining TCs.

E. Thresholds

The authors in [7] list different scenarios regarding the number of terms to be recognized and distinguish between those with a predetermined number of terms (cut-off value) and those in which the number of terms is determined by the algorithm. A hard threshold means that candidates with scores less than a threshold are not accepted. A top list, on the other hand, takes a certain number of candidates into account, while there is also a top percentage version which takes a certain percentage of top candidates into account. The authors in [7] also enumerate scenarios based on the length of term candidates.

In this research, the length of the candidates is not restricted. Although different cut-off values are explored, the paper also reports results when the complete TC lists are taken into account.

F. Bilingual ATE

When it comes to bilingual ATE or bilingual glossary compilation, there are many more challenges that need to be tackled compared to the monolingual ATE, e.g. the usage of terms is often not harmonized, especially in case of more translators or authors, mostly due to the differences and particularities of two systems. However, ATE can serve a valuable purpose of highlighting terminology issues and facilitating harmonization of terminology [20]. The evaluation should be conducted by assessing whether a target phrase is the translation of a source term or not [6].

G. Related Work

Although there is a huge line of work on ATE, in this subsection only those that at least distinctly relate to this study are presented. One of the first limitations in conducting this type of study is appropriate language support. The existing tools or frameworks support only a handful of languages.

The authors in [21] evaluate nine terminology extraction tools from a translator's perspective. They distinguish three classes of tools – standalone terminology extraction tools (e.g. Termsuite), web-based terminology extraction tools (e.g.

SketchEngine), and frameworks (e.g. Rainbow). This study is limited to a handful of tools since a massive approach would not contribute neither to comprehensibility nor to drawing clear conclusions. However, one tool representative of each category based on the criterion of user-friendliness is included. The clarity of making conclusions is affected even with such a limitation imposed, as shall be seen later in the paper.

By underlining the quality issues that ATE often exhibits, the results in [18] suggest that bare frequency may not be sufficient to extract even correct single-word terms (SWTs), but also that single words which occur frequently have a high chance of being identified as terms specific to that corpus. Furthermore, the performance of a POS tagger plays a great role in detecting terms, especially with regard to recall. While SWTs are usually too polysemous and too generic, multi-word terms (MWTs) often represent finer concepts in a domain [22]. Although MWTs might predominate in some languages thus making SWT extractors almost useless [6], [9], in some studies majority of domain specific terms turn out to be compound nouns [23]. The authors in [24] also acknowledge that in German texts the use of SWTs is frequent. More precisely, they report that the proportion of SWTs in three different subject domains of technical documents ranges from 57 to 94%. For that reason, the focus is put on hybrid tools, although one statistical tool is also included in this research for the sake of comparison.

Majority of the related work on German employs a hybrid approach. For example, the research presented in [24] combines linguistic filtering techniques with a statistical technique in order to extract noun-phrases from technical texts. The authors in [16] also combine a linguistic filter by using the list of pre-defined patterns with the weirdness ratio in the domains of chemical protection suits and of alcohol and drug detection. The authors in [25] extract nominal candidates in DIY domain. Statistical measures are combined to rank the TCs by the domain specificity after extracting TCs based on POS patterns and filtering out syntactically invalid ones and those occurring only embedded in other candidates. The comparison between the results of well-known statistical measures in the domain of grammar in [8] shows that measures based on corpus comparison outperform all others. The authors attribute this to the fact that German word formation allows for complex compound-unigrams (almost 83% of manually extracted terms are SWTs), which causes weaker performance of algorithms designed to identify MWTs. Three different approaches to ATE are compared, which are also representatives of three different categories of tools.

In this research study the same corpus is used both for manual and automatic term extraction, since one of the related works reveals that 35% of the false positives in the top 500 candidates qualify for the inclusion into the gold standard which is created without full access to the corpus used for automatic extraction [16]. While the authors in [8] manually extract the gold standard from a subset of the corpus and obtain best results for measures based on corpus comparison, the manual extraction in this study is conducted on the whole corpus.

There is little work on the effects that the domain has on ATE. The authors in [1] show that domain has an "impact on the performance of algorithms", as exemplified by the comparative study in the domains of biology and medicine. Moreover, both, language and domain, affect term length as illustrated in [15]. However, according to the findings over 80% of all the terms irrespective of the language and the domain belong to one of eight POS patterns - single nouns (N), a noun and an adjective (N+A), a single adjective (A), a named entity (NE), two nouns (N+N), two nouns separated by a preposition (N+P+N), two adjectives and a noun (N+A+A) or a single verb (V). While there are many N and N+A patterns, substantial differences can be observed between different domains and, even more so, between different languages. The trade-off between precision and recall can be determined by applying a cut-off value since [16] find almost 50% of terms in the gold standard in the top 500. The author in [3] demonstrates that there is no method which performs best on all datasets.

