Extracting Bilingual Terms from the Web

Robert Gaizauskas, Monica Lestari Paramita, Emma Barker, Mārcis Pinnis, Ahmet Aker and Marta Pahisa Solé

In this paper we make two contributions. First, we describe a multi-component system called BiTES (Bilingual Term Extraction System) designed to automatically gather domain-specific bilingual term pairs from Web data. BiTES components consist of data gathering tools, domain classifiers, monolingual text extraction systems and bilingual term aligners. BiTES is readily extendable to new language pairs and has been successfully used to gather bilingual terminology for 24 language pairs, including English and all official EU languages, save Irish. Second, we describe a novel set of methods for evaluating the main components of BiTES and present the results of our evaluation for six language pairs. Results show that the BiTES approach can be used to successfully harvest quality bilingual term pairs from the Web. Our evaluation method delivers significant insights about the strengths and weaknesses of our techniques. It can be straightforwardly reused to evaluate other bilingual term extraction systems and makes a novel contribution to the study of how to evaluate bilingual terminology extraction systems.

Keywords: comparable corpora, domain classification, term extraction, cross-language term alignment, machine translation, evaluation of term extraction

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 and terminological resources, particularly in specialist subject areas or domains, such as biomedicine, information technology, 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. Consequently, 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[1] has created a cloud-based terminological service that makes available bilingual terminological resources for all EU languages. These resources include both existing terminological resources and resources harvested 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 in order to enhance the quality and coverage of the term resources available on the platform. An overview of the TaaS system, including a description of how automatically harvested bilingual terms are exploited within it, is presented in Gornostay and Vasiļjevs (2014). However, in this paper we focus solely on the TaaS approach to automatic extraction of bilingual terminology from the Web. Specifically we do two things. First, we describe the novel Bilingual Terminology Extraction System (BiTES) developed in TaaS, which has enabled us to gather bilingual terminological resources for 24 language pairs. BiTES's principal strengths are the ease with which new language pairs may be incorporated within it and its component architecture that allows individual components to be replaced with more specialised or improved components as they become available without requiring their availability from the outset. For example, BiTES generalised approach to part-of-speech (POS) tagging and term grammar acquisition means there is no need to develop bespoke part-of-speech taggers and term grammars for each language, though these can be taken

[1] Information about the project can be found here: www.taas-project.eu

advantage of if they exist. Second, we describe the comprehensive methodology we developed to evaluate each of the components of BiTES and the insights gained from an evaluation across six languages. This methodology can be straightforwardly reused to evaluate other bilingual term extraction systems and makes a novel contribution to the study of how to evaluate several components of bilingual terminology extraction systems, including domain classification, term boundary determination and bilingual term alignment.

2. System Components

The main function 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.

As shown in Figure 1, 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 the TaaS term 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.

