Making Term Extraction Tools Usable

Gregor Thurmair
Comprendium GmbH
Balanstr. 57, 81541 Munich
Germany
gregor.thurmair@comprendium.de

Abstract

The paper reviews the extraction of terminology from corpora. It identifies three possible applications, terminology, translation, and retrieval. They differ in the requirements for relevancy of terms to be extracted. Standard evaluation methods based on recall and precision critically depend on the notion of relevancy of a term, which is questionable and possibly outweighed in favour of criteria of usability in practical applications. Two tools, TermExtract and BiExtract, are presented as examples of the integration of extraction tools into different workflows.

1 Definitions

Terms and general vocabulary: Term Extraction means to identify term candidates from a text corpus. Terms are a "*designation of a defined concept in a special language by a linguistic expression. A term may consist of one or more words*." [ISO 1087, 1st edition, 1990]. This definition refers to a difference between general and special language, assigning terms to the special language area. However, this is not always what a term extraction tool should be restricted to, as will be shown.

Term extraction and term recognition: There are two phases to be distinguished, both sometimes called term extraction: The first one is the term extraction proper, also called term acquisition [VBM02], i.e. the identification of term candidates in a corpus; the second one is term recognition [VBM02], i.e. comparison of the extraction result with some dictionary / term bank resource in order to identify known / unknown terms. There are applications of term extraction which do not need a term recognition.

Single and multiword terms: There is a difference between single word terms and multiword terms; a multiword term consists of several words but forms a semantic unit (as opposed to collocations, the meaning of which still can be determined compositionally). Most of the terms which a term extraction component is supposed to find, like *nuclear power plant*, are multiword terms; single word term extractors are seldom useful.

Monolingual and bilingual extraction: Some extraction tools work monolingually, i.e. identify term candidates in just one language. Others operate bilingually, by trying to find translation equivalents, usually in aligned corpora. Bilingual term extraction either works on source and target side simultaneously (as in Trados' ExtraTerm), or it starts from a monolingual glossary and find target equivalents for it (as in Comprendium's BiExtract).

2 Applications

In evaluating term extraction tools, the evaluation criteria should depend on the purpose of term extraction, as specified in the ISO norm 9126 for suitability: "*provide an appropriate set of functions for specified tasks and user objectives*" [cit. SAU02]. Therefore, a closer look at the intended applications for which term extraction is necessary before an evaluation can be planned. It quickly becomes clear that there are several different purposes for which such tools can be used; and they impose different requirements to the tools.

2.1 Terminology

The “classical” application of term extraction is to identify term candidates, in the sense of special language concept designators (e.g. [SQU02]). Terminology has always had a strong normative component, but is sometimes based on empirical considerations e.g. to investigate which designators are really used in the relevant special purpose language. Term extraction systems in this area focus on the difference between general language and special language expressions, as only the latter are relevant for the intended purpose.

A special case of this application are tools to verify controlled language. Such tools have close links to term recognition as their task is to identify words which neither belong to the general language nor to the special domain language defined for a special branch (like AECMA) or even enterprise. In order to compare such controlled terms they must first be extracted [THU00].

The main characteristics of this application in relation to term extraction is to

- identify a special subset of the vocabulary of a text, and
- restrict it to a language of a special domain.

Most of the term extraction task here is monolingual, and of course it must cover single and multiword terms; bilingual term extraction is also used. Term recognition is essential esp. for controlled language applications.

2.2 Translation

When term extraction is used for translation, be it human or machine translation, the purpose is to identify unknown words, be it terms or general vocabulary words. This is the focus of current commercial tools like Trados’ Extraterm, Multitrans, Xerox TermFinder, etc. (for Slowene cf. [VIN00]). However, the relevancy of candidates is a highly idiosyncratic issue. In reality, glossaries produced by translators depend on what they already know, what they consider their colleagues should know etc.; they do not distinguish between terms and general vocabulary, and they are seldom really consistent in covering a given domain (this can often be seen when attempting to build an ontology (or thesaurus) from a given glossary). As a result, it is hard to define a priori which results a term extraction should deliver, and the practical requirement is usually to find everything the system does not know yet.

Term recognition is therefore an essential component in such applications, and the purpose of the whole enterprise is to find unknown words; clearly so in machine translation but also in human translation. [LIE02].

So the main characteristics of term extraction in the translation area are to

- identify a special subset of the vocabulary of a text, but
- this subset is defined not absolutely (as defined by a language for special domain) but relatively, as a comparison with what is already there or known.

