Assigning Terms to Domains by Document Classification

Robert Gaizauskas, Emma Barker, Monica Lestari Paramita and Ahmet Aker
Department of Computer Science, University of Sheffield, United Kingdom
{r.gaizauskas,e.barker,m.paramita,ahmet.aker}@sheffield.ac.uk

Abstract

In this paper we investigate a number of questions relating to the identification of the domain of a term by domain classification of the document in which the term occurs. We propose and evaluate a straightforward method for domain classification of documents in 24 languages that exploits a multilingual thesaurus and Wikipedia. We investigate and provide quantitative results about the extent to which humans agree about the domain classification of documents and terms also the extent to which terms are likely to "inherit" the domain of their parent document.

1 Introduction

In an increasingly interconnected world, characterised by high international mobility and globalised trade patterns, communication across languages is ever more important. The demand for translation services has never been higher and there is constant pressure for technological solutions, e.g., in the form of machine translation (MT) and computer-assisted translation (CAT), to increase translation throughput and lower costs. One requirement of these technologies is bilingual lexical resources, i.e. dictionaries, particularly in specialist subject areas or domains, such as biomedicine, information techology, or aerospace. While in theory statistical MT approaches need only parallel corpora to train their translation models, there is never enough parallel material in technical areas or for minority languages to support high quality technical translation, so specialist bilingual terminological resources are very important. Similarly, human translators using CAT systems need support in the form of bilingual terminological resources in specialist areas about which they may know very little.

The EU FP-7 TaaS project has created a cloud-based terminological service, which makes available bilingual terminological resources for all EU languages. These resources include both existing terminological resources and resources derived automatically from parallel and comparable corpora available on the web. Additionally, the service's user community is able manually to supplement or correct these resources. Like many other terminology resources (e.g. IATE[1], Eurotermbank[2]), terms in TaaS have *domains* associated with them. This is done for a number of reasons: (1) *Computational Feasiblity*: While in theory a translator faced with a translation task could provide the set of documents to be translated to a system that dynamically assembled a bespoke terminological resource specific to this task, this is not computationally feasible, at least not in a time-frame a user is likely to accept. Much more feasible is to collect bilingual terminology off-line and store it within a term repository with an associated domain or domains. Then, an on-line user, having identified the domain of the document(s) to be translated, searches for terms within that domain or may have terms from the domain into which his documents are automatically classified made available to him. (2) *Sense Disambiguation*: Term expressions, or their translations, may have multiple senses, but these are likely to be in different domains. By restricting the domain when looking up terms, sense confusions are less likely to occur. (3) *User Preference*: Our

[1] http://iate.europa.eu

[2] http://www.eurotermbank.com

Proceedings of the 4th International Workshop on Computational Terminology, pages 11–21,
Dublin, Ireland, August 23 2014.

discussions with technical translators show they are used to and comfortable with the notion of domains and prefer terminological resources structured by domain.

Assuming, therefore, that term resources are to be structured into domains, the question arises as to how this is to be done automatically for automatically acquired terms. While the notion of domain is inherent in most definitions of "term"[3], most term extraction systems identify terms using grammatical patterns and/or statistical occurrence information applied to and gathered from corpora deemed to be either in-domain or general/multi-domain. I.e. such tools do not have any inherent notion of domain, but instead rely on the external provision of documents pre-selected by domain to determine the domain of the extracted terms. But how valid is this procedure?

In this paper we explore several questions related to the assignment of terms to domains. These questions were addressed within the evaluation of that component of the TaaS platform which automatically creates bilingual term resources (the Bilingual Term Extraction System, aka BiTES). Specifically:

1. How well can a simple vector space classifier built from a multilingual thesaurus automatically classify documents into domains prior to assigning these domains to the terms within the documents?

2. To what extent do humans agree about the assignment of terms to domains?

3. How accurate is the assumption that terms can be assigned to the domains of the documents in which they are found?

The rest of the paper is structured as follows. Section 2 gives a brief overview of the BiTES system as a whole and the domain classification component in somewhat more detail. In section 3 we describe the evaluation of those parts of BiTES relevant to the questions above, detailing the evaluation tasks, participants and data used and as well as the results of the evaluation. Section 4 provides analysis and discussion of results. Section 5 discusses related work. We conclude in Section 6.

