Knowledge-poor and Knowledge-rich Approaches for Multilingual Terminology Extraction

Béatrice Daille[1] and Helena Blancafort[2]

[1] University of Nantes, LINA, 2 Rue de la Houssinière,
BP 92208, 44322 Nantes, France,
beatrice.baille@univ-nantes.fr
[2] Syllabs, 53 bis rue Sedaine, 75011 Paris, France,
blancafort@syllabs.com

Abstract. In this paper, we present two terminology extraction tools in order to compare a knowledge-poor and a knowledge-rich approach. Both tools process single and multi-word terms and are designed to handle multilingualism. We run an evaluation on six languages and two different domains using crawled comparable corpora and hand-crafted reference term lists. We discuss the three main results achieved for terminology extraction. The first two evaluation scenarios concern the knowledge-rich framework. Firstly, we compare performances for each of the languages depending on the ranking that is applied: specificity score vs. the number of occurrences. Secondly, we examine the relevancy of the term variant identification to increase the precision ranking for any of the languages. The third evaluation scenario compares both tools and demonstrates that a probabilistic term extraction approach, developed with minimal effort, achieves satisfactory results when compared to a rule-based method.

1 Introduction

Identifying terms within a specific domain has been an active field of research since the early nineties [1,2]. Thanks to this research, several methods and tools have been developed for various applications such as information retrieval, information extraction, science and technology watch, and ontology acquisition. Twelve terminology extraction tools were described and compared in [3]. The various methods differ in how they process the corpus used as input, using anything from tokenization to syntactic analysis. Moreover, the tools differ in how they handle the output, e.g. with a manual validation process through dedicated interfaces. However sophisticated the processing of the input or output may be, all methods imply two steps that make up the core of the extraction process, namely:

1. **Step 1**: Identifying and collecting candidate terms (CT), i.e. term-like units in the texts (mostly multi-word phrases).
2. **Step 2**: Ranking the extracted CT to keep the most representative of the specialised field and the most useful for the target application.

Most of the tools are designed for one language with the exception of Termostat[4][3] which processes the Romance languages (French, English, Spanish, Italian and Portuguese) and Acabit[5][4] which works with French, English and Japanese. Concerning step 1, the CT can be made up of either multi-word terms only, as is the case with Acabit, or of both single (SWT) and multi-word terms (MWT), as is the case with Termostat. Acabit is the only tool that takes MWT variations into account i.e. the relation between basic and extended terms, as well as several forms of pattern switching. An example of pattern switching in French is the transformation of noun phrases with a Noun Adjective structure to a Noun Preposition Noun structure, e.g. as with the synonym terms *excès pondéral* ↔ *excès de poids*, both meaning 'overweight'.

The objective of step 2 is to measure the termhood of a CT, i.e. the degree in which a CT is related to a domain-specific concept [6]. Several methods have been proposed for this task: the C-value method [7], based on the frequency of occurrence and term length weights the termhood of MWT according to their nested occurrences. The more an item is part of longer terms, the more it is likely to be a term. This measure applies only to MWT. But most of the work carried out uses statistical measures to compute the termhood of the CT. They are based on frequency counts and frequency distributions in the domain-specific corpora from which the CT are extracted [5]. Another research line compares the frequency of a CT in a domain-specific corpus and a language-general corpus [8,4]. The potential of different statistical measures (including an n-gram model) was evaluated by [9] to distinguish terms from non-terms in a CT list. They concluded that the number of occurrences ($freq$) is a very good indicator of the quality of a CT as well as the domain-specificity score (d_s).