This fact has a significant influence on how relevancy is handled in evaluation. A term recognition phase is always part of this application.

Three issues are worth mentioning here:

- There is the problem in identifying **unknown multiwords** consisting of known parts, especially in machine translation. Even if *power* and *plant* are both in the dictionary the term *power plant* may not be; and not finding this term leads to mistakes in translation. Therefore simple term recognition (i.e. dictionary lookup) is not sufficient, and some term extraction needs to be performed as well.
- There is a special interest in multilingual or **bilingual extraction**, often based on the use of translation memories with aligned sentences. This helps translators to find translations which are already in use; special tools can be developed to define if those translations are wanted (canonical) or unwanted in a controlled environment (multilingual term verification, cf. [THU00]). Again only tools which do both single and multiword term extraction and alignment are really useful.
- A special field of research is the investigation of **collocations** ([HEI99], [GOW01]). Collocations are not terms (which are semantic units) but are situated on a pragmatic level: They define preferred ways of expressing things. As a result, they contain many more verbal parts than terms, which are mainly nominal constructions; and a result of this is that the linguistic variance of (verbal) collocations is much higher than that of terms, and more

elaborate linguistic analysis procedures must be used.

2.3 Information Retrieval

The purpose of term extraction in information retrieval is different again. The purpose here is to get an overview of the searchable vocabulary of an application (assuming it is a special purpose retrieval engine, not a global internet search tool), which is identical to determining the searchable topics. Term extraction in this context is a first step towards the definition of linguistic resources for query expansion, query translation, ontology building, etc. There is a growing interest in this area, cf. [VBM02], [XKP02], [LEM02], [FFR02].

The basic requirement for term extraction here is that users should find all searchable topics, whether these are terms or general vocabulary words. Completeness is rather important in this application, and term recognition is not really relevant as the resources to compare the terminology against are often incomplete [VBM02]. Whatever is in the text data must become an object of searching, and must therefore be represented in the resource for query expansion / translation: What is not there will not be translated, and will also not be found.

So the characteristics of this application for term extraction are

- to identify all searchable vocabulary, whether general or special, and whether known or not.
- in addition, to identify relations between the concepts found for use in ontologies and search support systems.

The application primarily requires monolingual extraction, and adds multilingual aspects in the case of cross-lingual retrieval. Bilingual term extraction in this application does not have to locate “correct” or canonical translations; the emphasis must be on translations which are actually used translations, as only these will ensure successful searching.

It is thus clear that term extraction is used for quite different purposes, and only the first step (identification of term candidates, or term acquisition) is common to all of them; what follows (term recognition, comparison to some known resources, use of extracted terms etc.), is different from application to application. This must be taken into account in evaluation.

3 Technology

A look at the technology used for term extraction is relevant because the selection of the technology has a direct impact on the extraction quality, and depends on the purpose of the extraction. The mainstream technology in term extraction follows a pattern of combining statistics with linguistic processing (details cf. [DAI96] [CEV01]). It identifies possible candidates, and determines their relevance.

3.1 Identification of possible candidates

Each term extraction system should have a means of identifying **single word** term candidates. Whichever application is intended, **base form reduction** should be a basic step in analysis as inflections are just contextual variants from a conceptual point of view.

Using **stop lists** may also be advisable, the content of the stop lists may differ, however, depending on which application is intended. For terminology, all general vocabulary words could be stopped; for cross-lingual retrieval, all non-search terms would be blocked.

Term extraction systems must also provide **multiword** term identification in all applications. Different techniques are used for this purpose:

- linguistic filters, based on **category patterns** with shallow (NP-oriented) syntactic analysis [ARP95][BOU95] [FAM00], are suitable for term identification, as terms have typical linguistic structures. Even if they are non-adjacent [DEA00] , they can be described linguistically.

  Our own analysis based on the content of the Siemens TEAM term base with about 1.5 mio entries (English and German) showed that the most frequent about 50 patterns cover about 80% of the terms; these patterns are nearly exclusively standard NP structures, the only “non-trivial” pattern being conjunctions (like *checks and balances*). (Similar results on the efficiency of such patterns are reported by [ARP95]). This observation justifies the selection of a rather direct analysis approach for terms.
- more complex filters are needed to describe collocations, including verbal elements. Reliable term extraction will have to use **full parsing**, given the amount of variants in which

such collocations can occur [HEI99] [GOW01].