2 System Components

2.1 BiTES overview

The Bilingual Term Extraction System (BiTES) uses different workflows, each comprising a set of tools run in sequence, to collect bilingual term pairs. Each new bilingual term pair found by BiTES is fed into a database for later retrieval. The workflows consist of four different types of tools:

1. tools for collecting Web resources, such as parallel and comparable corpora from which the bilingual terms are extracted;

2. tools for performing document classification into pre-defined categories or domains;

3. tools for extracting terms from or tagging terms in monolingual documents collected from the Web;

4. tools for bilingual alignment of tagged terms in parallel or comparable document pairs collected from the Web.

Each workflow can be run in an offline and periodic manner and starts with document collection from the Web followed by document classification. The output of the document classifier is passed to the monolingual term extractor. Term-tagged document pairs are fed to the bilingual term alignment processor to extract bilingual terms. The main goal of BiTES within the TaaS platform is to automatically collect large numbers of bilingual term pairs off-line that are then stored in a database for later retrieval by users. This database of automatically collected terms is consulted when other pre-existing, and presumed higher quality, manually gathered terminological resources, such as, EuroTermBank or IATE, which are also available in the TaaS platform, do not contain translations for terms the user seeks.

[3]For example Bessé et al. (1997) define term as "a lexical unit consisting of one or more than one word which represents a concept inside a domain"; ISO 1087-1:2000 defines term as "verbal designation of a general concept in a specific subject field".

In this section we detail only the domain classification component of BiTES as it is the component that has the most direct implications for the research questions addressed in the paper and as the underlying methods and performance of the other tools used in BiTES have been reported elsewhere (Aker et al., 2012; Pinnis et al., 2012; Su and Babych, 2012; Skadiņa et al., 2012; Aker et al., 2013; Aker et al., 2014b; Aker et al., 2014a).

2.2 Domain Classification

2.2.1 Domain classification scheme

Despite the existence of various domain classification schemes, the TaaS project has created its own domain classification for several reasons. First, the TaaS platform requires a suitable classification system which is easy to use, yet provides broad coverage of the topics that are of greatest interest to users working in terminology management and machine translation. The project conducted a user study to identify the set of required domains. Various classification systems were considered, including the Dewey Decimal Classification (DDC) and Universal Decimal Classification (UDC). These schemes, however, are too complicated to be used by terminologists (the latter uses 10 level-1 domains and more than 60,000 level-2 domains) yet still did not sufficiently cover relevant subject fields identified by our users, such as IT, medicine and mechanical engineering. The Internal Classification for Standards (ICS) scheme was considered next, as it covers technical subject fields, but it was lacking with respect to legal and humanities domains. Intially, therefore, the TaaS project decided to adopt the domain structuring used in the EuroVoc thesaurus, which includes a broad range of domains. However, with 21 level-1 domains and 127 level-2 domains, it too is quite complex and focuses more on European Union domains than the industry-related domains identified in our user study. Therefore, various modifications to the EuroVoc domain scheme were performed to merge and delete various domains so as to increase the scheme's suitability for the project and also improve its practicality and ease of use. This resulted in what we here refer to as the TaaS domain classification scheme, which contains 11 level-1 domains and 66 level-2 domains[4]. A mapping from EuroVoc level-1 and -2 domains to TaaS level-1 and -2 domains was manually established.

2.2.2 Document classifier

Many approaches to document classification have been proposed in the literature – see Agarwal et al. (2014) for a survey. Our domain classifier uses the well-explored vector space approach. For each language, each domain is represented by one vector and each document to be classified by another vector. The cosine similarity measure (Salton and Lesk, 1968) is calculated between the vector representation of the input document and the vector representation of a domain and serves as a measure of the extent to which the document belongs to that domain. The highest scoring domain may be chosen if hard classification is required, or a vector of scores, one per domain, may be returned, if soft classification is needed. The advantage of this approach in our setting is that we can exploit an existing multilingual, domain-structured thesaurus to build our domain vector to deliver domain classifiers for 11 domains in 24 languages, without the need for collecting training data.

