Bootstrapping Term Extractors for Multiple Languages

Ahmet Aker, Monica Lestari Paramita, Emma Barker, Robert Gaizauskas
Department of Computer Science, University of Sheffield
Sheffield, S1 4DP, United Kingdom
a.aker@dcs.shef.ac.uk, m.paramita@sheffield.ac.uk, e.barker@dcs.shef.ac.uk, r.gaizauskas@sheffield.ac.uk

Abstract

Terminology extraction resources are needed for a wide range of human language technology applications, including knowledge management, information extraction, semantic search, cross-language information retrieval and automatic and assisted translation. We report a low cost method for creating terminology extraction resources for 21 non-English EU languages. Using parallel corpora and a projection method, we create a General POS Tagger for these languages. We also investigate the use of EuroVoc terms and Wikipedia to automatically create a term grammar for each language. Our results show that these automatically generated resources can assist the term extraction process, achieving similar performance to manually generated resources. All POS tagger and term grammar resources resulting from this work are freely available for download.

Keywords: POS Tagger, term grammar, EU languages

1. Introduction

Term extraction tools are important for a wide range human language technology applications, including knowledge management, information extraction, semantic search, cross-language information retrieval and automatic and assisted translation. In translation and cross-language information retrieval applications, the requirement is typically for *bilingual* terminologies. However, such terminologies are commonly built by following a *symmetric* approach (Moore, 2003), where for each document pair in a parallel or comparable corpus, the source and target documents are first independently processed by a monolingual term extraction tool, after which some technique is used to pair the extracted terms to form a list of bilingual terms (see, e.g., Aker et al. (2013)).

Monolingual term extraction tools are, therefore, extremely important language resources. However, at present such resources exist only for a relatively small number of resource-rich languages, such as English, German and French. Furthermore, many term extraction approaches rely upon the prior existence of part-of-speech (POS) taggers and term grammars (typically just sequences of POS tags that syntactically characterise terms) and these too are not available for many languages. If these resources must be manually developed for each new language in order to build term extraction capability, then building term extractors for new languages is a considerable undertaking.

In this paper we propose a low cost method for creating terminology extraction resources for new languages. Our method exploits a number of existing resources: POS taggers for English, parallel corpora, cross-language word alignment tools and a small existing multilingual terminology thesaurus. It also relies on the conjecture that fined-grained POS tagging is not needed for term extraction. We illustrate the approach by developing term extraction resources for 21 non-English EU languages, all of which we make freely available for download.

In brief our method is as follows: first, we follow the approach of Das and Petrov (2011) to induce a generalised POS tagger for each non-English language T by taking the DGT-TM (Steinberger et al., 2012) English-T parallel data. We then tag the English side of the corpus using an available English POS tagger and map the POS tags on the English side to generalised tagsets. Using word-to-word alignment information obtained through GIZA++ (Och and Ney, 2000; Och and Ney, 2003), we project the generalised English POS tags in each English sentence to the target language sentence and train a new POS tagger for T on the tagged target language sentences. Next, we induce term grammars for T. We use an existing small scale terminology resource for T and project the terms in it onto Wikipedia articles in T, marking sequences of words in the articles that match terms. We then POS tag the sentences containing term matches, using the newly created POS tagger for T, and record the POS sequences of those word sequences marked as terms. The resulting POS sequences are taken as the term grammar. Finally, we supply the induced POS tagger and term grammar for T as arguments to an existing freely available term extractor (Pinnis et al., 2012), which applies them on texts in T to extract terms.

We first describe our POS tagger induction method (Section 2) and then discuss the term grammar creation (Section 3). We integrate the POS tagger and term grammar into a term extraction tool (Section 4) and evaluate the performance (Section 5). Finally, we describe the resources we publish with this paper (Section 6) and conclude with Section 7.

2. Creation of POS Taggers

To create a POS tagger for a target language, we make use of parallel data. The parallel data consist of *source sentences* from a resource-rich language, such as English, for which gold standard POS-tagged training data exist, and *target sentences* from an under-resourced language for which we aim to create a new POS tagger. We follow the approach of POS tag projection reported in various studies (Yarowsky et al., 2001; Das and Petrov, 2011). First, we POS-tag the source sentences (i.e. English) using a POS tagger trained on gold standard training data and then project the source POS tags to the target sentences. We describe the parallel data we use and our method for POS tag projection in the following sections.