- **statistical devices** to coordinate words into multiwords [PAL 01] usually face the problem that they generate significantly more noise than the approaches just mentioned as there are many strings in a corpus which occur with some frequency (like *this can be* or *in general, however*). Some filters are used to reduce the number of term candidates. The simplest filter is to limit the length of the pattern to two [PAL01] or three content words (plus prepositions, determiners etc.). However, the only cases where statistical analysis seems to be superior to linguistic filtering [DEA00] appear to go together with a suboptimal selection of linguistic term patterns.

Possible term candidates are best if they are (linguistically) meaningful concepts. All other proposals just add noise for the evaluators.

3.2 Weighting and Relevancy Determination

There are several proposals on how the relevant terms can be identified in a set of candidates. Most of them are based on frequency considerations, and use different measures (cf. [EAG98]).

However, considering the applications discussed above, it is difficult to specify what the target of a relevancy determination should be:

- Only in the case of terminology is there an objective criterion, namely to identify members of a special domain language.
- In the case of translation, relevancy depends on the existing material (all non-known words), and
- In the case of retrieval, relevancy is defined by the searchable concepts of a corpus.

In the first case, a comparison of relative frequencies of term candidates in a special domain corpus and a general vocabulary corpus seems to be adequate [SQU02], but the problem that many terms (like *belt*, *fault* etc.) are homonymous between a term reading and a general reading must also be solved. So this approach works only if combined with some word sense disambiguation procedure. In itself it does not seem to improve the recognition quality [HEV99].

In the other applications (i.e. translation and retrieval) it is difficult to optimise the term candidate list more than by providing simple frequency information; it is difficult to know the direction into which to optimise.

3.3 Bilingual Extraction

If existing bilingual dictionaries (both human and machine readable) are consulted then all combinations of source language single word / multiword terms with target language single word / multiword terms are represented, cf. fig. 1. Therefore all these combination possibilities must be supported in term extraction; otherwise the approach lacks descriptive adequacy.

German term	English translation	pattern
Hund	dog	sw -> sw
abblenden	[car] dim the headlights	sw -> mw
Kernkraftwerk	nuclear power plant	sw -> mw
springender Punkt	crux	mw -> sw
beschleunigte Hinterbliebenenrente	accelerated death benefits	mw -> mw

Fig. 1: Translation examples

Again, a proper linguistic characterisation of the term candidates on both sides is a prerequisite for a successful correlation of source and target term candidates, so all techniques described above for base form production and multiword term filtering must be applied here as well, otherwise a significant amount of noise must be expected.

In bilingual extraction, the goal of the extraction can be expressed on an intuitive basis: To find a translation equivalent for a given term. Whether a target phrase is the translation of a source term or not can be assessed (e.g. by consulting dictionary material), and be used as an evaluation criterion.

Approaches differ in whether they carry out source and target extraction simultaneously (Trados, Multitrans), or in whether they use a source language glossary and then try to identify matching target expressions (BiExtract, cf. below). Most approaches use aligned text (i.e. translation memories) as a linguistic basis although this resource is not always available; e.g. in cross-lingual retrieval the standard case is to have unaligned texts, but on the same topics.

3.4 Additional Issues

There are several issues which influence the analysis results. Some of them have to do with decisions about what should be considered to be a term candidate in special cases:

- It is unclear whether or not some **token classes** should show up in a term candidate list: numbers, filenames, system commands like *fgrep* or *rm* and others, abbreviations, URLs and other tokens.
- **Proper names** are another issue, particularly if news texts are used for term extraction. Person names like *Bill Monroe* would not be considered as terms but are of interest in a search environment and should be kept there. For some others it is more difficult to decide, like *Lufthansa Frequent Flyer program* where the name is part of a term. Determining proper names adds an additional level of complexity to a term extraction program; examples are given e.g. in [MIK94].
- There is also the problem of **embedded terms**. These are terms which occur as parts of larger terms: if *power plant* always occurs in the phrase *nuclear power plant*, should it then be proposed as a term candidate in its own right or not? This issue is discussed in both [FAM00] and [DAI96]; results are mixed.

In these cases, as the decision seems to be dependent on the purpose and the domain of the extraction, the best solution is to offer a possibility of selection to the users, and let them decide on a case-by-case basis.

Finally, there is the problem of **base form creation**. If a term always occurs in plural form only (like *checks and balances*) there is no point in creating a singular base form. Sometimes term parts must not be reduced to their base form (esp. participle forms like *secured connection* do not have *secure connection* as base form; but also adjective forms in German where the base form of *logischer Verbindung* is another adjective inflection (*logische Verbindung*) and not *logisch Verbindung*). It is not always straightforward to determine the correct base form of a term, and term interchange formats like OLIF [OLI02] provide special guidelines on this topic.