To create a vector representation for an input document, the document is first pre-processed and stop words and punctuation are removed from it. The TaaS project covers 23 of the 24 official EU languages[5] as well as Russian. For each of these languages we took the entire dump of Wikipedia and weighted each word in the articles using $tf * idf$ (Manning et al., 2008). Any word whose idf is below a predefined threshold is used as a stop word. Using this method we collected stop word lists for all 24 languages. To identify punctuation we used simple rules covering the major punctuation symbols. After filtering out stop words and punctuation, the remaining words in the input document are stemmed. We adopted Lucene stemmers for all languages for which these resources are available in and implemented new stemmers for Latvian, Lithuanian and Estonian. Finally, term frequency counts for the stems in the input document are gathered, idf scores are taken from the Wikipedia dump and $tf * idf$ weights are computed and stored to create the vector representation of the input document.

[4] A full specification of the scheme is available at: `https://demo.taas-project.eu/domains`.

[5] The omitted language is Irish, for which insufficient data was available for training our tools.

To create domain vectors we did the following: (1) For each domain and language, we manually downloaded the relevant EuroVoc term file from the EuroVoc website[6]. (2) We used the EuroVoc-to-TaaS mapping described in Section 2.2.1 above to map all terms belonging to a specific EuroVoc domain (level-1 or -2) to the corresponding TaaS domain (level-1 or -2). (3) For each TaaS domain (in each language) we built a domain-specific vector from the set of newly derived TaaS terms in the domain. Since our vector elements correspond to single words, we convert any multi-word term in the domain into multiple single word representations. To do this we process each multi-word by splitting it on whitespace, removing any words that are stop words and finally stemming the remaining words. For any single word terms we simply take their stems. Finally, all the word stems so derived are stored in a vector. We use simple term frequency, measured across the bag of stemmed words derived from all terms in the domain, as a weight for each stem. In the experiment below we report results only for classification into the 11 level-1 TaaS domains – see Table1.

Level-1 Domain	**Level-2 Domain**
Agriculture and foodstuff	Agriculture, forestry, fisheries, foodstuff, beverages and tobacco, and food technology.
Arts	Plastic arts, music, literature, and dance.
Economics	Business administration, national economics, finance and accounting, trade, marketing and public relations, and insurance.
Energy	Energy policy, coal and mining, oil and gas, nuclear energy, and wind, water and solar energy.
Environment	Climate, and environmental protection.
Industries and technology	Information and communication technology, chemical industry, iron, steel and other metal industries, mechanical engineering, electronics and electrical engineering, building and public works, wood industry, leather and textile industries, transportation and aeronautics, and tourism.
Law	Civil law, criminal law, commercial law, public law, and international law and human rights.
Medicine and pharmacy	Anatomy, ophthalmology, dentistry, otolaryngology, paediatrics, surgery, alternative treatment methods, gynaecology, veterinary medicine, pharmacy, cosmetic, and medical engineering.
Natural sciences	Astronomy, biology, chemistry, geology, geography, mathematics and physics.
Politics and administration	Administration, politics, international relations and defence, and European Union.
Social sciences	Education, history, communication and media, social affairs, culture and religion, linguistics, and sports.

Table 1: TaaS Domains

3 Evaluation

To evaluate the BiTES system we devised a set of four human assessment tasks focussed on different aspects of the system. These tasks were designed to assess the domain classifier, the extent to which terms found in a document judged to be in a given domain were in the domain of their document, the accuracy of the boundaries of extracted terms in context and the accuracy of system proposed bilingual term alignments. In this paper we focus on the first two of these tasks only. As noted above the TaaS project addressed 24 languages in total. Evaluation of all these languages and language pairs was clearly impossible. We chose to focus on six languages – English (EN), German (DE), Spanish (ES), Czech (CS), Lithuanian (LT) and Latvian (LV) – and five language pairs EN-DE, EN-ES, EN-CS, EN-LT and EN-LV. This gave us exemplars from the Germanic, Romance, Slavic and Baltic language groups.

3.1 Human assessment tasks

3.1.1 Domain classification assessment

In the domain classification assessment task we present participants with a document and the TaaS set of domain classes (see Table 1), and ask them to select the TaaS level-1 domain that in their judgement best represents the document. We provide a brief set of guidelines to help them carry out this task.