The experimental study presented in the remainder of the paper is designed based on the aims set out in this section and on the findings of the related work presented.

III. Experimental Study

In this experimental study a black box evaluation of the selected terminology extraction tools is conducted from the translators' point of view. Since bilingual term extraction is the most useful feature in the eyes of translators, one tool from each category distinguished in [21] is chosen but only if it supports bilingual term extraction. However, this study is limited to monolingual terminology extraction only. This is done purposefully in order to check what level of quality can be expected in a more simplistic scenario.

The topic chosen for this work is Brexit due to its relative novelty and creativity in word coinage. Its linguistic impact is recognized in [26].

The choice of tools is made logically by satisfying the criteria that the tool has support for German, that it supports bilingual term extraction, and that it has user friendly interface, where a user is considered to be a translator and not a developer. The gold standard list is lemmatized and after term extraction, the same procedure is repeated for each term candidate list. The lemmatized forms are obtained with the python package spaCy. The results are presented in terms of precision and recall at five different levels and in terms of total precision, recall, and average precisions.

A. Language and Corpus

In this study a monolingual German corpus is compiled on the topic of Brexit. A geopolitical change known under the term Brexit has occurred as of recent. The term itself first appeared almost ten years ago. As a result, a multitude of other creative coinages and compounds appeared [26]. German lies somewhere between configurational languages–which encode grammatical relations through the position of constituents–and case languages–which encode grammatical relations through morphological marking [27]. The leading way of word formation in the contemporary German language is compounding [28]. An example is

Meinungsforschungsinstitut, which consists of Meinung (opinion), Forschung (research) and Institut (institute) connected with the letter‘s’. The connecting letter makes it easier to see where one word ends and another begins. Another example of a compound is a German neologism Brexit-Schock (shock caused by Brexit), which is a hybrid made up of one English and one German noun connected with a hyphen. In general, such compounds are hard to translate into other languages and usually require paraphrasing.

The corpus is compiled from German newspaper articles from three different sources–Frankfurter Allgemeine Zeitung, Süddeutsche Zeitung and Zeit, and consists of 50 articles on Brexit in the period of one month. The corpus has 20.409 words in total. The topic of Brexit is chosen since it is very specific for the domain of political newspaper articles, as well as relatively novel. Moreover, as the authors in [29] warn, new or upcoming domains are usually characterized by terminological variation, which may affect the overall results. Due to its specificity and to the design of the experiment elaborated further on, the size of the corpus is relatively small compared to the usual size.

B. Manual Extraction Task

A Reference Term List is manually extracted to serve as the gold standard for the evaluation of monolingual TCs automatically extracted from German texts in the newspaper domain on the topic of Brexit. The extracted list consists of SWTs and MWTs. A terminologist is purposefully omitted from the task since the analysis focuses on the translation purposes. In that scenario, having domain experts compile the list is a more realistic scenario. Two experts are asked to extract all linguistic terms regardless of their structure, similarly to [8].

The experiments are conducted separately with the union and with the intersection of the obtained lists. Almost 79% of terms in the union of the lists are unigrams or SWTs (535 out of 681), and the remaining 21% are MWTs (bigrams 11%, trigrams 6%, fourgrams 3%, fivegrams and sixgrams both less than 1%; in counts 77, 43, 21, 2, and 2, respectively). The terms extracted manually from the above-described corpus cover nouns, verbs, and adjectives, and belong to different registers in the German language. Unlike the authors in [8], who do manual extraction on a subset of their corpus, in this paper a small-sized corpus is chosen in order to process the whole corpus manually, as this is considered beneficial according to [16]. A random sample of terms is shown in Table I. The manually extracted lists contain quite a big percentage (almost 77%) of terms which consist of nouns or nouns and adjectives (Table II). The top 10 patterns account for 92% of manually extracted terms.

C. Automatic Extraction Task

Three tools are used and evaluated in the task of ATE – Rainbow, which is part of the Okapi framework; Termsuite which is a tool developed within TTC project; and Sketch Engine, which is not a terminology extraction tool per se, but a leading web service for corpus analysis. The reference corpus used in the evaluation is the German web 2013 (deTenTen) corpus from the TenTen family [30], which consists of 16.5 billion words. In the first part of the experiment, different cut-off values are applied to ATE results, similarly to [18]. The selected cut-off values are 50, 100, 200, 500, and 1000. In the second part of the experiment, no cut-off value is applied. The differentiation is made between the lists obtained without a minimum frequency threshold and those obtained with the minimum frequency threshold set to three.