4 Evaluation

4.1 Evaluation Criteria Reviewed

Fundamentally, evaluation is always a comparison. The EAGLES 7-step recipe [EAG99] defines the dimensions of possible comparisons; based on the intended purpose of the tool / component, a requirements model must be designed which can evaluate the different tools, utilising defined metrics.

In term extraction, apart from compliance to documentation, existence of user support, ease of installation etc. [SAU02], supported languages, presentation of contextual information (cf. [LIE02] for a very detailed list), the central criterion is the quality of the extraction (cf. HEV99]).

4.2 Recall and Precision

The basic quality measure in term extraction is recall and precision (cf. [HEV99], [SAU02], and most others). Both are defined in terms of relevancy, i.e. which relevant terms are found in the total document set and in the retrieved term set.

The fundamental problem with this approach is that it is very difficult to define what relevant terms are. As pointed out above, the concept of relevancy differs strongly depending on the intended application. And while there is an intersubjective intuition of what a possible translation equivalent of a given term in the target language is, there is no such intuition as to which terms should considered to be relevant.

Some examples may be given from an automotive text that contained about 10300 tokens. Term extraction produced about 260 multiword term candidates, the most frequent of which are shown in fig. 2 in the Appendix.

If people were asked which entries to consider relevant, they would possibly all agree that *first time* should be eliminated while *peak torque* should be kept as a term; but what about *sports seat*? *petrol engine*? *ground clearance*? *soft top*?

Apparently, relevancy consideration, which is fundament to all recall / precision evaluations, is based on different criteria:

- in terminology, relevancy of a term is determined by its position in an LSP domain;
- in translation, relevancy of a term is determined by its status vis-à-vis an existing term bank or dictionary, the entries of with have an unclear status and coverage themselves;
- in retrieval, relevancy of a term is determined by the possibility to search for it, which holds for basically all terms of the corpus.

While *sports seats* would possibly not go into a translator's term bank, it would surely go into an

ontology for a search system, because users would want to be able to search for *cars* with *sports seats*.
Most articles on term extraction which refer to recall / precision evaluation do not consider this basic problem, and mainly evaluate term relevancy against a set of given terms, without questioning its theoretical status. And the results are therefore not very helpful:

- Should UNIX commands be terms? If so, the system finds approval in [HEV99]. Should proper names be? If so, it is not liked by [HEV99] and [SAU02] but liked by [XKP02] as an instance of an ontology node.
- What if a system does not find a term which is in the glossary but not in the corpus (in the AVENTINUS project, only 16% of the terms of the law enforcement glossary were represented in the corpus; similar results for medical domain [cf. VBM02]): Does this affect the f-measure, although it is a weakness of the glossary to contain useless terms? What if the system gives a term in base form but the reference term bank has it as inflected (plural) form? etc.

For the single term extraction tools, this has the effect that the best thing they can do is to provide meaningful linguistic concepts which could be semantic units; this is the least common denominator of term extraction. What to do with them, tools cannot decide. The same terms are used in all applications, what differs is just their status.

A possible quality measure could then be the ratio between really possible concepts (semantic units), like *power plant*, and candidates which will never be used in any application, and are only noise, like *this is* or *first time*. Such a quality measure addresses the issue of usability, which will finally decide on the quality of term extraction tools (some points here are made in [LIE02]).

4.3 Usability Revisited

From a usability point of view, there are three aspects of term extraction which are really important:

1. Does the tool provide all candidates?
This is a kind of recall orientation, because finding missing terms costs more time than eliminating noise. Users have the alternative either to scan through the whole corpus and mark all term candidates by hand, or to run a tool which does this for them. But the tool is only useful if they do not have to manually go through the corpus again. So the more term candidates that are missing the less useful a term extraction tool tends to be.

2. How fast can people be in defining the right term (sub)set?
Given the fact that determining the relevancy is a difficult job, and requires user intervention in any case, the question of usability is how easy users can determine what they consider relevant. Most current term extractors produce significant amounts of noise, and the more time users spend in scanning through term candidate lists the less useful the tool is.

Usability here depends on two factors, namely noise production and speed of editing.

- How much noise does the tool produce? I.e. how many irrelevant candidates need to be eliminated? This again depends on two other factors, namely how many candidates are produced, and how many candidates can be filtered by some term recognition component which knows what is already there (or simply stops noisy words). So noise reduction is an important factor in usability comparisons.
- How fast can a correct term list be produced? This involves editing issues (e.g. if a term candidate is already in base form it needs not be edited) and also easy deletion of noise candidates. A term extraction tool should provide good support for doing this.