[6] `http://eurovoc.europa.eu`

We encourage participants to select a primary domain wherever possible – i.e. a single domain that best represents the document. But we allow them to select multiple domains from the list provided, if they believe the text spans more than one domain and they are unable to decide upon a primary domain. If they do opt to select multiple domains we ask them to keep the number of selected domains to a minimum. For example, the Wikipedia article entitled "Hydraulic Fracturing" [7] discusses a wide range of topics, including the process of hydraulic fracturing and its impacts in the geological, environmental, economic and political spheres. For this document, which we use in our guidelines for the task, we recommend assessors choose "Energy" as a primary domain and possibly also "Industries and Technology", since these two domains best represent the overall document content, which is chiefly concerned with what is described as a "mechanical" process in the "industrial sector of mining", the products being natural gas and oil. But we would limit our selection to these two.

The aim is for participants to select domains from the list we provide. However, in the event that they are unable to do so, we provide an option "none of the above", which they may select and then provide a domain of their own. In the guidelines we ask them to spend some time reviewing potential domain candidates, and combinations of candidates, before opting to provide an as yet unspecified domain. I.e. they should only select the option "none of the above" if they have genuinely exhausted all the possibilities using one or more domains from our list.

3.1.2 Term in domain assessment

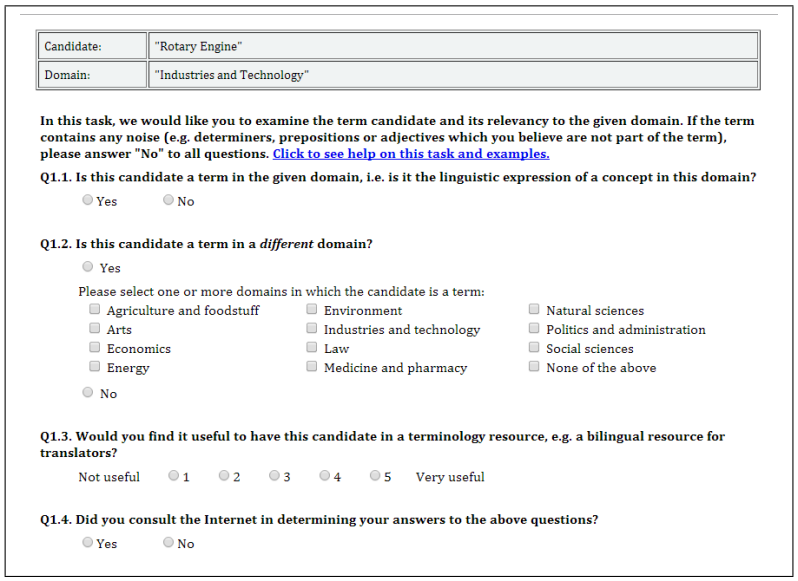

Figure 1: Judging a Term Candidate in a Domain

This is the first of two tasks assessing the (monolingual) extraction of terms. It assesses whether an automatically extracted term candidate is a term in a proposed, automatically determined, domain. Assuming the candidate is a term, a subsequent task assesses whether the boundaries of the term candidate, when taken in their original document context, are correct.

In this task (see Figure 1) we present assessors with a term candidate and a domain and then ask them to judge if the candidate is a term in the given domain or if it is a term in a different domain. If they judge the term to be in a different domain we ask them to specify the alternate domain(s). In this question the candidate and the domain category are assessed together but we do not provide any specific context, such as the source sentence or source document. As with the previous task we provide a brief set of guidelines to help assessors carry out the task.

We ask assessors to base their judgement on the entire candidate string. If the string contains a term but also contains, additional words that are not part of the term then they should answer "no". For

[7] Aka "fracking", see `http://en.wikipedia.org/wiki/Hydraulic_fracturing`

example, consider the candidate "excessive fuel emissions" and the domain "Industries and Technology". Although most people would agree that "fuel emissions" is a term, Q1.1 and Q1.2 should be answered "no" in this case since the candidate also contains noise, i.e. the word "excessive". Superfluous articles, determiners and other closed class words are also considered "noise" in this context.