2.1. Bilingual parallel corpora

We use the DGT-TM (Steinberger et al., 2012) parallel data which are available for 22 EU language pairs when English is taken as the source language. These are all the official EU languages, at time of writing. However, we exclude English-Irish because the amount of parallel data available in DGT-TM for this pair is insufficient for our approach. Table 1 shows the number of sentence pairs available in the DGT-TM corpora for the remaining 21 language pairs.

Language Pair	Sentence Pair
EN-BG	1,810,612
EN-CS	3,633,782
EN-DA	3,179,359
EN-DE	3,207,458
EN-EL	3,016,402
EN-ES	3,175,608
EN-ET	3,652,963
EN-FI	3,135,651
EN-FR	3,692,787
EN-HU	3,789,650
EN-IT	3,221,060
EN-LT	3,736,907
EN-LV	3,722,517
EN-MT	2,130,282
EN-NL	3,164,924
EN-PL	3,665,112
EN-PT	3,620,006
EN-RO	1,781,306
EN-SK	3,721,620
EN-SL	3,689,972
EN-SV	3,248,207

Table 1: DGT-TM parallel data

2.2. Projection technique

For POS tag projection, we first train an English POS tagger using the PennTreeBank corpus (Marcus et al., 1993). First, we replace the POS tags within this corpus by more general tags: *NOUN, VERB, DET, ADP, ADJ, PRT, ADV, NUM, CONJ, PRON,* . and *X*, using the "universal tagset" and mapping approach described in Petrov et al. (2011). After all tags are replaced by more general tags, we train a bi-gram HMM to obtain an English POS tagger. We use the HMM implementation in LingPipe[1] with the default features. We tested the performance of the English POS tagger on the ConNLL testing data (containing the universal tags) and obtained 97% accuracy. Finally, we POS-tag the English sentences in the parallel corpora using this new tagger. Note that in place of the HMM approach, any other supervised machine learning approach could equally well have been adopted instead, e.g., CRFs (Sha and Pereira, 2003), SVMs (Giménez and Marquez, 2004), etc. Since our aim is not to compare different machine learning approaches and because the LingPipe implementation of an HMM tagger fitted well with our experimental set-up, we selected the HMM tagger.

To perform POS-tag projection, word alignment information between the source and target words in the parallel sentences is required. We obtain this alignment information using the Giza++ toolkit (Och and Ney, 2000; Och and Ney, 2003). We run it in both directions (source-to-target, followed by target-to-source) and then refine the alignments using the "grow-diag-final-and" strategy. When projecting the POS tags from the source language to the target language, we consider only three alignment types:

1. **One-source-word-TO-one-target-word:** We project the source word POS to the target word.
2. **Many-source-words-TO-one-target-word:** We insist that all English words projected to the same target word must have the same POS type and project this POS tag to the target word.
3. **One-source-word-TO-many-target-words:** We project the POS tag of the source language to all the target words.

Source POS tags are projected to the target side subject to these three conditions. The result is a set of

[1] http://alias-i.com/lingpipe/

POS-tagged target sentences which we can then use to train a POS tagger in the target language. We refer below to POS taggers trained on projected POS tags as *General POS Taggers* or *GenTaggers*. We use the same HMM implementation as for the English POS tagger.

3. Creation of Term Grammars

After POS taggers for each language have been implemented, we implement an automatic method to create term grammars for these languages. To induce a term grammar for a target language T, we use the EuroVoc terms (Steinberger et al., 2012) and project them onto Wikipedia articles in the T-language. The number of EuroVoc terms found in the Wikipedia corpus for each language is shown in Table 2. For each EuroVoc term in language T, we find at most 10 sentences containing an exact match of the term in T-language Wikipedia. We extract such sentences and mark the terms in them with a special tag. An example of this process is shown in English in Table 3. The original version of each such sentence is POS-tagged with a T-language POS tagger. For every word sequence marked as a term, we take its POS tag sequence and include it in our term grammar. For cases in which the same term (appearing in multiple sentences) is tagged using a different POS tag sequence, the majority tag sequence is used to represent the term. Using this approach, we are able to create term grammars for the 21 languages automatically.