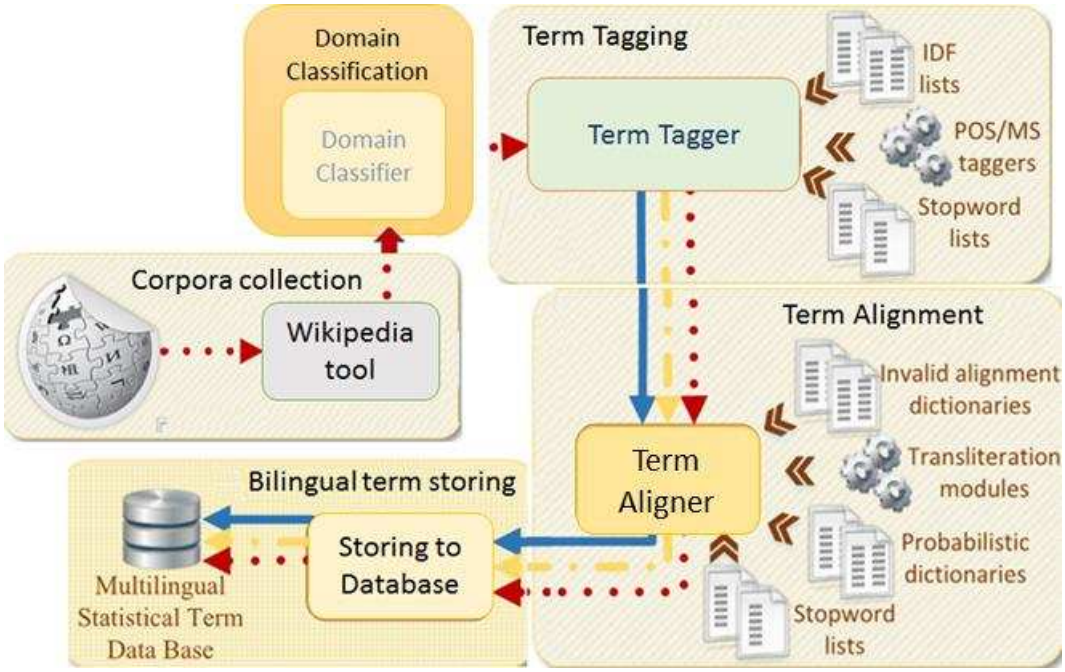


Figure 1: BiTES overview

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. In the following sub-sections we detail these components. BiTES successfully extracts bilingual term pairs for 24 language pairs – English plus X for all official EU languages X, except Irish – too few web texts available at present, and with the addition of Russian. We refer to these 25 languages as the TaaS languages.

2.1. Collecting comparable corpora

Of the tools used for collecting Web resources, we concentrate here only on the tool for gathering comparable corpora from Wikipedia (freely available at www.taas-project.eu). Three other corpus collection tools were developed to collect parallel corpora from the Web, crawl RSS news feeds in multiple languages, and gather comparable document pairs from arbitrary web sources given a set of seed terms. Space precludes discussing each of these tools, though

we compare term extraction using them in section 4. In any case, Wikipedia proved the best source of terms, both in breadth and quality.

Wikipedia contains a large number of documents on various topics and in different languages. When two articles in different languages are on the same topic, they are connected by inter-language links, enabling a comparable corpus to be extracted that is already aligned at the document level. Using the Wikipedia comparable corpus collection tool to exploit these links, we created twenty-four Wikipedia comparable corpora, pairing English with each of the other TaaS languages.

When run for the first time for a given language pair, the comparable corpus collection tool downloads the latest monolingual Wikipedia dumps for the specified languages[2] and extracts plain-text versions of the articles for both languages, deleting infoboxes, images, tables and URLs. The tool also downloads the Wikipedia inter-language links file and uses it to identify linked document pairs (Paramita et al. 2012). Once the comparable corpus is ready it is passed to the next tool within the workflow – the document classifier.

2.2. Domain classification

Like many other terminology resources (e.g. IATE 2014, EuroTermBank 2015), bilingual terms in the TaaS repository have domains associated with them. This is done for several reasons: (1) Computational Feasiblity: While in theory a bespoke terminological resource specific to a particular translation task could be dynamically assembled from a user-supplied set of documents to be translated, this is not computationally feasible, at least not in an acceptable time-frame. Much more feasible is to collect bilingual terminology off-line and store it within a term repository with an associated domain or domains. Then, a user, having identified the domain of the document(s) to be translated, searches for terms within that domain or has terms from the domain into which his documents are automatically classified made available to him.

---

[2] Available from http://dumps.wikimedia.org/

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

In BiTES, therefore, terms are assigned to one or more domains. This is done by assuming terms 'inherit' the domain of the document in which they are found and using a document-level domain classifier (described below) to assign domains to documents. This validity of this assumption is discussed in detail in Gaizauskas et al. (2014) and some of the results of that study are summarised below.