The two terminology extraction tools that are presented in this paper encompass the main capabilities of current state-of-the-art tools. To complete step one, they handle SWT and MWT. In addition to this, the knowledge-rich method also processes SWT and MWT variation. To achieve the goal of step 2, they rank the CT according to specificity using a general language comparison corpus. The domain specificity d_s of a CT as defined by [8] is the quotient of its relative frequencies in both the monolingual comparable corpus (the domain corpus) rf_d and a general language corpus rf_g.

$$d_s(ct) = \frac{rf_d(ct)}{rf_g(ct)} = \frac{\frac{freq(ct)}{\sum_w freq(w)}}{\frac{freq(ct)}{\sum_{w'} freq(w')}} \tag{1}$$

Furthermore, the algorithms for steps 1 and 2 are formulated in a language-independent fashion. For the knowledge-rich approach, the language is a parameter: basic term and term variant patterns are formulated in terms of POS tags adopting the Multext POS tag annotations[5]. This language-independency allows

[3] http://olst.ling.umontreal.ca/~drouinp/termostat_web/doc_termostat/doc_termostat.html

[4] http://www.bdaille.com/

[5] http://aune.lpl.univ-aix.fr/projects/multext/

us to integrate a new language either by training the probabilistic tool, or by providing a defined set of language resources. In the next sections, we describe in detail the knowledge-poor approach, followed by the knowledge-rich approach.

2 Knowledge-poor Approach for Term Extraction

The knowledge-poor approach is based on a probabilistic tool. In contrast to a knowledge-rich tool that needs a POS tagger and hand-written rules to identify term candidates, the probabilistic tool simply requires a large raw corpus and a second smaller corpus with manually annotated sentences (noun phrases). This small corpus can be annotated by a linguist in a single day. The knowledge-poor approach is interesting for languages for which a POS tagger is not available. This can be the case when developing tools in an industrial context where open-source resources cannot be used because of license restrictions. It is also useful for under-resourced languages, for which annotated corpora are rare. It is possible today to compile a corpus for an under-resourced language from the web and use it as training material.

2.1 Training a Pseudo POS Tagger

Part-of-speech induction is the task of clustering words into word classes (or pseudo-POS) in a completely unsupervised setting. No prior knowledge such as a morphosyntactic lexicon or annotated corpus is required. The only resource needed is a relatively large training corpus. As in [10] and based on [11], we use Clark's tool[6] [12]. This tool for POS induction uses a distributional clustering algorithm and includes morphological information. The clustering algorithm is based on a cluster bigram model [13]. It is the highest performing system in almost every language, and one of the fastest methods. Performance and speed are important factors in an industrial context. The pseudo-POS tagger was trained using 50 clusters, after having run experiments with 20 to 100 clusters.

2.2 Corpora to Train the Pseudo POS Tagger

The input to the pseudo POS tagger is a tokenized corpus. For English, French and Spanish corpora we used the newstrain-08 corpora, monolingual language model training datasets which were provided for the WMT'09 translation task. Their size is approximately 2.5 GB for 500 million tokens (English), 1 GB for 175 million tokens (French), and 250 MB for 50 million tokens (Spanish). The German pseudo-POS tagger was trained on the German Wortschatz (350 MB for 60 million tokens), which performed a little better than the German newstrain-08 corpus. For Latvian we used a web-based corpus provided by an industrial partner.

[6] Available here: http://www.cs.rhul.ac.uk/home/alexc/pos2.tar.gz

2.3 CRFs to Train a Term Candidate Extractor

The terminological extraction task is close to the definition of the noun phrase chunking task, which is itself a subtask of the more general shallow parsing task. Traditional approaches in shallow parsing rely on a pre-processing step with a POS tagger. As in [10], the tool adopts the strategy of [14], who achieved near state-of-the-art results on the English supervised shallow parsing task using Conditional Random Fields (CRFs) [15]. CRFs enable a large number of features to be added in a flexible way. We used the CRF++ implementation, distributed under the GNU Lesser General Public License and new BSD License. The CRF model is trained on a tokenized corpus where sentences are separated by empty lines. Each line contains a word of the sentence together with its noun-phrase chunk tag. The tag is either B, I or O. B indicates the beginning of the noun phrase. I stands for inside the noun phrase. O represents tokens that do not belong to any phrase. In addition, the pseudo POS tag was used as one of the training features.