Based on the Okapi Framework, Rainbow is an open-source platform-independent term extraction tool written in Java, which implements purely statistical methods, and can thus be applied to any language. Since a token grouping method is applied for the extraction, terms are not reduced to their stems. The only linguistic knowledge provided is a list of stop words. This means that a sequence of words is discontinued if a stop word is found in-between. Due to the fact that almost no linguistic knowledge is utilized, one experiment is conducted in which the corpus is lemmatized prior to the extraction to explore the effects that lemmatization has on the process.

Termsuite is an open-source tool developed within TTC project. Term extraction in Termsuite is a two-step procedure. A pattern-based candidate identification from the first step is followed by ordering by decreasing domain-specificity. Term candidates are, beside SWTs, restricted to bigrams and trigrams in accordance with previously identified POS patterns [29]. Since Termsuite enables also morphological compound detection and term variant processing, it is expected to be best suited for German.

TABLE I. A RANDOM SAMPLE OF MANUALLY EXTRACTED TERMS

German Terms
Austrittsdatum
Brexit-Gruppe des Europaparlaments
Brexit-Schock
Chaos-Brexit
EU-Botschafter
Gestaltung der Grenze
No-Deal-Szenario
Steuersenkung
Vereinigte Staaten
Zwangspause des Parlaments

TABLE II. TOP 10 POS PATTERNS IN MANUALLY EXTRACTED TERMS

Pattern	**Frequency**
N	484
ADJ N	40
V	28
ADJ	22
NN	15
N ART N	10
N PREP N	7
PREP N V	6
ART N	5
N V	4
PREP N	4

Sketch Engine is named after one of its key features–word sketches. It employs a contrastive two-step approach to terminology extraction–first the grammatical validity of a phrase (unithood) is assessed using the term grammar, next the normalized frequencies of TCs from the focus corpus are contrasted (termhood) with those in the reference corpus [31] by using the 'Simple Math' (with an add-N parameter of one) statistics [32]. Sketch Engine implements ATE as two separate processes, dependent on the user needs–keywords extraction and multiwords extraction. Keyword designates a word typical of a corpus in comparison to a general corpus and it is determined by the keyness score given in eq. 4, where $f_{pm_{focus}}$ stands for the normalized frequency (per one million words) of the word in the focus corpus, and $f_{pm_{ref}}$ for the normalized frequency (per one million words) of the word in the reference corpus. The default value of n is set to one. Terms are multiword expressions and the same score is used for their extraction, except that the absolute frequency counts are used. While keywords can be extracted from any corpus, a prerequisite for extracting terms is the existence of a term grammar since term extraction requires tagged and lemmatized corpora.

$$\frac{f_{pm_{focus}} + n}{f_{pm_{ref}} + n} \quad (4)$$

IV. Results and Discussion

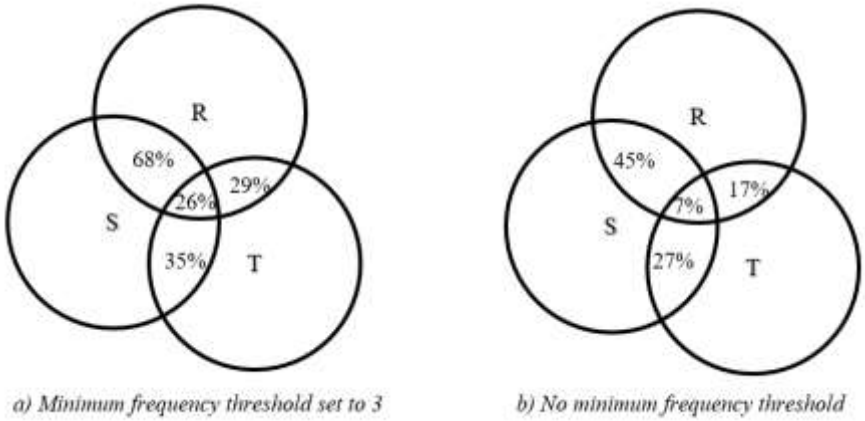
In this paper intrinsic evaluation is applied [14] and precision, recall, and average precision quality indicators are used. The comparison with the gold standard is implemented as a strict string matching of the list entries as well as a lemmatized string matching of entries. Only the latter scores are reported as these prove to be slightly superior. The counts of terms manually extracted and extracted by each tool are given in the Table III.