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 that 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 and Universal Decimal Classification. 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 scheme was considered next, as it covers technical subject fields, but it was lacking with respect to legal and humanities domains. Initially, therefore, the TaaS project decided to adopt the domain structuring used in the EuroVoc thesaurus (Steinberger et al. 2002), which includes a broad range of domains (21 level-1 and 127 level-2 domains). However, it focuses more on EU-related domains than the industry-related domains identified in our user study. Therefore, various modifications to the

EuroVoc domain scheme were made to increase the scheme's suitability for the project. 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 (Table 1). A mapping from EuroVoc level-1 and -2 domains to TaaS level-1 and level-2 domains was manually created.

2.2.2. Document classifier

Many approaches to document classification have been proposed – see Agarwal and Mittal (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 (Manning et al. 2008) 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. It is to be expected that this simple approach will produce results below the state-the-art as compared with a supervised classifier for any specific language. However, the advantage of this approach is that we can exploit an existing multilingual, domain-structured thesaurus – EuroVoc – to build our domain vectors to deliver domain classifiers for 11 domains in 24 languages, without the need to produce training data.

To create a vector representation for an input document, the document is first pre-processed and stop words and punctuation are removed. For each of the TaaS languages we took the entire dump of Wikipedia and computed inverse document frequency (idf) for each word in this corpus. 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. After filtering out stop words and punctuation, the remaining words in the input document are stemmed. We used Lucene stemmers where available and implemented new stemmers for Latvian, Lithuanian and Estonian. Finally, word frequency (tf) counts for the stems in the input document are gathered

and, using the idf scores from Wikipedia, tf*idf weights (Spärck Jones, 1972) are computed to create the vector representation of the input document. 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 (EuroVoc, 1995). (2) We used the EuroVoc-to-TaaS mapping described in Section 2.3.1 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.[3] 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 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 Table 1.

2.3. Term extraction

We performed term tagging for each Wikipedia article using Tilde's Wrapper System for CollTerm (TWSC) (Pinnis et al. 2012). TWSC identifies terms using a linguistically, statistically, and reference corpus-motivated method in the following four steps:

1. The document is POS-tagged (or morpho-syntactically tagged if morpho-syntactic taggers are available).

[3] Currently we use single words as vector elements. However, terms could be incorporated into the vector representation of both the input document and the domain. This could take the form of using terms only in the vectors and/or combining terms with single words.

2. N-grams ranging from one to four tokens in length are extracted and filtered using term patterns (i.e., regular expressions of valid parts-of-speech or morpho-syntactic tag sequences) and stop-word lists. The linguistic filtering ensures that, for morphologically rich languages, morpho-syntactic agreements between tokens of multi-word term phrase candidates are valid. The term patterns have been created either manually (e.g., for Latvian and Lithuanian) or in a semi-automatic manner by statistically analysing POS tag sequences of occurrences of terms from the EuroVoc thesaurus in the Wikipedia corpora (Aker et al. 2014).
3. The linguistically valid term candidates are then filtered using minimum frequency filters and ranked using (a) different statistical co-occurrence measures, such as the Dice coefficient and point-wise mutual information and (b) the reference corpus-motivated tf*idf measure. Here (a) acts to establish unithood while (b) is an indicator of termhood. Uni-gram terms are ranked using only the tf*idf measure. Filtering thresholds were tuned so that TWSC achieves higher F-measure using a gold standard (human annotated data set) for Latvian, Lithuanian, and English. For the remaining languages the same thresholds as for English were used.
4. Finally, the term phrase candidates are tagged in the source document by prioritising longer and higher ranking n-grams.

Table 1: TaaS Domains

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

2.4. Term alignment

For term alignment, we use the context-independent term mapping tool MPAligner (Pinnis 2013). MPAligner identifies which terms from a term-tagged document pair are reciprocal translations.

For each term pair candidate MPAligner tries to find the maximum content overlap between the two terms by building a maximised character alignment map. The identification of content overlap is performed in two steps. First, each word is pre-processed by (1) translating and transliterating it into the opposite language using probabilistic dictionaries and character-based SMT transliteration systems (Pinnis 2014), and (2) romanising it using romanisation rules. Then, for each word of the source term the method identifies the target word with which it has the highest content overlap using string similarity methods (the longest common substring and Levenshtein distance). The same process is repeated for each word of the target term. The

separate word-to-word overlaps are combined into a character alignment map that represents the content overlap between the two terms so that the content overlap is maximised. Finally, the term pair candidate is scored based on the proportion of the content overlap. If the overlap exceeds a threshold, the term pair is considered a reciprocal translation.