Language	EuroVoc terms found
BG	941
CS	1,292
DA	1,462
DE	1,517
EL	920
ES	4,601
ET	1,213
FI	959
FR	4,784
HU	1,286
IT	1,492
LT	1,200
LV	1,119
MT	256
NL	1,567
PL	977
PT	1,579
RO	1,322
SK	986
SL	1,091
SV	1,245

Table 2: Wikipedia dataset

We distinguish between three type of term grammars: *openNLP-auto*, *openNLP-auto-generalTagSet* and *general-auto*. The *openNLP-auto* term grammar is obtained by running the existing OpenNLP POS tagging tools[2] on the Wikipedia articles and recording the POS sequences obtained for the EuroVoc terms as a term grammar, as described above. This results in term grammars such as $\langle NN, NN \rangle$. We also map the specific POS types from the *openNLP-auto* term grammar set to general POS types using the mapping rules described in Petrov et al. (2011). This leads to the second term grammar sets: *openNLP-auto-generalTagSet*. Note that these two term grammar sets cover only the German, Italian, French, Dutch and Spanish languages, since OpenNLP provides POS taggers only for these languages. Finally, we run the General POS Taggers on the Wikipedia sentences as done with the OpenNLP POS taggers and obtain the *general-auto* term grammar rules.

4. Integration

Once we have a POS tagger and a term grammar for a new language T, we integrate them in TWSC (Pinnis et al., 2012), a freely-available term extraction tool for tagging terms in plain-text documents which uses linguistically and statistically motivated term extraction methods. It was originally developed for Latvian and Lithuanian but can be extended to a new language, if provided with a POS tagger and term grammar for the new language. Given a plain-text file, it tags the part-of-speech and identifies possible term candidates using the term grammar. This tool is freely available for download under the Apache 2.0 license.

5. Term Extraction Evaluation

Due to the extensive language coverage of these resources, it was not feasible to create a Gold-Standard dataset (i.e. term tagged corpus) to evaluate the quality of these resources for all 21 languages. Instead, we make use of available resources to conduct this evaluation. Firstly, we perform an automatic recall evaluation using EuroVoc terms. Secondly, we use gold standard data available for three languages in order to perform a precision and recall evaluation.
We create four different settings for TWSC, each representing a different combination of POS tagger and term grammar as described in more detail below:

1. **Setting 1**: In this setting, we make use of the OpenNLP POS Tagger which is available in 5 languages (i.e. DE, ES, FR, IT and NL). We use

[2]https://opennlp.apache.org/

Lang	EuroVoc Term	Sentences
DE	Finanzierungsmittel	Die <TERM>Finanzierungsmittel</TERM>, die um die Bundesmittel aus dem Finanzausgleich ergänzt wurden, stammten aus dem Haushaltsplan des Ministeriums für Ernährung, Landwirtschaft und Forsten.
EN	telecommunications industry	This has led to accusations of the organisation's complicity with the mobile <TERM>telecommunications industry</TERM> in keeping information about mast locations secret.
ES	vida institucional	Fue su hermano el Doctor Benjamín Aceval, distinguida personalidad dentro de la cultura y la <TERM>vida institucional</TERM> de este país.
FR	médecine du travail	Il s'agit de l'atteinte la plus répandue en France en <TERM>médecine du travail</TERM> parmi les troubles musculosquelettiques.
IT	nave cisterna	La SS Marine Sulphur Queen é stata la prima <TERM>nave cisterna</TERM> al mondo per il trasporto dello zolfo liquido.
NL	scheiding der machten	In een presidentieel systeem is aldus de <TERM>scheiding der machten</TERM> sterker dan in een parlementair systeem.

Table 3: Example Wikipedia sentences for EuroVoc terms in various languages

the *openNLP-auto* term grammar which is automatically generated as described in Section 3.

2. **Setting 2**: In this setting, we make use of the General POS Tagger which has been generated for all languages. We use the *openNLP-auto-generalTagSet* term grammar, which is generated by first using the OpenNLP tagger, as in Setting 1, and then mapping the resulting tag sequences to the general tag set. Since these term grammars rely on the availability of the OpenNLP POS-Tagger, they are only available for the 5 languages above.
3. **Setting 3**: In this setting, we make use of General POS Tagger and the *general-auto* term grammar, which has been automatically created using the General POS Tagger. This setting is available for all 21 languages.
4. **Setting 4**: For the last setting, we integrate a language-specific POS-Tagger and a manually generated term grammar (called *manual* below). These resources are very limited and are only available in a small number of languages. In this study, we use the following language specific POS-taggers: TreeTagger (Schmid, 1995), TILDE LV Tagger[3], and HU POS Tagger (Halácsy et al., 2007), for DE, LV and HU, respectively.

5.1. Automatic recall evaluation

In Section 3, we described the process of retrieving Wikipedia sentences that contain EuroVoc terms. Here we report a limited form of recall evaluation of the TWSC term extractor including automatically acquired resources. First, we gathered those Wikipedia documents that contained the sentences with EuroVoc terms that we used to create our term grammars in Section 3 (we do this because the statistical component of TWSC requires full documents to be tagged, not just isolated sentences). Then we used TWSC to extract terms within this dataset using Setting 1 - Setting 3 as described above. We do not include Setting 4 due to its limited availability on all languages.