We encourage assessors to search the Internet, as translators and terminologists might do, to help determine whether the entire candidate is indeed a term in the given domain. Web searches can provide examples of real world uses of a candidate in different domains. We also allow assessors to consult existing terminological or dictionary resources, online or otherwise, during the evaluation task. However, participants are encouraged not to assume that such resources are complete or entirely correct and advised that such resources be used with some consideration and caution.

Finally, if assessors have answered "yes" to one of Q1.1 or Q1.2, they will also be asked to indicate the utility of the term candidate in Q1.3, however this aspect of the assessment is not of interest here and will not be discussed further.

3.2 Participants

We recruited experienced translators to participate in the evaluation tasks. For English and for each language pair, three assessors carried out each of the evaluation tasks. In total our study involved 17 assessors – one assessor took part in DE only, EN-DE and EN only tasks. All assessors had an excellent background in translation in a wide variety of domains, with an average of 8.5 years translation experience in the relevant language pairs. All assessors who evaluated the English, Lithuanian and Latvian data were native speakers. For each of the remaining languages (Czech, German and Spanish), 2 were native speakers whilst 1 was a fluent speaker with over 54 years, 15 years and 12 years experience (respectively) in using these languages as a second language.

3.3 Data

3.3.1 Domain classification

For the domain classification task, we selected a set of documents to be evaluated using the following approach. First, we gathered all articles from the August 2013 Wikipedia dump in each of the assesment languages and extracted the main text paragraphs, i.e. tables, images, infoboxes and URLs were filtered out. The number of articles ranged from 50,000 (for Latvian) to 4,000,000 (for English). We then ran our domain classifier over each document in this dataset and assigned to each document the top domain proposed by the classifier, i.e. the domain with the highest score according to our vector space approach (Section 2.2.2). During processing we filtered out documents whose top domain scores were below a previously set minimum threshold and those whose document length was below a minimum length. Finally, for each domain D, we sorted the documents classified into D based on their scores, divided this sequence into 10 equal-size bins and selected one document from each bin. Since we were classifying documents into one of the 11 level-1 TaaS domains, this resulted in 110 documents for each language[8].

3.3.2 Term extraction

For the term in domain assessment task, we narrowed the task to focus on two domains only – "Industries and Technology" and "Politics and Administration" – since we could not hope to assess sufficient terms in all domains in all languages. We extracted terms from all documents contained in the top bin of the domain classifier, i.e. the 10% of documents in the domain with the highest similarity score to the domain vector, using TWSC as the term extractor tool (Pinnis et al., 2012). Next, we selected 200 terms from both domains, choosing terms of different word lengths: 50 of length 1, 70 of length 2, 50 of length 3 and 30 of length 4. This distribution was chosen in order to approximate roughly the distribution of term lengths one might expect in the data[9]. This process was repeated for each of our six languages.

[8]The Latvian set contains a slightly smaller set (i.e. 106 documents) due to a fewer number of documents found in one of the domains (i.e. 6 documents in the "Energy" domains).

[9]This distribution was chosen after analysing term lengths in the EuroVoc thesaurus and in the term extractor results, which indicated that terms length 2 are the most common, followed by terms length 1 and 3, and terms length 4 are found to be the least common. We boosted slightly the numbers of length 4 terms in our test to try to eliminate very small number effects.

3.4 Results

3.4.1 Domain classification assessment

A total of 656 documents (in 6 languages) were assessed and on average 1.2 domains were selected for each document. Regarding human-human agreement, at least 2 assessors fully agreed on their domain selections (including cases where more than one domain was selected) on 78% of the cases. When considering cases where at least 2 assessors agreed on at least one domain, agreement increases to 98%.

Regarding human-system agreement, since 3 assessors participated in each assessment, we produced two types of human judgments: *majority* (i.e. any domains selected by at least two assessors) and *union* (i.e. any domains selected by at least one assessor). We computed the agreements between the classifier and both the majority and the union human judgments. Results averaged over all domains and languages show the system's proposed top domain agreed with the majority human judgment in 45% of cases and with the union of human judgments in 58% of cases. Broken down by language, agreement with the majority judgment ranged from a low of 35% (EN) to a high of over 53% (DE) while agreement with the union of judgments ranged from a low of 48% (EN) to a high of over 64% (CS). By domain, agreement with majority judgment ranged from just over 12% (Agriculture and foodstuff) to 88% (Medicine and pharmacy) while agreement with the union of judgments ranged from 23% (Agriculture and foodstuff) to over 91% (Social sciences).