The approach allows the mapper to map multi-word terms and terms with different numbers of tokens in the source and target languages.

For the experiments reported below, MPAligner was executed with a consolidation threshold of 0.7 (empirically set), which means that after term mapping with a simple threshold of 0.6, MPAligner performs an analysis of the results and groups the term pair candidates into clusters of inflectional variants. The grouping conditions differ depending on how much linguistic information (lemmas, part-of-speech tags, morpho-syntactic tags, normalised forms, etc.) is available for each term in the term-tagged documents. The aim of the consolidation process is to keep low scoring term variants in a highly scoring group (possibly correct term pair), while removing high scoring variants in low scoring groups (possibly incorrect term pair).

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 (1) the accuracy of the domain classifier (2) the extent to which terms found in a document judged to be in a given domain were in the domain of their document (3) the accuracy of the boundaries of extracted terms in context and (4) the accuracy of system-proposed bilingual term alignments. As noted above, the TaaS project addressed 25 languages in total. Evaluation of each of 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.

While we used this evaluation to assess the components of BiTES, there is nothing system-specific about it, and this evaluation setup, or parts of it, could be easily reused for any comparable system.

3.1. Human assessment tasks

3.1.1. Task 1: Domain classification assessment

To assess domain classification, we present participants with a document and a list of TaaS domains (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. We encourage participants to select a primary domain– i.e. a single domain that best represents the document – but allow them to select multiple domains if they believe the document content spans more than one domain and cannot choose a primary domain. If they do opt to select multiple domains we ask them to keep the number selected to a minimum. For example, the Wikipedia article entitled "Hydraulic Fracturing"[4] (Wikipedia, 2014) 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 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. But we would limit our selection to these two.

The aim is for participants to select domains from the TaaS level-1 domains. 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 carefully review potential domain candidates, and combinations of candidates, before opting to provide a new domain.

---

[4] Aka "fracking".

3.1.2. Task 2: Term in domain assessment

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 an automatically proposed domain.

In this task (see Figure 2) 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 is a term in a different domain. If they judge the term to be in a different domain they are asked to specify the alternative domain(s). Here the candidate and the domain category are assessed together but we do not provide any specific context, such as a sentence in a document from which the term was extracted. As with the previous task we provide 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 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" 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.

While no specific source context is given, 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 advised not to assume that such resources are complete or entirely correct and to use them with caution and carry out further checks and searches (as they would in normal practice) to confirm the results.