3. How can people make use of the term extraction output?
This aspect is important because term extraction is only the first step in a larger processing chain, which usually consists in cleaning up the extraction results, and then importing the “good” terms into some tools, be it a term bank or a machine translation dictionary. Therefore it is important to have open interfaces and support of standards for easy integration of term extraction results into other tools.

The de-facto-standard for terminology is a tab-delimited format, the most widely used terminology tool is Excel [LIA02] or other table-based applications. Standards like TBX or OLIF are used much less frequently. Therefore, term extraction tools should at least provide interfaces to formats from where they can easily imported (Trados

Multiterm as well as Comprendium Translator and other tools offer tab-delimited import functionality).

There is an additional aspect to this: Some tools, esp. for machine translation, require annotations from their imported terms, like part-of-speech, subject area, gender, inflection type and the like, which can be calculated from corpus analysis to some extent. Term Extraction tools are all the more useful the more information of this kind they can provide. This would then have an impact on the exchange format (support of the richer interfaces like OLIF, instead of the simple tab-delimited ones).

5 Examples: TermExtract and BiExtract

This section describes two term extraction tools, one for monolingual extraction (TermExtract), and one for bilingual extraction (BiExtract) which were developed on the basis of the above considerations.

5.1 TermExtract

TermExtract is a tool for monolingual term extraction, with the following features:

1. It analyses corpora consisting of one or multiple files. Corpora usually consist of several files, and users must not be bothered with the task of comparing and merging similar term candidate lists.
2. It analyses single words and multiwords. It uses hybrid technology (linguistic normalisation, frequency analysis, noun phrase multiword analysis) to determine term candidates. It does this for all 11 official languages of the European Union, plus some Slavic languages (Russian, Serbo-Croatian etc.).
3. It does not use sophisticated means of relevancy determination. It basically counts frequencies for linguistically normalised forms. However, it reduces noise in its output by linguistic filtering and normalisation, as described, by applying special filters on some parts of terms to eliminate e.g. prepositions which rarely occur in terms (like *as*, *in_front_of*, *near*, *southwest_of*), and by allowing users to specify a stop list of terms which they do not want to see. The goal is to avoid users having to scan through what they might regard as useless material.
4. It presents the results in form of a tab-delimited list which can be easily displayed, printed, and edited in standard tools like Excel. (An output example is given in fig. 2 above). In addition, it supports the OLIF exchange standard format, to allow for richer annotations. Output contains the term, its frequency, its part of speech, and a user-definable number of context sentences, plus some additional information users can select from.
5. It allows easy scanning through output lists in Excel, and marking of the lines which are not considered to be terms. Working through such a file from the most frequent term downwards ensures that a limited amount of editing time is spent in the most efficient manner.

Output of this tool is a glossary list of terms, in tab-delimited or OLIF format, for integration into further processing flows as shown in fig. 4.

5.2 BiExtract

BiExtract takes a glossary file, as e.g. produced by TermExtract, and tries to identify translation equivalents in aligned text.

- The memory material is assumed to be either in Ascii or in TMX [LIS02], to be flexible on the input side.
- Identification of translation equivalent candidates is done between single word and multiword candidates in all combinations. It is based on an iterative procedure which estimates the probability of each target candidate co-occurring with a given source term in the relevant memory segments, taking also into account the position and the orthography of the candidates. Noise is reduced by doing linguistic normalisation of candidates on both the source and the target side.
- Output is a file containing glossary terms for which no translation equivalent candidates could be found, and another file containing the translation candidates. Users can select to see just the best proposals, or the n best proposals, or all of them. They are offered frequencies and example contexts for the respective proposals. An output example is given in fig. 3 in the appendix.

- Output can be reviewed just like in the TermExtract case by using a tool like Excel, and the final result can be brought into either a tab-delimited form or a OLIF format for further integration.

5.3 Integration

The term extraction tools were used in two main work flows, as shown in fig. 4.