Recall (Section 3.3.1) that our test data includes documents from different similarity score bins. This enables us to analyse the agreement between the assessors and the classifier in more detail. In general we see a monotonically increasing agreement with both the majority judgement and union of judgments as we move from the lowest to highest scoring bin. The highest agreement is achieved in bin 10 which represents the 10% of documents "most confidently" classified to a given domain, i.e. those documents with the highest similarity score to the domain vector. Just under 80% of these documents (77.27%) are included in the union of assessors data and 63% are included in the majority. I.e. for approximately 77% of the documents most confidently classified to a domain by our classifier, at least one in three humans will agree with the domain classification and for about 63% the majority of humans will agree.

3.4.2 Term in domain assessment

Term length	Total	Term in the given domain	Term in a different domain
All length	457	88%	12%
1	144	88%	12%
2	182	87%	13%
3	84	92%	8%
4	47	91%	9%

Table 2: Terms with different term length

Languages	Total	Term in the given domain	Term in a different domain
All languages	457	88%	12%
CS	103	86%	14%
DE	79	82%	18%
EN	80	88%	13%
ES	54	80%	20%
LT	47	98%	2%
LV	94	97%	3%

Table 3: Terms of different languages

A total of 1,200 candidate terms in 6 languages were assessed by 3 assessors and the majority judgments (i.e. cases where at least two assessors agree) show that 38% terms were assessed to be candidate terms in the given domain, 5% terms were assessed to be candidate terms in a different domain, and the rest (57%) were deemed not to be terms.

This indicates that out of all candidate terms which were identified to be correct terms (43% of the data), 88% were assessed to be in the same domain as the documents they were extracted from. Further analysis showed that the 57% of candidates judged not to be terms could be further broken down into 33% which contain an overlap with a term, i.e. term boundaries were incorrectly identified, and 24% which neither are nor overlap with a term.

Of the 43% candidate terms that were judged to be terms, we examined the variation in extent to which they were judged to be terms in the given domain across term lengths and across languages. These figures are shown in Tables 2 and 3. We also examined variation in the extent to which these terms were judged to be terms in the given domain across the two domains we were investigating: in "Industries and

Technology" 92% of the terms were judged to be in the given domain and 8% in another domain, while for "Politics and Administration" these figures were 85% and 15% respectively.

For the 43% of the term candidates that were identified as correct terms (457 terms), all three assessors agreed about the domain of the term, i.e. they either accepted the domain proposed by the system for the term or they agreed on an alternative(s), in 45% of the cases. In 54% of the cases there was not universal agreement but at least two assessors agreed on at least one domain they assigned to the term. Only in 1% of the cases was there no overlap in judgment about term domain.

4 Analysis and Discussion

Let us now return to the research questions we raised in Section 1. Our first question was: *How well can a simple vector space classifier built from a multilingual thesaurus automatically classify documents into domains prior to assigning these domains to the terms within the documents?* First, we have to view system performance in the context of human performance. Results in the last section show that 2 out of 3 humans agree 78% of the time on exact assignment of (possibly multiple) domains to documents and 98% of the time if only one of the domains they assign to a document need to match. Over all languages and domains our classifier achieves only 45% agreement with the majority judgment and 58% with the union of judgments. However, if we restrict ourselves to the highest confidence domain assigments, then the picture is much better: 63% agreement with the majority judgment and 77% with the union of judgments. This restriction reduces the number of documents from which terms could be mined from if accurate domain classification is important – but so long as there are lots of documents to mine terms from this may not be important. Furthermore note that our classifier could easily be used to select multiple domains, perhaps, e.g., when the differences in scores between highest scoring domains is small. This would make the comparison with the human figures fairer (now the system can only propose one domain per document while the humans can propose several) and could only result in higher system figures relative to human ones. We conclude that the vector space classifier utilizing domain representations derived from a pre-existing multingual thesaurus has much to recommend: it is simple, it needs no training data, it is straightforwardly applicable to multiple (24 in our case) different languages and its performance is adequate if it is suitably constrained.