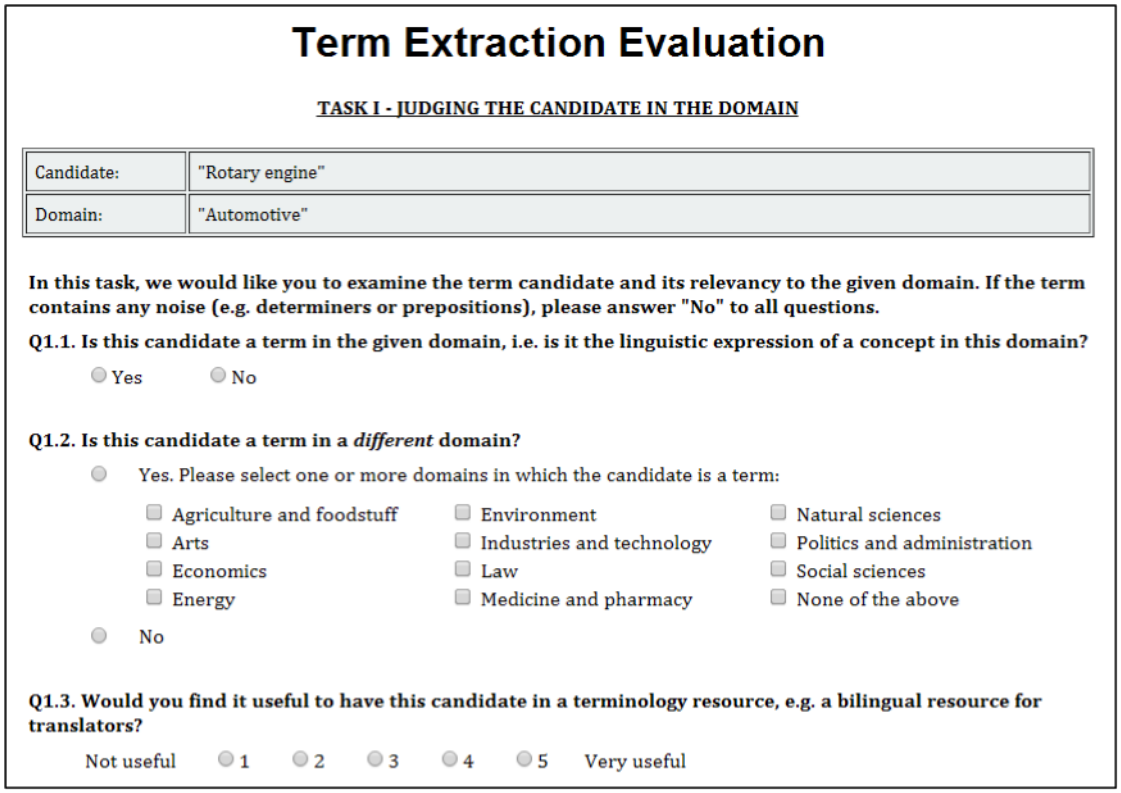


Figure 2: Task 2: Judging a term candidate in a domain

Finally, if assessors have answered “yes” to one of Q1.1 or Q1.2, they are also asked to indicate the utility of the term candidate in Q1.3. However this aspect of the assessment is not discussed further.

3.1.3. Task 3: Term boundaries in context

The second monolingual term extraction assessment task is to determine whether the boundaries of an automatically extracted term candidate, when taken in its original document context, are correct.