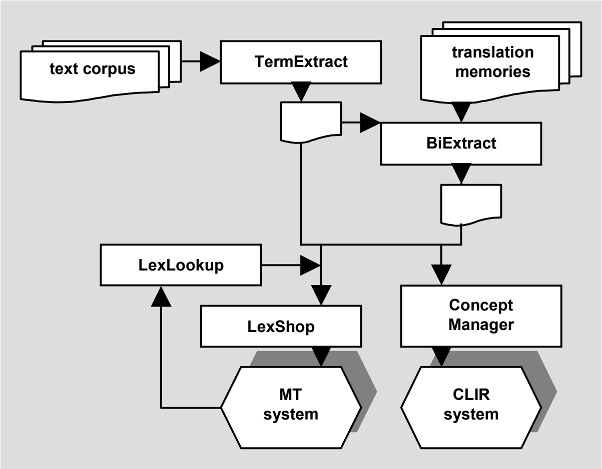

Fig. 4: Workflows

1. Machine Translation
The task was to build up a machine translation dictionary for a particular domain of a financial application for which complete glossaries did not yet exist [ROE02]. TermExtract and BiExtract are used to identify term candidates and their equivalents in the corpus material. Candidates then undergo a term recognition phase (dictionary lookup) to find the new / missing ones. These candidates are imported into the LexShop tool which is the coding environment of the Comprendium MT system [BGM01], to be verified and fully annotated. A tab-delimited exchange format is used for this.

2. Crosslingual Retrieval
Here the task was to set up a concept net for a cross-lingual retrieval system in the domain of law enforcement, for all European languages. Term Extraction tools were used to create the ontology backbone terms in a pivot language (English), BiExtract was used to find equivalents on aligned texts, and the result was imported into the ConceptManager which is a tool to verify and administer conceptual networks [JEA02]. Exchange format here is OLIF.

6 Acknowledgements

The tools were designed and implemented by Thilo Will. Special thanks to St.McLaughlin for many useful comments.

References

Arppe, A., 1995: Term Extraction from Unrestricted Text. Proc. NODALIDA 1995, Helsinki [ARP95]

Bernardi, U., Gieselmann, P., McLaughlin, St., 2001: A taste of MALT. Proc. MT-Summit VIII, Santiago di Compostela. [BGM01]

Bourigault, D., 1995: LEXTER, a Terminology Extraction Software for Knowledge Acquisition from Texts. In Proceedings of the 9th KAW. Banff, Canada. [BOU95]

Cabré Castellvi, M.T., Estopá Bagot, R., Vivaldi Palatresi, J., 2001:Automatic term detection: A review of current systems. In: Bourigault, D., Jacquemin, Chr., L'Homme, M.-Cl. (eds.), Recent Advances in Computational Terminology. 2001 [CEV01]

Dias, G., et al., 2000: Combining Linguistics with Statistics for Multiword Term Extraction: A Fruitful Association? Proc. RIAO 2000 [DEA00]

EAGLES Evaluation Working Group, 1999: The EAGLES 7-step recipe. http://www.issco.unige.ch/projects/eagles/ewg99/7steps.html [EAG99]

EAGLES, 1998: Preliminary Recommendations on Semantic Encoding: Multiword Recognition and Extraction. http://www.ilc.pi.cnr.it/EAGLES96/rep2/node38.html [EAG98]

Ferret, O., Fluhr, Chr., Rousseau-Hans, Fr., Simoni, J.-L., 2002: Building domain-specific lexical hierarchies from corpora. Proc. LREC 2002, Gran Canaria [FFR02]

Frantzi, K., Ananiadou, S., Mima, H., 2000 Automatic recognition of multi-word terms: the C-value / NC-value method. Int. J. Digit. Lib. 3, 2000: 115-130 [FAM00]

Goldmann, J., Wehrli, E., 2001: FipsCo: A Syntax-based system for Terminology Extraction. http://www.federation-nlp.uqam.ca/publications/01/goldman.pdf [GOW01]

Heid, U., 1999: Extracting Terminologically Relevant Collocations from German Technical Texts. In: Sandrini, Peter (ed.): Terminology and Knowledge Engineering (TKE '99), Innsbruck. Wien: TermNet, 241-255. [HEI99]

Heidemann, B., Volk, M., 1999: Evaluation of Terminology Extraction Tools. Report Univ. Zürich [HEV99]

Jackson, A., et al., 2002 ConceptManager: Pflege multilingualer Ontologien im crosslingualen Retrieval. Proc. ISI, Regensburg, 2002 [JEA02]

Le Moigno, S., et al., 2002: Terminology extraction from text to build an ontology in surgical intensive care. Proc. ECAI 2002 [LEM02]

Lieske, Chr., 2002: Pragmatische Evaluierung von Werkzeugen für die Term-Extraktion. in: Mayer, F., Schmitz, Kl.-D., Zeumer, J., eds.: eTerminology, Köln 2002 [LIE02]