Our second question was: *To what extent do humans agree about the assignment of terms to domains?* Our results show that in less than half the cases do all three human assessors agree with the assignment of a term to a particular domain. However, in 99% of the cases at least two of three assessors concur on at least one domain to which the term belongs. This suggests that using overlap with two of three human assessors is a good approach to measuring automatic domain assignment to terms.

Our third question was: *How accurate is the assumption that terms can be assigned to the domains of the documents in which they are found?* Tables 2 and 3 show that on average 88% of terms are judged to be in the domain of the document in which they are found. Furthermore there is relatively little variation in this figure – it ranges from a low of 80% (ES) to a high of 98% (LT) and a low of 87% for terms of length 2 to a high of 92% for terms of length 3. This suggests that assigning domains to terms based on the domain of the document the term is found in is a relatively safe thing to do, but is by no means perfect: just over 10% of terms will have their domains incorrectly assigned by making this assumption.

5 Related Work

There has been extensive work on the development of automated techniques to extract terminology from document collections. Such term extraction approaches can be grouped into three categories based on the information used to extract terms: approaches using purely linguistic information, approaches using purely statistical information and those using combinations of both. An analysis of different approaches is given by Pazienza et al. (2005). For the most part, however, such approaches make the assumption that domain-specific, and perhaps also non-domain-specific, collections of texts are available. Justeson and Katz (1995), for example, assume that term frequency of a limited sort of noun phrases in domain-specific texts is sufficicent to indicate termhood. Others such as Chung (2003) and Drouin (2004) look at statistical contrasts between domain-specific and general comparison or reference corpus. See also

(Kim et al., 2009; Marciniak and Mykowiecka, 2013; Kilgariff, 2014). By contrast our approach does not presuppose the existence of documents pre-classified by domain (though we could benefit from this). Rather our approach starts by classifying a document into a domain and then extracting terms from it and assigning them the domain of the document.

Utsuro et al. (2006) and Kida et al. (2007) extract terms from web-documents. The domain specification of a term is determined in two stage approach. In the first stage for a term under inspection web-documents which mention the term are collected. Then these documents are divided into two sets: domain relevant and domain-irrelevant documents. A document whose content similarity to a domain specific corpora is above a predefined threshold is regarded as relevant. Any other document is regarded as irrelevant. In the second stage a ratio of times the term occurs in the relevant and the irrelevant set is computed. This ratio is used to determine whether the extracted term belongs to the domain in hand or not. Again, a domain-specific corpus is assumed for this approach to proceed.

Benedictis et al. (2013) use bootstrapping to collect domain specific terms. They start with some manually selected domain specific seed terms, perform web-search to obtain documents, extract further terms and re-start the process with the new terms. The documents returned by the search engine are assumed to belong to the domain in hand and so are the extracted terms. By contrast our approach does not require manually selected terms, but instead uses an existing domain structured multingual thesaurus.

6 Conclusion

In this paper we have investigated a number of questions relating to the identification of the domain of a term by domain classification of the document in which the term occurs. We proposed and evaluated a straightforward method for domain classification of documents in 24 languages which uses a multilingual thesaurus to construct "domain vectors". We investigated the extent to which humans agree about the domain classification of documents and terms. And, we investigated the extent to which terms are likely to "inherit" the domain of their parent document. Our results show that the domain classification method has significant merit, that humans generally, but by no means universally, agree about domain classification of documents and terms, and again that terms are generally, but certainly not universally, likely to be of the same domain as the document in which they occur.

7 Acknowledgments

The authors would like to acknowledge funding from the European Union FP-7 programme for the TaaS project, grant number: 296312. We would also like to thank the human assessors without whose careful work the results reported here would not have been obtained. Finally we thank our project partners in the TaaS project for user studies with translators and terminologists, contributions to the TaaS system, and development of the TaaS domain classication scheme and the EuroVoc-to-TaaS mapping.

References

Basant Agarwal and Namita Mittal. 2014. Text classification using machine learning methods-a survey. In *Proceedings of the Second International Conference on Soft Computing for Problem Solving (SocProS 2012), December 28-30, 2012*, pages 701–709. Springer.