2.4 Training the CRF-based Term Candidate Extractor

To train the probabilistic CRF-based term candidate extractor, we used manually annotated corpora in each language. Small corpora with 300 to 600 sentences in French, English, Spanish and German were first automatically annotated with a symbolic term extractor and then manually corrected by a linguist. For Latvian, the corpus was manually annotated from scratch, as no rule-based system was available at that time. Table 2 gives more detailed information about the size and type of corpora used (general language corpus vs. domain-specific language corpus, raw corpus vs. manually annotated corpus).The domain-specific corpora were compiled using the focused web crawler Babouk [16].

3 Knowledge-rich Term Extraction Framework

The knowledge-rich approach requires linguistic knowledge to identify the CT (step 1). The following resources are needed:

- tools for the linguistic processing of the specialised texts: tokenizers, POS taggers and lemmatisers;
- hand-crafted patterns for the identification of single and multi-word CT based on POS tags;
- hand-crafted rules for the grouping of term variants.

The two first resources are mandatory, the last one is optional but needed to handle term variation.

For the linguistic processing step, we decided to use the TreeTagger[7] [17], because it performs both POS annotation and lemmatisation for 15 languages. The choice of the TreeTagger was thus determined by the number of languages available.

[7] `http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/`

Table 1. MWT: syntagmatic compounds of the noun category

	Pattern	Example	English translation
English	N N	*rotor blade*	
	A N	*renewable energy*	
French	N A	*énergie renouvelable*	*renewable energy*
	N S:p N	*caisse de résonance*	*sounding-box*
German	A N	*fossiler Energieträger*	*fossil energy source*
	N S:p N	*Netzintegration von Windenergie*	*grid integration of wind energy*
Latvian	A N	*meteoroloģiskā stacija*	*meteorological observing station*
	N:g N	*gaisa blīvums*	*air density*
Russian	A N	*метеорологический станция*	*meteorological station*
	N N:g	*выработка энергия*	*production of energy*
Spanish	N A	*energía eólica*	*wind energy*
	N S:p N	*fuente de energía*	*energy source*

3.1 Patterns for Candidate Term Identification

The term candidates and term variants are identified by means of patterns using Multext word classes. SWT are nouns or adjectives. MWT are noun phrases of length 2 or of length 3. A MWT of length 2 is a noun phrase with a head noun and a dependent of level 1, either an argument or a modifier (noun, adjective, etc.). To illustrate the patterns, the two main patterns for each language are provided (see Table 1).

3.2 Patterns for Term Variation

The term variant grouping functionality is optional and takes place once the CT has been annotated as a SWT or MWT. Several methods are implemented depending on the linguistic operation involved. There are 3 sub-functionalities: the detection of spelling term variants based on string distances, the detection of morphological variants based on monolingual lists of affixes and the detection of syntactic variants based on pattern rules on feature structures.

The spelling variants such as *air flow* ↔ *airflow* are detected by means of the edit distance. Morphological variants are handled by the Treetagger lemmatiser. Syntactic term grouping based on pattern rules consists in checking binary relation satisfactions between a pair of terms. For example, a binary relation is made between the MWT *énergie éolien* 'wind energy' and *énergie renouvelable éolien* 'renewable wind energy' according to the specifications of the French variant grouping pattern. The rule refers to a modification variant and expresses that a term whose components are a noun and an adjective should be related to any terms whose components are a noun and two adjectives, if and only if, they have the two same nouns and adjectives on the borders. Such a grouping pattern is written in a language-dependent grouping pattern specification file as follows:
Original term: N_0 A_1/ Variant: N_0 A_2 A_1
The term and the variant elements that are shared are numbered with the same

Table 2. Size of the domain-specific corpora in the domains of wind energy and mobile communication.

Language	De	En	Es	Fr	Lv	Ru
Wind energy: nb tokens	358,602	313,954	454,095	314,551	220,823	323,946
Mobile technologies: nb tokens	474,316	303,972	474,534	437,505	306,878	318,225

values. For example, in the above rule, the N_0 lemma of the term and of the variant are identical (same with A_1). The grouping patterns are thus not oriented: they are symmetric. The base term is defined as the most frequent of both items. Syntactic term grouping patterns cover the following syntactic phenomena: modification, coordination, compounding, decompounding. There is an average of 14 MWT patterns and 10 MWT grouping rules per language.