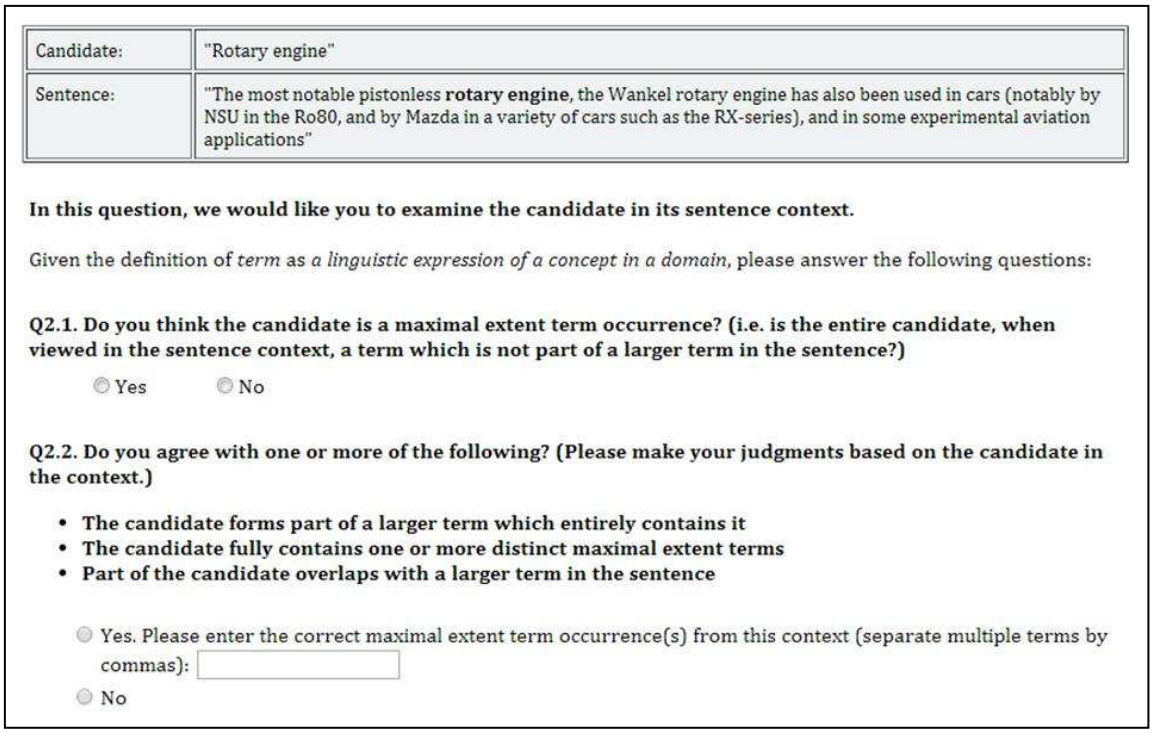


Figure 3: Task 3: Judging a term candidate in context

In this task (see Figure 3) we present assessors with a term candidate and a sentence from which the candidate was extracted. Here, we do not specify a domain, but provide the following statement: "a term is a linguistic expression of a concept in a domain". We then ask assessors to judge whether the candidate is a maximal extent term occurrence, i.e. a term occurrence that in context is not part of a larger term. If they decide that the candidate in context is i) part of a larger term, ii) overlaps with a term, iii) contains one or more terms, or iv) a combination of i)-iii), we ask them to provide the correct term extent(s). In this example, the term candidate rotary engine is a part of a larger term that entirely contains it (i.e. pistonless rotary engine); therefore, the assessors would answer "no" to Q2.1 and "yes" to Q2.2 and enter the correct maximal extent term occurrence: pistonless rotary engine.

As in the previous task, assessors are allowed to search the Internet to help determine whether the term candidates are indeed terms and to consult existing terminological or

dictionary resources, online or otherwise, during the evaluation task. The same caveats as mentioned in 3.1.2 apply.

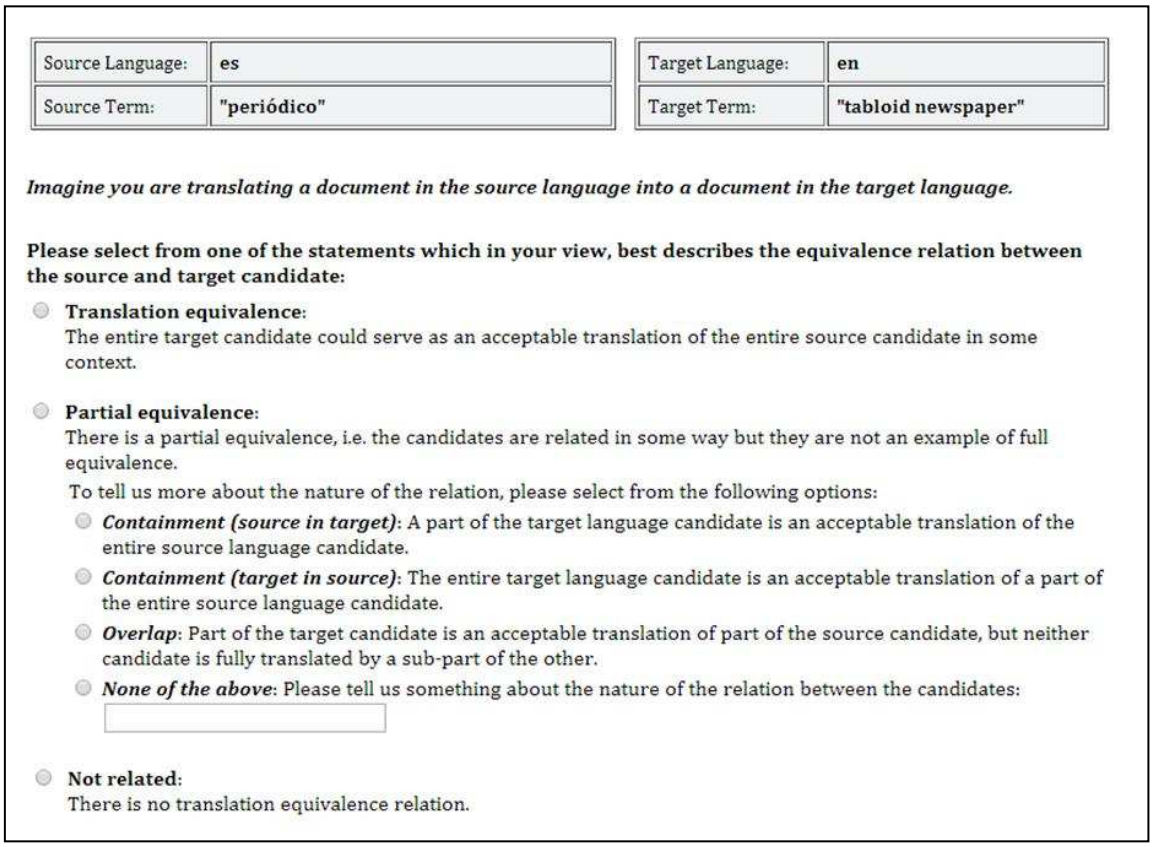
3.1.4. Task 4: Bilingual term alignments