LISA, 2002a: Translation memory Exchange (TMX) http://www.lisa.org/tmx/ [LIS02]

LISA, 2002b: Results of the Terminology Survey www.lisa.org/2001/termsurveyresults.html [LIA02]

Mikheev, A., 1994: Periods, Capitalized Words, etc. Computational Linguistics 3,1994 [MIK94]

OLIF, 2002: Open Lexicon Interchange Format http://www.olif.net/ [OLI02]

Pantel, P., Lin, D., 2001 A statistical corpus-based term extractor. http://www.cs.ualberta.ca/~lindek/papers/ai01.pdf [PAL01]

Röthlisberger-Käser, M., 2002: CLS Workflow – a translation workflow system. Proc. ASLIB 2002, London [ROE02]

Sauron, V., 2002: Tearing out the Terms: Evaluating Terms Extractors. Proc. ASLIB 2002, London [SAU02]

System QUIRK: System QUIRK, Language Engineering Workbench. http://www.mcs.surrey.ac.uk/SystemQ/ [SQU02]

Thurmair, Gr., 2000: TQPro: Quality Tools for the Translation Process. Proc. ASLIB 2000, London [THU00]

Valderrábanos, AS., Belskis, A., Moreno, L.I., 2002: Multilingual Terinology Extraction and Validation. Proc. LREC 2002, Gran Canaria [VBM02]

Vintar. S.: 2000: Using parallel corpora for translation-oriented term extraction http://www2.arnes.si/~svinta/babel.rtf [VIN00]

Xu, F., Kurz, D., Piskorski, J., Schmeier, S., 2002: Term Extraction and Mining od Term Relations from Unrestricted Texts in the Financial Domain. in: Proc. BIS 2002, Poznan [XKP02]

Appendix

term (normalized)	terms (textform)	freq	examples	types
fuel consumption	-> fuel consumption -> Fuel consumption	16	"... is a common-sense car that is anything but dull; its fuel consumption is low, but it is still ..."	common noun
top speed	-> top speed -> Top speed	13	... : it sprints from 0 to 100 km/h in 14.9 seconds and touches a top speed of 168 km/h.	common noun
manual gearbox	-> manual gearbox	13	... results from its distinctly more complex 3-litre technology with automatically controlled manual gearbox, start/stop system, aluminium engine and ...	common noun
front-wheel drive	-> front-wheel drive	12	"Both three-door and five-door versions with front-wheel drive have a luggage capacity of 350 litres, which increases to up ..."	common noun
power output	-> power output	12	"{ \ b - } 2.7 T V6, power output increased to 184 kW (250 bhp)"	common noun
Petrol engine	-> Petrol engines -> petrol engine -> petrol engines	9	Petrol engines:	common noun
ground clearance	-> ground clearance	8	The answer is obvious: a specialist vehicle for all kinds of surfaces needs variable ground clearance.	common noun
sports seat	-> sports seats -> sports seat	8	"The leather option can also be combined with front sports seats, which provide a most satisfactory degree of lateral ..."	common noun
four-cylinder engine	-> four-cylinder engine -> four-cylinder engines	8	The 1.9-litre four-cylinder engine with pump-injector fuel supply and an output of 96 kW (130 bhp) extends the ...	common noun
steering wheel	-> steering wheels -> steering wheel	7	"Together with one of the leather-covered steering wheels that are also available as optional extras, this colour scheme adds ..."	common noun
optional extra	-> optional extra -> optional extras	7	... is regulated depending on the position of the sun is an optional extra for this car that will appeal ...	common noun
centre console	-> centre console	6	"As well as the two Space Floor boxes and the centre console storage compartment with integral cup	common noun

			holder, the ..."	
shock absorber	-> shock absorber -> shock absorbers	6	New spring and shock absorber settings and modified bearing elasticities	common noun
first time	-> first time	6	"... series it can now be combined with five different engines, including for the first time a TDI unit."	common noun
soft top	-> soft top	5	The soft top with its large heated glass rear window can be opened in a single-stage process and stowed behind ...	common noun
pulling power	-> pulling power	5	... 66 kW (90 bhp) 1.9-litre TDI also surpasses the pulling power of many rivals with a petrol engine ...	common noun
peak torque	-> peak torque	5	The 1.2-litre TDI ' s peak torque of 140 Nm is produced at between 1800 and 2400 rpm.	common noun
cylinder head	-> cylinder head	5	"... TDI is the first direct-injection diesel engine to have a light-alloy cylinder head, but that is not all: ..."	common noun