4 Resources

To assess the knowledge-poor and the knowledge-rich approaches, we use manually-checked comparable corpora, as well as hand-crafted reference term lists (RTL). The comparable corpora have been collected with a focused web crawler [16]. The corpus size varies from 300,000 to 400,000 tokens, depending on the domain and language. The RTL of around 130 terms in a specialised domain have been compiled to serve as a GOLD STANDARD for the evaluation of the tools. It should be noted that the wind energy domain corpora used to make our experiments are subsets of the corpora used to build the RTL with the exception of Latvian. For the mobile domain, the corpora are the same as those used for the compilation of the RTL with the exception of French, that differs slightly in terms of size.

4.1 Domain-specific Corpora

Because the tools for terminology extraction are particularly useful for new domains with poor terminological resources, the corpora used are related to two emerging domains: wind energy, a subdomain of renewable energy, and mobile technologies, a subdomain of computer science.

4.2 Reference Term Lists

The terms and variants are listed in the lemma form provided by the TreeTagger. The Reference Term Lists (RTL) were created manually to serve as a gold standard for the evaluation of the term extractors. They include both single (SWT) and multi-word terms (MWT) with their corresponding base terms and variants. One of the constraints is occurence in the corpus: a minimum term frequency of occurence was fixed, 5 for MWT and 10 for SWT. To decide on the termhood of

Table 3. Size of the RTL and corresponding corpora

Language	De	En	Es	Fr	Lv	Ru
Wind energy corpus: nb tokens	1,700,000	750,855	453,953	710,702	220,823	2,328,609
RTL size - Terms	132	128	136	126	129	107
RTL size - Variants	25	59	65	75	76	11
Mobile technologies: nb tokens	474,316	308,263	473,273	302,634	306,878	372,459
RTL size - Terms	159	140	137	130	139	103
RTL size - Variants	2	17	55	19	57	13

a term with respect to the domain, several linguistic criteria were applied [18]. Moreover, large terminology banks or specialised dictionaries (e.g. TERMIUM[8], Grand Dictionnaire Terminologique, IATE[9] and EuroTermBank[10] were used to check the terms.

Table 3 gives the RTL size expressed by the number of terms for each language and each domain. It recalls the size of the monolingual corpora used to build the RTL. The number of reference terms does not include the number of variants which are listed on a separate line. Compilation of these RTL is described in detail in [19]. For all the languages, we get an average of 130 RT. The number of variants depends on the languages. There are a higher number of variants encoded in the French, Spanish, English and Latvian lists, and nearly none for Russian for the wind energy domain. The wind energy domain has more variants than the mobile technologies domain. Spanish and Latvian display a large number of variants for both domains.

4.3 Reference Corpus of General Language

To calculate the domain-specificity (see equation 1), a general language corpus is needed. The general language corpora are a compilation of newspaper and Europarl data with 10 to 15 million words depending on the language. As an example, the German reference is based on the German newspaper TAZ and contains 20 millions tokens.

5 Results and Discussion

In this section, we evaluate the tools using the F-measure [20]. Precision is the percentage of the number of reference terms (RT) over the total number of canditate terms (CT) acquired from the corpus.

$$Precision = \frac{count_{RT=CT}}{count_{CT}} \tag{2}$$

[8] `http://termiumplus.gc.ca`

[9] `iate.europa.eu/`

[10] `http://www.eurotermbank.eu`, etc.