Ahmet Aker, Evangelos Kanoulas, and Robert J Gaizauskas. 2012. A light way to collect comparable corpora from the web. In *Proceedings of Eighth International Conference on Language Resources and Evalution (LREC)*, pages 15–20.

Ahmet Aker, Monica Paramita, and Robert Gaizauskas. 2013. Extracting bilingual terminologies from comparable corpora. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL 2013)*.

Ahmet Aker, Monica Paramita, Emma Barker, and Robert Gaizauskas. 2014a. Bootstraping Term Extractors for Multiple Languages. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*.

Ahmet Aker, Monica Paramita, Mārcis Pinnis, and Robert Gaizauskas. 2014b. Bilingual dictionaries for all EU languages. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*.

Teresa Mihwa Chung. 2003. A corpus comparison approach for terminology extraction. *Terminology*, 9(2).

Flavio De Benedictis, Stefano Faralli, Roberto Navigli, et al. 2013. Glossboot: Bootstrapping multilingual domain glossaries from the web. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 528–538.

Bruno de Bessé, Blaise Nkwenti-Azeh, and Juan C. Sager. 1997. Glossary of terms used in terminology. *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication*, 4:117–156(39).

Patrick Drouin. 2004. Detection of domain specific terminology using corpora comparison. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC2004)*.

John S. Justeson and Slava M. Katz. 1995. Technical terminology: Some linguistic properties and an algorithm for identification in text. *Natural Language Engineering*, 1(1):9–27.

Mitsuhiro Kida, Masatsugu Tonoike, Takehito Utsuro, and Satoshi Sato. 2007. Domain classification of technical terms using the web. *Systems and Computers in Japan*, 38(14):11–19.

Adam Kilgariff. 2014. Finding terms in corpora for many languages with the Sketch Engine. *14th Conference of the European Chapter of the Association for Computational Linguistics*.

Su Nam Kim, Timothy Baldwin, and Min-Yen Kan. 2009. An unsupervised approach to domain-specific term extraction. In *Australasian Language Technology Association Workshop 2009*, page 94.

Christopher D Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to information retrieval*, volume 1. Cambridge university press Cambridge.

Małgorzata Marciniak and Agnieszka Mykowiecka. 2013. Terminology extraction from domain texts in polish. In *Intelligent Tools for Building a Scientific Information Platform*, pages 171–185. Springer.

Maria Teresa Pazienza, Marco Pennacchiotti, and Fabio Massimo Zanzotto. 2005. Terminology extraction: an analysis of linguistic and statistical approaches. In *Knowledge Mining*, pages 255–279. Springer.

Mārcis Pinnis, Nikola Ljubešić, Dan Ştefănescu, Inguna Skadiņa, Marko Tadić, and Tatiana Gornostay. 2012. Term extraction, tagging, and mapping tools for under-resourced languages. In *Proceedings of the 10th Conference on Terminology and Knowledge Engineering (TKE 2012), June*, pages 20–21.

Gerard Salton and Michael E Lesk. 1968. Computer evaluation of indexing and text processing. *Journal of the ACM (JACM)*, 15(1):8–36.

Inguna Skadiņa, Ahmet Aker, Nikos Mastropavlos, Fangzhong Su, Dan Tufis, Mateja Verlic, Andrejs Vasiļjevs, Bogdan Babych, Monica Paramita, Paul Clough, Robert Gaizauskas, and Nikos Glaros. 2012. Collecting and using comparable corpora for statistical machine translation. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC), Istanbul, Turkey*.

Fangzhong Su and Bogdan Babych. 2012. Measuring comparability of documents in non-parallel corpora for efficient extraction of (semi-) parallel translation equivalents. In *Proceedings of the Joint Workshop on Exploiting Synergies between Information Retrieval and Machine Translation (ESIRMT) and Hybrid Approaches to Machine Translation (HyTra)*, pages 10–19. Association for Computational Linguistics.

Takehito Utsuro, Mitsuhiro Kida, Masatsugu Tonoike, and Satoshi Sato. 2006. Collecting novel technical terms from the web by estimating domain specificity of a term. In *Computer Processing of Oriental Languages. Beyond the Orient: The Research Challenges Ahead*, pages 173–180. Springer.