The overlaps between ATE results are given in Fig. 2. Please note that R stands for Rainbow, S for Sketch Engine and T for Termsuite, and these labels will be used in figures henceforth. Furthermore, R(V1) is used when referring to the Rainbow results on the non-lemmatized corpus, and R(V2) when referring to the Rainbow results on the lemmatized corpus. Since R(V2) version results in a degradation, it is discarded from all further experiments. Although Sketch Engine and Rainbow give more similar lists, their scores differ greatly, as evident from Fig. 3 and Fig. 4, which give overall evaluation results. The overlaps between ATE results which are found in the gold union are shown in Fig. 5.

Termsuite results with the minimum frequency threshold set to 3 are superior up to the cut-off value of 200, after which it suddenly deteriorates and Sketch Engine gets better under the same settings. The results on the gold intersection follow the same pattern although the precisions are lower from the very beginning. Rainbow manages to recover a great number of terms in the gold intersection even with the minimum frequency threshold set to 3.

TABLE III. Counts of lemmatized terms in the gold standard and terms extracted automatically by each tool

	Number of terms					
	SWTs		***MWTs***		***Total min=3***	***Total N/A***
	min=3	***N/A***	***min=3***	***N/A***		
MTE union	536		145		N/A	681
MTE intersect	229		82		N/A	311
Rainbow	933	3655	343	6147	1276 (V1) 1217 (V2)	9802 (V1) 8731 (V2)
Sketch Engine	883	3222	14	317	897	3539
TermSuite	622	1384	134	359	756	1743



a) Minimum frequency threshold set to 3 b) No minimum frequency threshold

Fig. 2. Overlap between ATE Results.

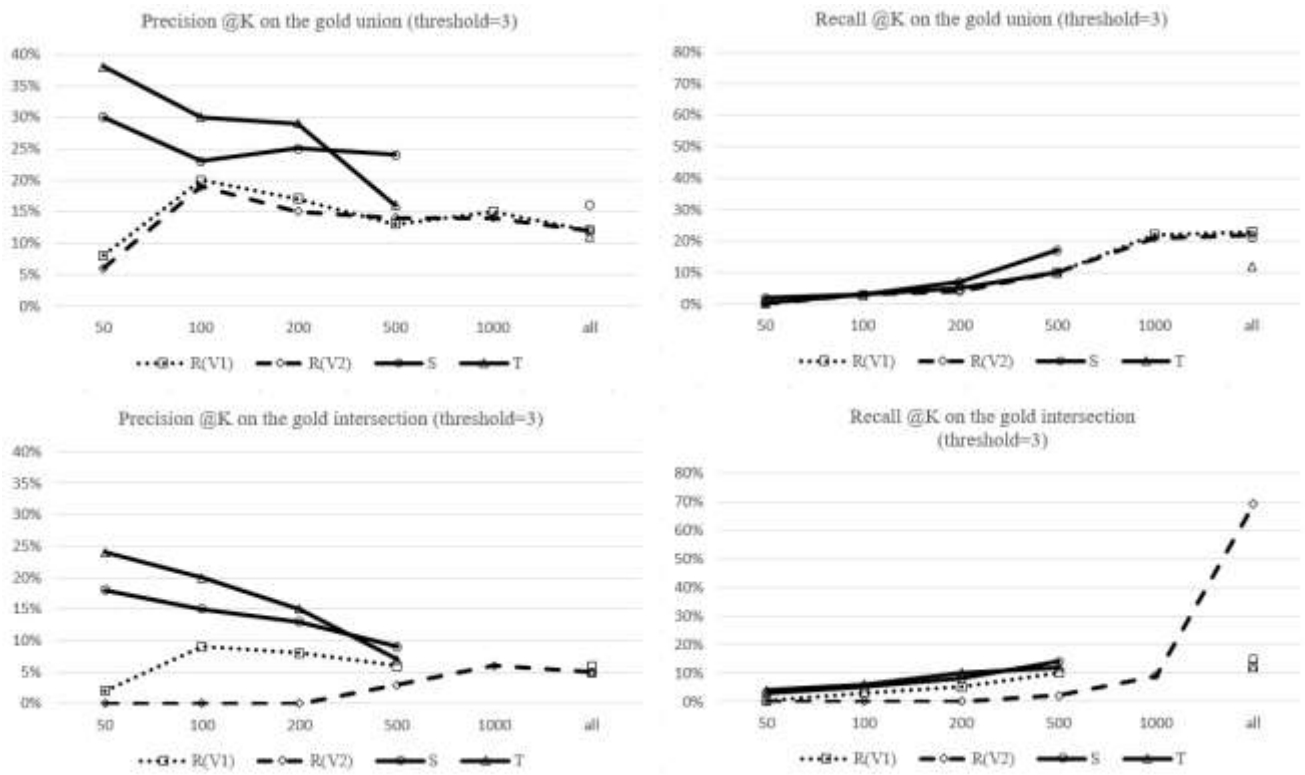


Fig. 3. Overall Lemmatized Evaluation Results on the Gold Union and Intersection.