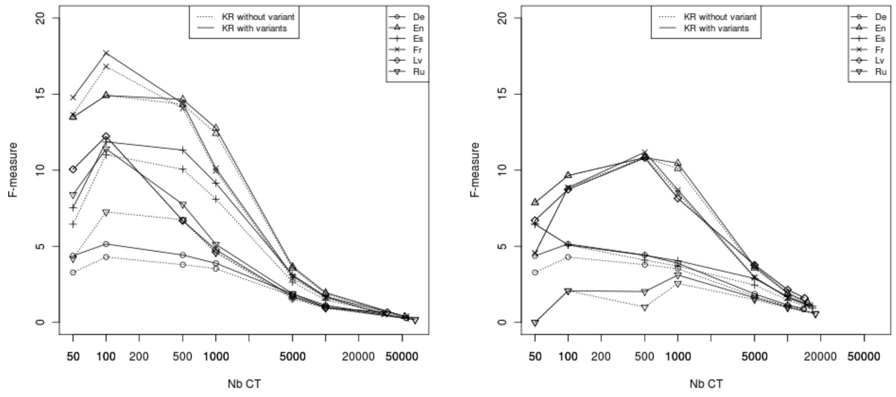

Fig. 1. F-measure ranking according to specificity (left) and the number of occurrences (right) for the wind energy domain

Recall is the percentage of the reference terms over the total number of terms from the reference term list contained in the corpus.

$$Recall = \frac{count_{RT=CT}}{count_{RT}} \tag{3}$$

The F-measure is defined as the harmonic mean of precision and recall. The F-measure has a value that is bound between 0 and 1, but we use here a percentage value.

$$F - measure = \frac{100 \cdot 2 \cdot precision \cdot recall}{precision + recall} \tag{4}$$

The first two evaluations focus on the knowledge-rich framework depending on two parameters (ranking and term variation), while the third evaluation compares both tools. The three evaluation scenarios are the following:

1. ranking of CT: the number of occurrences vs. the specificity (equation 1) (knowledge-rich framework);
2. term variants: the handling or not of term variation (knowledge-rich framework);
3. method: the knowledge-rich vs. the knowledge-poor approach

The CT ranking is given by the specificity value (d_s) or the number of occurrences ($freq$) in decreasing order. Table 4 illustrates the ranking of the CT alone and Table 5 with term variant recognition.

We consider a CT as correct if it matches a RT included in the RTL. The matches are made between lemmas. For the term extraction tools without variant recognition, we compare the CT and the RT. If they match, we return the CT rank. For the term extraction tools with variant recognition, we compare the CT or one of its variants and the RT, if either the term or one of its variants matches, we return the CT rank.

Table 4. CT ranking without variant recognition

Rank	Term or variant
1	wind project
2	wind energy project
3	aerodynamic
4	wind energy
5	wind turbine energy
6	onshore wind energy
7	energy from wind
8	small-scale wind energy

Table 5. CT ranking with variant recognition

Rank	Term or variant	Term or variant
1	T	wind project
1	V	wind energy project
2	T	aerodynamic
3	T	wind energy
3	V	wind turbine energy
3	V	onshore wind energy
3	V	energy from wind
3	V	small-scale wind energy

5.1 Ranking of Candidate Terms: Specificity vs. Occurrences

For the knowledge-rich approach, we compute the F-measure (see equation 4) and then we compare the ranking provided by the specificity and that provided by the number of occurrences for all the languages. We only use the terms of the RTL.
For all languages with the exception of German, and whether term variation is applied or not, the ranking of the specificity outperforms that of the number of occurrences. In a good ranking reflecting the termhood [6], the terms should appear on the top of the list: we clearly see a difference of shape between the ranking of the specificity and the occurrence. The higher results are obtained until the top 100 to top 500 candidates, with a clear decrease afterwards, although the ranking by occurrence does not show a stark contrast. This is striking for French, where the F-measure has nearly doubled from 9 to 17 points, as well as for Russian, where the F-measure rises from 2 to 12 for the top 100 CT. The only language for which there is no difference is German: this could only be explained by the empty intersection between the terms of the wind energy domain and the terms of the general language domain, which is the result of the compounding process. Figure 1 gives the F-measure on the wind energy corpus according to the specificity ranking for all the languages.

For the following two evaluation scenarios, we will only consider the ranking based on specificity.