Fig. 2: Term candidates extracted by the Comprendium TermExtract

source term	target term	freq.	SL context	TL context
Centre of Excellence	centre d' excellence	8 of 8	The EDU is regarded as a centre of excellence in this field attracting many visits from academic and law enforcement personnel .	L ' UDE est considérée comme un centre d ' excellence dans ce domaine et attire de nombreux visiteurs envoyés par les universités et les services répressifs .
Clandestine Immigration Network	filière d' immigration clandestine	8 of / 9	…7 Crimes involving clandestine immigration networks 7 Trafficking in human beings 7 Money laundering 8 Illicit vehicle trafficking 8 Special techniques …	… 8 Filières d ' immigration clandestine 9 Traite des êtres humains 9 Blanchiment de capitaux 9 Trafic illicite de véhicules …
Commission representative	représentant de la Commission	1 of 1	However , the Management Board may decide to meet without the Commission representative .	Le conseil d ' administration peut toutefois décider de délibérer en l ' absence du représentant de la Commission .
Committee of Ministers of the Council of Europe	Comité des ministres du Conseil de l'Europe	2 of 2	… shall take account of Recommendation No R (87) 15 of the Committee of Ministers of the Council of Europe of 17 September 1987 concerning the use of personal data in the police sector .	… la recommandation R (87) 15 du Comité des ministres du Conseil de l ' Europe , du 17 septembre 1987 , sur l ' utilisation des données à caractère personnel par la police .
Council Act	Acte du Conseil	2 of 5	Council Act drawing up the Convention based on Article K . 3 of the Treaty on European Union , …	Acte du Conseil portant établissement de la convention sur la base de l ' article K . 3 du traité sur l ' Union …
Council Act	décision du Conseil	2 of 5	By Council Act of 29 April 1999 , Mr . Jürgen Storbeck (Germany) was appointed Director of Europol …	Par décision du Conseil du 29 avril 1999 , M . Jürgen Storbeck (Allemagne) a été nommé directeur d ' Europol …
Council of Europe Convention	convention du Conseil de l'Europe	6 of 6	The collection , storage and processing of the data listed in the first sentence of Article 6 of the Council of Europe Convention of 28 January 1981 …	La collecte , le stockage et le traitement des données qui sont énumérées à l ' article 6 première phrase de la convention du Conseil de l ' Europe du 28 janvier 1981 …
Council of Heads of State	Conseil de ministres , du Conseil de chefs d'Etat	1 of 1	Forthcoming advice , decisions and priorities arising from experts ' meetings … , the Council of Heads of State and the Multidisciplinary Group on Organised Crime ….	Les conseils , décisions et priorités qui découleront des réunions d ' experts , … , du Conseil de chefs d ' Etat et du groupe multidisciplinaire sur la criminalité organisée ….
Council of Minister of Justice and Home Affairs	Conseil des ministres de la Justice et des Affaires intérieures	1 of 1	Europol is accountable to the Council of Ministers of Justice and Home Affairs .	Europol est responsable devant le Conseil des ministres de la Justice et des Affaires intérieures .
Council of Ministers	Conseil des ministres	5 of 7	* Accountability , Supervision and Management Europol is accountable to the Council of Ministers for Justice and Home Affairs .	* Obligations , contrôle et gestion Europol est responsable devant le Conseil des ministres pour la Justice et les Affaires intérieures .
Council of the European Union	Conseil de l' Union européenne	4 of 5	4 . The Secretary - General of the Council of the European Union shall notify all Member States of the date of entry into force of the amendments . contents	4 . Le Secrétaire général du Conseil de l ' Union européenne notifie à tous les Etats membres la date d ' entrée en vigueur des modifications . contents
Court of Auditors	Cour des comptes	3 of 3	…carried out by the Joint Audit Committee composed of three members appointed by the Court of Auditors of the European Communities .	…soumis à un contrôle effectué par le comité de contrôle commun composé de trois membres nommés par la Cour des comptes des Communautés européennes .
Criminal Police	police criminelle	1 of 1	… 7) the International Criminal Police Organization , forward the relevant information to it by whatever means may be appropriate .	… 7) à l ' Organisation internationale de police criminelle , de lui transmettre les informations correspondantes par tous moyens appropriés
Deputy Director	directeur adjoint	17of19	It appoints the Director and the Deputy Directors and adopts the budget .	Il lui incombe de nommer le directeur , les directeurs adjoints et d ' adopter le budget .

Fig. 3: Example Output BiExtract (example context shortened)