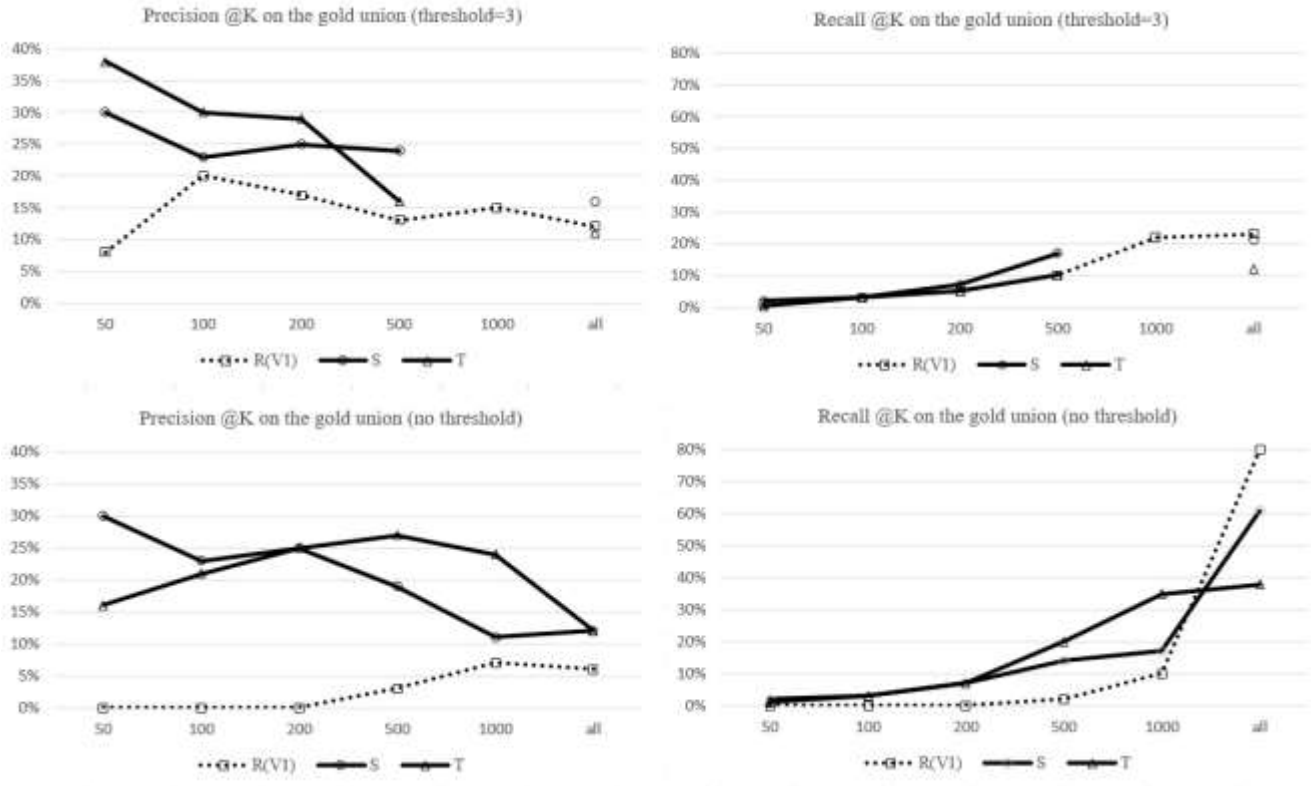


Fig. 4. Overall Lemmatized Evaluation Results on the Gold Union with and without Minimum Frequency Threshold.