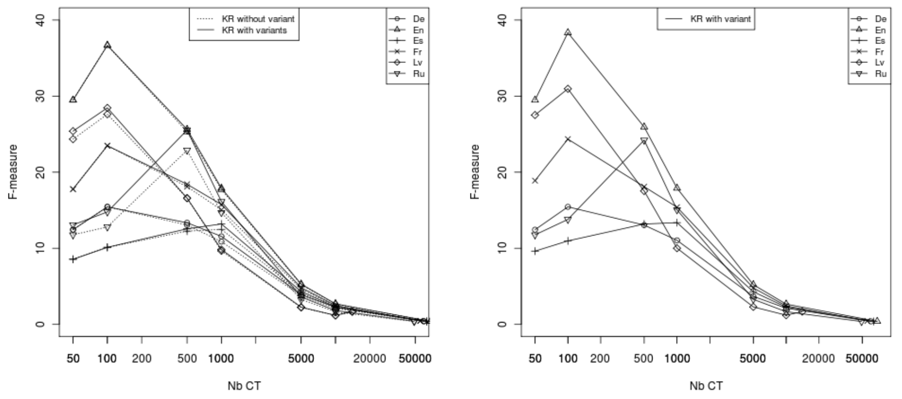

Fig. 2. F-measure ranking the CT (left) and the CT with variants (right) according to specificity for the mobile communication domain

5.2 Impact of Term Variation: with and without Variants

Here we examine the impact of term variant recognition on terminology extraction when using a knowledge-rich approach. When dealing only with the recognition of terms (see Fig. 1 left and 2 left), the detection of variants increases the F-measure for the highest ranks: this is the case for almost all languages with the exception of English and Latvian. We compare the recognition of terms alone using the RTL with terms only, and the recognition of terms and variants using the RTL with terms and variants. Fig. 2 shows the results obtained for both scenarios for the mobile domain. When we associate a set of synonymic variants to a term, the F-measure increases for almost all languages, with the exception of Russian (See Fig. 2 right). Looking through the Russian results, it appears that the recognition of terms increased but not the recognition of variants. This means that the knowledge-rich approach is able to correctly identify variants and that some variants appear in the list before the term to which they are related to. This result has however to be approached with care as RTL do not contain the same number and kind of variants.

5.3 Knowledge-poor vs. Knowledge-rich approach

In this subsection we compare the results provided by the knowledge-poor and the knowledge-rich approaches by examining the F-measure results obtained for the domains of wind energy and mobile communication. CT are ranked by the specificity score. Figures 3 show the results for English, Spanish and German. In English, the knowledge-poor method obtains similar results for both domains. The knowledge-poor approach performs better than the knowledge-rich approach for Spanish for the first 100 CT in the mobile domain but not for the wind energy

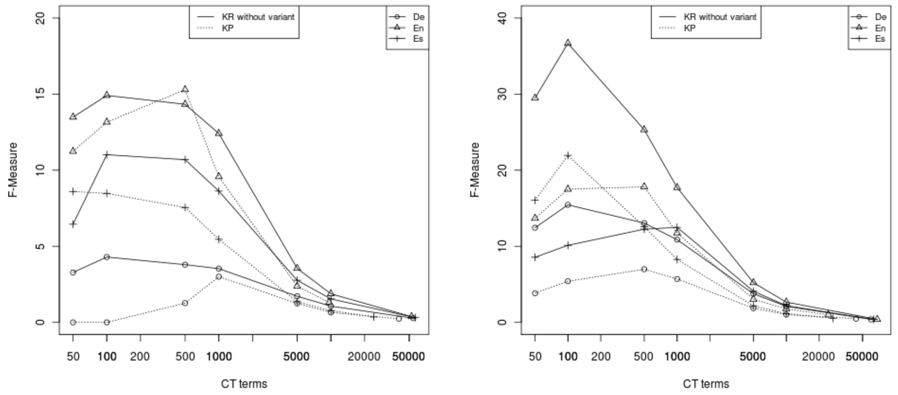

Fig. 3. F-measure ranked the CT according to specificity for the wind energy (left) and the mobile domain (right) either with knowledge-rich or knowledge-poor approaches