Recall scores are calculated for each language as shown in Table 4.

Language	Setting		
	1	2	3
DE	0.47	0.26	0.26
ES	0.47	0.40	0.38
FR	0.48	0.46	0.45
IT	0.48	0.46	0.44
NL	0.46	0.45	0.44
BG	–	–	0.37
CS	–	–	0.42
DA	–	–	0.38
EL	–	–	0.38
ET	–	–	0.37
FI	–	–	0.36
HU	–	–	0.34
LT	–	–	0.45
LV	–	–	0.40
MT	–	–	0.50
PL	–	–	0.32
PT	–	–	0.39
RO	–	–	0.41
SK	–	–	0.40
SL	–	–	0.42
SV	–	–	0.37

Table 4: Recall score

The results show that term extractors using a General POS Tagger (Settings 2 and 3) achieve only slightly lower recall scores – with German as an exception – than those using OpenNLP taggers (Setting 1), which

[3]This tagger is included in TWSC (Pinnis et al., 2012).

Lang	Setting	POS-Tagger	Term Grammar	Precision	Recall	F-measure
DE	1	OpenNLP	openNLP-auto	59.63%	22.49%	32.66%
DE	2	GenTagger	openNLP-auto-generalTagSet	**59.92%**	**51.21%**	**55.22%**
DE	3	GenTagger	general-auto	59.68%	**51.21%**	55.12%
DE	4	TreeTagger	manual	50.67%	39.45%	44.36%
HU	3	GenTagger	general-auto	35.37%	**51.53%**	**41.95%**
HU	4	HU Tagger	manual	**35.69%**	30.92%	33.13%
LV	3	GenTagger	general-auto	**54.11%**	25.11%	34.30%
LV	4	TILDE	manual	46.73%	**44.84%**	**45.76%**

Table 5: Performance of POS-Taggers.

rely on a more fine-grained tagset and language-specific manually annotated training data. This is evidence that general pos taggers provide sufficient information for term extraction purposes. We also investigated the difference between creating a term grammar automatically using General POS Tagger (Setting 3) versus using an OpenNLP POS Tagger (Setting 2) and found that the performances are very similar to each other. These findings show that a General POS Tagger and a *general-auto* term grammar can be used effectively in cases where linguistic resources are not available.

We note that recall scores are lower overall than might be expected given that the term grammars are being run on sentences from which they were derived. This is due to several features of the term extractor: (1) TWSC ignores terms of length greater than 4; (2) the statistical filtering component of TWSC rejects term candidates that fail to meet certain criteria, such as term frequency and inverse document frequency thresholds.

5.2. Manual precision/recall evaluation

The automatic recall evaluation shows promising results regarding the use of General POS Taggers and *general-auto* term grammars in extracting EuroVoc terms from a Wikipedia corpus. However, these results do not provide any indication of term extraction accuracy. To evaluate this further, we make use of a Gold-Standard (GS) dataset which was developed within the TaaS project[4] by asking terminologists and translators to identify terms found in a set of documents. This dataset is available in three languages: German (DE), Hungarian (HU) and Latvian (LV). Each language contains a large article (over 2,000 words) in the "automotive" domain whose terms were tagged by two assessors. For each language, an average of 361.67 unique terms of length 1 to 6 were tagged in the documents, which – even though limited in size – provides a Gold-Standard dataset containing a large variety of terms.

[4]http://www.taas-project.eu/

As in the recall evaluation, we used TWSC to identify terms in this dataset using all available settings. We compared the automatically tagged documents and the manually tagged documents (GS) to calculate precision, recall and F-measure as shown in Table 5.

The results show that the automatically generated resources – the generalized POS taggers induced by cross-language projection and the automatically derived term grammars – perform comparably to the currently available POS taggers. For DE, Setting 2 (i.e. GenTagger and *openNLP-auto-generalTagSet* term grammar) results in the highest precision and recall, closely followed by Setting 3. Moreover, results in this language also show that the automatically generated resources (Setting 2 and 3) manage to achieve signifantly higher precision (59.92% and 59.68% respectively) than using a manually created term grammar (Setting 4), which achieves just 50.67%.

The performance for Hungarian term extraction achieves similar results. Using a HU-specific POS Tagger and *manual* term grammar results in marginally better precision than using automatically generated resources (35.69% compared to 35.37%). However, the latter achieves much higher recall.

The performance for Latvian, however, is slightly different. Using Setting 3 (GenTagger and *general-auto* term grammar), TWSC is able to identify terms with higher precision. However, it achieves significantly lower recall than using manually generated resources.