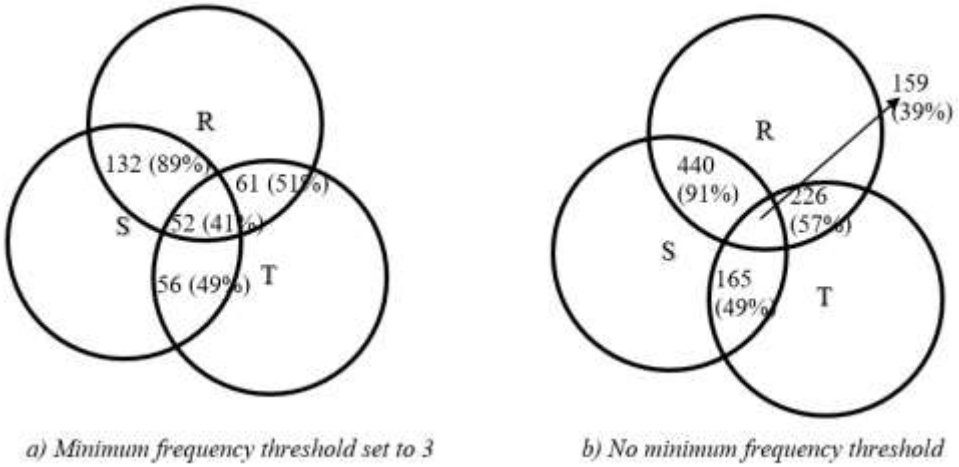


Fig. 5. Overlap between ATE Results per Tools and the Gold Standard.

Without the minimum frequency threshold, the situation with Sketch Engine and Termsuite gets somewhat reversed, i.e. the result is a draw at 200, while Termsuite beats Sketch Engine when the cut-off value is set to 500 or higher. While Sketch Engine performs similarly as in the scenario with the minimum frequency threshold, it can be concluded that Termsuite is affected by the threshold settings. Finally, when there is no cut-off value, the two tools perform the same. The results regarding recall are more stable when the minimum frequency threshold is set. When the minimum frequency threshold is removed, all the tools gain better recall scores at the expense of precision, except for Termsuite which has a 1% increase in precision alongside 26% increase in recall.

Since this paper presents an analysis of terminology extraction tools from a translator's perspective, and since Rainbow is, expectedly, the worst scoring tool regarding precision, average precisions are calculated only for the remaining two tools (Table IV).

Although the average precision at rank 1000 which takes recall into account is better for Termsuite (9%) than for Sketch Engine (4%), they both do the same, i.e. 9% according to the overall results.

Regarding MWUs, it is worth noting that in the scenario with the minimum frequency threshold applied, the number of recovered MWUs is 12, 7, and 5 for Rainbow, Sketch Engine, and Termsuite, respectively.

TABLE IV. AVERAGE PRECISIONS FOR SKETCH ENGINE AND TERMSUITE

	avgP@500	***avgP(+recall)@500***	***Total avgP(+recall)***
Sketch Engine	25%	4%	5%
Termsuite	28%	3%	3%
Sketch Engine (no threshold)	27%	3%	9%
Termsuite (no threshold)	28%	5%	9%

The average precisions at rank 500 are 25% and 28% for Sketch Engine and Termsuite, respectively. By looking at average precisions which take recall into account, it is evident that only 5% overall average precision is obtained for Sketch Engine versus 3% for Termsuite. In the scenario without the minimum frequency threshold, average precisions at rank 500 are 27% and 28% for Sketch Engine and Termuite, respectively.

Another thing worth noting is that less than 2% of the terms extracted by Sketch Engine in the scenario with the threshold are MWUs, while that percentage goes up to 10% in the scenario with no threshold. Termsuite, on the other hand extracts pretty similar percentage of MWUs in both scenarios, i.e. 21% and 26%, respectively. Rainbow doubles the percentage from 37% in the first scenario to 63% in the second. On the other hand, even 50% of the MWU candidates extracted by Sketch Engine in the scenario with the threshold are correct, while those percentages are as low as 3.5 and 1.5% for Rainbow and Termsuite, respectively. The results on Rainbow assert the fact that due to data sparseness in small-sized specialized corpora statistical measures that use the candidate's frequency in a domain-specific corpus perform much better on SWTs than on MWTs [25]. A general conclusion can be made that linguistic approaches seem to be more suitable for translators as translators do not want to go through huge lists of term candidates and find only a handful of real terms. Although the performance of Sketch Engine and Termsuite is competitive, results speak slightly in favor of Termsuite with German as the language and Brexit as the domain. Regarding the overlap in the correct terms between Sketch Engine and Termsuite, there seems to be potential in combining their outputs.

V. CONCLUSION AND FUTURE WORK

The research described in this paper is conducted on the German corpora. The study does not propose a novel approach to automatic terminology extraction, but compares performance of three well-known extraction methods under the same experimental settings.