domain. For the other languages, French, German and Latvian, the knowledge-rich approach outperforms the probabilistic approach. However, the results for German are as low in both cases. This demonstrates the limits of the multilingual framework that applies the same symtagmatic approach for all languages. In addition to using syntactic patterns, morphological analysis is required for a language with productive compounding, e.g. German. In German, morphological compounds are much more frequent than MWT: 52% of nouns were reported to be compounds by [21] in the renewable energy domain. This means that a multilingual framework dedicated to terminology extraction should implement both morphological and syntactical processing. Concerning the knowledge-poor approach, the results are generally below the knowledge-rich approach. There are two reasons for this. First, the CT are not lemmatised, which is a severe obstacle for most of the languages, with the exception of English. As a matter of fact, the knowledge-poor approach delivers good results for English. Secondly, the knowledge-rich terminology extraction focuses on MWT of length 2 and 3, while the knowledge-poor approach, extracts MWT of unconstrained length. As the RTL do not include MWT with more than 3 tokens, longer MWT do not match with the terms in the RTL, e.g. the CT *small scale domestic wind turbine system* (Rank 428 in the CT list).

5.4 Related work

To our knowledge, no previous research has been done to use a probabilistic method for term extraction based on POS induction. An experiment on a similar task, namely on shallow parsing based on the English CoNLL 2000 corpus, is described in [10]. These experiments validate the knowledge-poor approach. They obtain interesting performances compared to a parser based on the Brill

[22] tagger. For the POS induction step, they use a corpus of 1 million words, and for the shallow parsing training, a few hundred annotated sentences. With the POS induction approach, they obtain an F-measure of 93.98 on noun phrase extraction against an F-Measure of 94.29 achieved with the Brill tagger. These experiments have been carried out in English. Work on CRF for shallow parsing is mainly carried out on English, probably because of the need to have a large amount of annotated data. [23] report work on chunking corpora using the Arabic Treebank and [24] report work on the UPENN Chinese Treebank-4. In our paper we include work on English, German and Spanish.

6 Conclusion

In this paper we have presented two terminology extraction tools that are designed to process a wide range of languages: a knowledge-poor and a knowledge-rich line. Both deal with SWT and MWT, and rank the CT according to domain specificity. The knowledge-poor approach is based on a probabilistic tool that performs pseudo POS tagging and thus could be an alternative for languages for which a POS tagger is not available. The knowledge-rich approach implements the main properties of state-of-the-art tools, and in addition handles term variation. Moreover, it is designed in a language-independent fashion: a language is a parameter where only term patterns and, optionally, term variation rules are required. We evaluated both approaches for two emerging domains and for six languages using hand-crafted reference term lists and manually-checked crawled comparable corpora. The results confirm that the specificity ranking outperforms the frequency of occurrence ranking and that the handling of term variants improves the ranking for the first candidate terms. Finally, the knowledge-poor approach provides satisfactory results with a minimal effort. In the future, it would be interesting to consider a scenario where a POS tagger is available and implement a method that uses a POS tagger but no hand-crafted rules, and then compare the results to the knowledge-rich and knowledge-poor tools that we have presented here.

References

1. Bourigault, D., Jacquemin, C., L'Homme, M.C., eds.: Recent Advances in Computational Terminology. John Benjamins (2001)
2. Kageura, K., Daille, B., Nakagawa, H., Chien, L.F.: Recent trends in computional terminology. Terminology **10**(2) (2004) 1–21
3. Cabré, M.T., Bagot, R.E., Platresi, J.V.: Automatic term detection: A review of current systems. In Bourigault, D., Jacquemin, C., L'Homme, M.C., eds.: Recent Advances in Computational Terminology. Volume 2 of Natural Language Processing. John Benjamins (2001) 53–88
4. Drouin, P.: Term extraction using non-technical corpora as a point of leverage. Terminology **9**(1) (2003) 99–117