The figures reported in Table 5 are calculated based on exact term matches and upon investigation, we find that many of the non-matching terms are supersets/subsets of each other. An example of these partial matches (in English) are: "multiple injection patterns" (automatically-tagged term) and "multiple injection" (manually-tagged term). These caused the scores of exact terms to be relatively low throughout the different languages. Our analysis identifies that most of these cases also represent terms and if these partial matches are considered to be correct, precision of all systems increases substantially (e.g. in Setting 3 precision scores increase to 66% (HU), 75% (DE) and

90% (LV) if the tagged term is either identical to or a superset or subset of the GS term).

6. Resources for Download

We have released the following resources for download. All data can be downloaded from http://www.taas-project.eu/.

- **POS tagger model**: We provide POS tagger models for each of 21 EU languages for free download. Each POS tagger model uses the same tag set. We will also provide a tool that performs POS tagging using the new models.
- **Term grammar**: We provide term grammars for each EU language for free download.

7. Conclusion

We have described a technique for bootstrapping term extractors for new languages based on inducing POS taggers and term grammars for new languages from existing language resources and have used it to generate term extractors for 21 EU languages. Our evaluation shows that these approaches perform competitively to those using language-specific taggers and manually crafted term grammars. The automatic recall evaluation shows that using GenTagger with a *general-auto* term grammar can achieve similar recall scores to a tagger using a more fine-grained tagset and trained on a language-specific annotated corpus. This is a very promising result, especially for EU languages for which no POS tagger is available. Further evaluation using a small GS dataset in three languages shows the quality of the automatically generated resources. Unfortunately, the limited availability of Gold-Standard data has so far enabled this evaluation to be performed only for a small number of languages and documents.

All resources – POS taggers and term grammars – which have resulted from this study have been made available for download. As future work, we plan to gather more manually created evaluation data from different languages to enable further analysis to be performed.

8. Acknowledgements

The research reported was funded by the TaaS project, European Union Seventh Framework Programme for research, technological development and demonstration under grant agreement no. 296312.

9. References

Aker, A., Paramita, M., and Gaizauskas, R. (2013). Extracting bilingual terminologies from comparable corpora. In *Association for Computational Linguistics*.

Das, D. and Petrov, S. (2011). Unsupervised part-of-speech tagging with bilingual graph-based projections. In *ACL*, pages 600–609.

Giménez, J. and Marquez, L. (2004). Svmtool: A general pos tagger generator based on support vector machines. In *In Proceedings of the 4th International Conference on Language Resources and Evaluation*. Citeseer.

Halácsy, P., Kornai, A., and Oravecz, C. (2007). Hunpos: an open source trigram tagger. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, pages 209–212. Association for Computational Linguistics.

Marcus, M. P., Marcinkiewicz, M. A., and Santorini, B. (1993). Building a large annotated corpus of english: The penn treebank. *Computational linguistics*, 19(2):313–330.

Moore, R. C. (2003). Learning translations of named-entity phrases from parallel corpora. In *Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics-Volume 1*, pages 259–266. Association for Computational Linguistics.

Och, F. J. and Ney, H. (2000). A comparison of alignment models for statistical machine translation. In *Proceedings of the 18th conference on Computational linguistics*, pages 1086–1090, Morristown, NJ, USA. Association for Computational Linguistics.

Och, F. J. O. and Ney, H. (2003). A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

Petrov, S., Das, D., and McDonald, R. (2011). A universal part-of-speech tagset. *arXiv preprint arXiv:1104.2086*.

Pinnis, M., Ljubešić, N., Ştefănescu, D., Skadiņa, I., Tadić, M., and Gornostay, T. (2012). Term extraction, tagging, and mapping tools for under-resourced languages. In *TKE (Terminology and Knowledge Engineering) Conference 2012*, pages 193–208.

Schmid, H. (1995). Treetagger| a language independent part-of-speech tagger. *Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart*, 43.

Sha, F. and Pereira, F. (2003). Shallow parsing with conditional random fields. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, pages 134–141. Association for Computational Linguis-

tics.

Steinberger, R., Eisele, A., Klocek, S., Pilos, S., and Schlüter, P. (2012). DGT-TM: A freely available translation memory in 22 languages. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012)*, Barcelona, Spain, May.

Yarowsky, D., Ngai, G., and Wicentowski, R. (2001). Inducing multilingual text analysis tools via robust projection across aligned corpora. In *Proceedings of the first international conference on Human language technology research*, pages 1–8. Association for Computational Linguistics.