Due to the differences in corpus selection, evaluation methodology, and scope of TCs included in the evaluation, comparisons of various research results are often intractable. One of the goals set for this research was to create a gold standard and thus facilitate performance comparisons of various terminology extraction tools. One of the strengths of the study presented in this paper is the fact that the gold standard is compiled by two domain specialists as this is a more realistic setting than having a terminologist at hand when doing terminology extraction for translation purposes. Both the union and intersection of the two lists compiled by domain specialists are used in the experiments. Although having two evaluators approve the term in order to include it in the gold standard somewhat improves recall for lower levels, it reduces the size of the gold standard, and consequently precision for at least 6%. The analysis conducted on the gold union reveals that altogether only three POS tags for unigrams and two POS patterns for bigrams account for over 86% of the terms. These tags and patterns are included in the list of eight most important patterns provided by [15]. Precision could thus be potentially improved by restricting the rules. A cut-off value of 500 per category is opted for since ATE systems produce significant amounts of noise and users in the role of translators are mostly unwilling to scan through TC lists. Another strength of this research is the fact that the gold standard is based on the same corpus of the exact same size which is used for testing purposes.

To conclude, the results confirm that fully automatic terminology extraction is still out of reach for computers. The choice of method should therefore depend on whether the application puts more importance on the precision or recall. In general, if the extracted lists are to be checked manually, precision should be considered more important to avoid the task being too tedious. Since terms are inherently semantically defined, the final confirmation of an expression's term status still has to be done manually by domain specialists.

There are several directions which open up for future work. In order to get more meaningful results, the size of the focus corpus should be increased. This has not been done for the purpose of this research in order to have both manual and automatic extraction conducted on the same corpus, which would not be feasible with a corpus greater in size. Human evaluators could be asked to judge the TCs which are, perhaps mistakenly, not included in the gold standard. Furthermore, besides testing other tools, the combination of different ATE results or a voting mechanism could be employed. From the results presented in this paper, there exists some potential in combining Sketch Engine and Termsuite outputs. In the future, the work will be extended with domain-specific terminology in fast developing domains, e.g. the Information and Communication Technology (ICT) domain and different issues will be highlighted which occur when the corpus is compiled by different authors in languages which do not have a harmonized terminology and thus exhibit inconsistencies at the lexical level.

REFERENCES

[1] Z. Zhang, J. Iria, C. Brewster, and F. Ciravegna, "A comparative evaluation of term recognition algorithms," Proc. 6th Int. Conf. Lang. Resour. Eval. Lr. 2008, pp. 2108–2113, 2008.

[2] Z. Zhang, J. Gao, and F. Ciravegna, "JATE 2.0: Java Automatic Term Extraction with Apache Solr," Proc. 10th Int. Conf. Lang. Resour. Eval. Lr. 2016, pp. 2262–2269, 2016.

[3] N. A. Astrakhantsev, "ATR4S: toolkit with state-of-the-art automatic terms recognition methods in Scala," Lang. Resour. Eval., vol. 52, no. 3, pp. 853–872, 2018.

[4] A. Repar, V. Podpečan, A. Vavpetič, N. Lavrač, and S. Pollak, "TermEnsembler : An ensemble learning approach to bilingual term extraction and alignment," Int. J. Theor. Appl. Issues Spec. Commun., no. January, 2019.

[5] K. Heylen and D. De Hertog, "Automatic Term Extraction," in Handbook of Terminology, 2015, pp. 276–287.

[6] G. Thurmair, "Making Term Extraction Tools Usable," in Proceedings of EAMT-CLAW, 2003.

[7] N. Astrakhantsev, D. G. Fedorenko, and D. Y. Turdakov, "Methods for automatic term recognition in domain-specific text collections: A survey," Program. Comput. Softw., vol. 41, no. 6, pp. 336–349, 2015.

[8] C. Lang, R. Schneider, and K. Suchowolec, "Extracting Specialized Terminology from Linguistic Corpora," Gramm. Corpora, pp. 425–434, 2018.

[9] M. L. Homme and L. Benali, "Definition of an evaluation grid for term-extraction software," Terminology, vol. 3, no. 2, pp. 291–312, 1996.

[10] G. Doğru, "Automatic Term Extraction from Turkish to English Medical Corpus," in Computational and Corpus-based Phraseology, 2019, pp. 157–166.

[11] R. Nazar, "Distributional analysis applied to terminology extraction," Terminol. Int. J. Theor. Appl. Issues Spec. Commun., vol. 22, no. 2, pp. 141–170, 2016.

[12] K. Kageura and B. Umino, "Methods of Automatic Term Recognition: A Review," Terminol. Int. J. Theor. Appl. Issues Spec. Commun., vol. 3, no. 2, pp. 259–289, 1996.

[13] G. Doğru, A. Martín, and A. Aguilar-amat, "Parallel Corpora Preparation for Machine Translation of Low-Resource Languages : Turkish to English Cardiology Corpora," in Proceedings of the LREC 2018 Workshop'MultilingualBIO: Multilingual Biomedical Text Processing', 2018, pp. 12–15.