5. Daille, B., Gaussier, E., Langé, J.M.: Towards automatic extraction of monolingual and bilingual terminology. In: Proceedings of the Fifteenth International Conference on Computational Linguistics (COLING-94), Kyoto, Japon (1994) 515–521
6. Kageura, K., Umino, B.: Methods for automatic term recognition: A review. Terminology **3**(2) (1996) 267–278
7. Frantzi, K.T., Ananiadou, S., Mima, H.: Automatic recognition of multi-word terms: the c-value/nc-value method. Int. J. on Digital Libraries **3**(2) (2000) 115–130
8. Ahmad, K., Davies, A., Fulford, H., Rogers, M.: What is a term? the semi-automatic extraction of terms from text. Translation Studies: An Interdiscipline (1994) 267–278 John Benjamins.
9. Drouin, P., Langlais, P.: Évaluation du potentiel terminologique de candidats termes. In: 8th Journées Internationales d'Analyse Statistique des Données Textuelles (JADT 2006), Besançon, France (2006) 379–388
10. Guégan, M., de Loupy, C.: Knowledge-poor approach to shallow parsing: Contribution of unsupervised part-of-speech induction. In Angelova, G., Bontcheva, K., Mitkov, R., Nicolov, N., eds.: RANLP, RANLP 2011 Organising Committee (2011) 33–40
11. Christodoulopoulos, C., Goldwater, S., Steedman, M.: Two decades of unsupervised pos induction: How far have we come? In: Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing. EMNLP '10, Stroudsburg, PA, USA, Association for Computational Linguistics (2010) 575–584
12. Clark, A.: Combining distributional and morphological information for part of speech induction. In: EACL, The Association for Computer Linguistics (2003) 59–66
13. Ney, H., Essen, U., Kneser, R.: On structuring probabilistic dependencies in stochastic language modelling. Computer Speech and Language **8** (1994) 1–38
14. Sha, F., Pereira, F.: Shallow parsing with conditional random fields. Technical Report CIS TR MS-CIS-02-35, University of Pennsylvania (2003)
15. Lafferty, J.D., McCallum, A., Pereira, F.C.N.: Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: Proceedings of the Eighteenth International Conference on Machine Learning. ICML '01, San Francisco, CA, USA, Morgan Kaufmann Publishers Inc. (2001) 282–289
16. de Groc, C.: Babouk: Focused Web Crawling for Corpus Compilation and Automatic Terminology Extraction. In: The IEEEWICACM International Conferences on Web Intelligence, Lyon, France (2011) 497–498
17. Schmid, H.: Probabilistic Part-of-Speech Tagging Using Decision Trees. In: Actes, International Conference on New Methods in Language Processing. (1994)
18. L'Homme, M.C.: La terminologie : principes et techniques. Les Presses de l'Université de Montréal (2004)
19. Loginova, E., Gojun, A., Blancafort, H., Guégan, M., Gornostay, T., Heid, H.: Reference lists for the evaluation of term extraction tools. In: Proceedings of the Terminology and Knowledge Engineering Conference (TKE'2012). (2012)
20. Chinchor, N.: Muc4 evaluation metrics. In: In Proceedings of the 4th conference on Message understanding. MUC4 '92, Stroudsburg, PA, USA, Association for Computational Linguistics (1992) 22–29
21. Weller, M., Gojun, A., Heid, U., Daille, B., Harastani, R.: Simple methods for dealing with term variation and term alignment. In: In Proceedings of the 9th International Conference on Terminology and Artificial Intelligence, TIA 2011. (2011)

22. Brill, E.: Transformation-based error-driven learning and natural language processing: A case study in part-of-speech tagging. Computational Linguistics **4**(21) (1995) 543–565
23. Diab, M., Hacioglu, K., Jurafsky, D.: Automatic tagging of arabic text: from raw text to base phrase chunks. In: In 5th Meeting of the North American Chapter of the Association for Computational Linguistics/Human Language Technologies Conference (HLT-NAACL04. (2004) 149–152
24. Chen, W., Zhang, Y., Isahara, H.: An empirical study of chinese chunking. In Calzolari, N., Cardie, C., Isabelle, P., eds.: ACL, The Association for Computer Linguistics (2006)