[14] J. Vivaldi, "Evaluation of terms and term extraction systems : A practical approach," Terminology, vol. 13, no. 2, pp. 225–248, 2007.

[15] A. Rigouts Terryn, V. Hoste, and E. Lefever, "A Gold Standard for Multilingual Automatic Term Extraction from Comparable Corpora: Term Structure and Translation Equivalents," in 11th International Conference on Language Resources and Evaluation (LREC 2018), 2018.

[16] A. Gojun, U. Heid, B. Weissbach, C. Loth, and I. Mingers, "Adapting and evaluating a generic term extraction tool," Proc. 8th Int. Conf. Lang. Resour. Eval. Lr. 2012, pp. 651–656, 2012.

[17] A. Šajatović, M. Buljan, J. Šnajder, and B. Dalbelo Bašić, "Evaluating Automatic Term Extraction Methods on Individual Documents," in Proceedings of the Joint Workshop on Multiword Expressions and WordNet (MWE-WN 2019), 2019, vol. 0, pp. 149–154.

[18] D. Inkpen, T. Sima Paribakht, F. Farahnaz, and A. Ehsan, "Term Evaluator: A Tool for Terminology Annotation and Evaluation," Int. J. Comput. Linguist. Appl., vol. 7, no. 2, pp. 145–165, 2016.

[19] P. Schone and D. Jurafsky, "Is knowledge-free induction of multiword unit dictionary headwords a solved problem?," Proc. 2001 Conf. Empir. Methods Nat. Lang. Process., pp. 100–108, 2001.

[20] M. Brkic Bakaric and I. Lalli Pacelat, "Parallel Corpus of Croatian-Italian Administrative Texts," in 2nd Workshop on Human-Informed Translation and Interpreting Technology (HiT-IT 2019), 2019, pp. 11–18.

[21] H. Costa, A. Zaretskaya, G. Corpas, and M. Seghiri, "Nine terminology extraction tools Are they useful for translators ?," Multiling. #159, vol. 27, no. 3, pp. 14–20, 2016.

[22] D. Bourigault, M. De Recherche, A. Machado, and C. Jacquemin, "Term Extraction + Term Clustering : An Integrated Platform for Computer-Aided Terminology Components of the Platform for Computer-Aided Terminology," in Ninth Conference of the European Chapter of the Association for Computational Linguistics, 1999, pp. 15–22.

[23] H. Nakagawa, "A Simple but Powerful Automatic Term Extraction Method," in COLING-02 on COMPUTERM 2002: second international workshop on computational terminology-Volume 14, 2002, pp. 1–7.

[24] M. Hong, S. Fissaha, and J. Haller, "Hybrid filtering for extraction of term candidates from German technical texts," Conférence TIA-2001, Nancy, 3 4 mai 2001 Hybrid, p. 10, 2001.

[25] J. Schäfer, I. Rösiger, U. Heid, and M. Dorna, "Evaluating noise reduction strategies for terminology extraction," CEUR Workshop Proc., vol. 1495, pp. 123–131, 2015.

[26] G. Lalić-Krstin and N. Silaški, "From Brexit to Bregret," English Today, vol. 34, no. 2, pp. 3–8, 2018.

[27] K. Ivanova, U. Heid, S. S. I. Walde, A. Kilgarriff, and J. Pomikálek, "Evaluating a German sketch grammar: A case study on noun phrase case," Proc. 6th Int. Conf. Lang. Resour. Eval. Lr. 2008, pp. 2101–2107, 2008.

[28] M. M. A. Gizi, "Word Formation in German Linguistics: Theoretical and Methodological Analysis," Open J. Mod. Linguist., vol. 08, no. 05, pp. 143–150, 2018.

[29] U. Heid and A. Gojun, "Term candidate extraction for terminography and CAT : an overview of TTC," Proc. 15th EURALEX Int. Congr., pp. 585–594, 2012.

[30] M. Jakubíček, A. Kilgarriff, V. Kovář, P. Rychlý, and V. Suchomel, "The TenTen Corpus Family," 7th Int. Corpus Linguist. Conf., pp. 125–127, 2013.

[31] M. Jakubíček, A. Kilgarriff, V. Kovář, P. Rychlý, and V. Suchomel, "Finding Terms in Corpora for Many Languages with the Sketch Engine," pp. 53–56, 2015.

[32] A. Kilgarriff, "Simple maths for keywords," in Proceedings of the Corpus Linguistics Conference, 2009.