For bilingual term alignment evaluation, we modify and extend the evaluation process described in Aker et al. (2013). In this task, we ask participants to make judgements on the nature of the semantic relation in a candidate translation pair (i.e. a pair of aligned text fragments, in different languages, where each fragment has been identified by our system as a candidate term phrase). Since the inputs are candidate terms output by automated term extractors, we can expect the aligned text fragments to contain noise, i.e., we cannot assume they always contain terms. To keep the assessment as simple as possible and focussed on a single question, we ask assessors to make their judgements based solely on the nature of the equivalence relation of the pair and irrespective of whether they believe the candidates contain terms or not (i.e. they could select the option "the candidates are equivalent" even if they believe the candidates are not terms).

To ground the task, we ask participants to imagine they are carrying out a translation job where they are translating a document in the source language into the target language. In addition we permit the assessors to search the Internet when assessing the candidate translation pairs (as translators might do), as Web searches can provide examples of language use in different languages, contexts and domains. As in previous tasks we allow assessors to consult existing dictionary resources, online or otherwise, during the evaluation with the same caveats as before.

The categories of possible semantic relation and the task instructions are shown in Figure 4. We present each term candidate pair (t1, t2) twice: first with t1 as source language candidate and then with t2 as source language candidate. For each candidate translation pair, the assessment interface prompts participants to select which of the three statements (i.e. "translation equivalence", "partial equivalence" and "not related") best describes the semantic relation between the source and target candidates. Note that for a term pair to be declared

translation equivalents we do not require substitutability in all contexts, but only in some context. In all judgements on translation candidates, we allow for inflectional variation, e.g., single vs. plural forms.



Source Language:	es	Target Language:	en
Source Term:	"periódico"	Target Term:	"tabloid newspaper"

Imagine you are translating a document in the source language into a document in the target language.

Please select from one of the statements which in your view, best describes the equivalence relation between the source and target candidate:

- **Translation equivalence**:
  The entire target candidate could serve as an acceptable translation of the entire source candidate in some context.
- **Partial equivalence**:
  There is a partial equivalence, i.e. the candidates are related in some way but they are not an example of full equivalence.
  To tell us more about the nature of the relation, please select from the following options:
  - ***Containment (source in target)***: A part of the target language candidate is an acceptable translation of the entire source language candidate.
  - ***Containment (target in source)***: The entire target language candidate is an acceptable translation of a part of the entire source language candidate.
  - ***Overlap***: Part of the target candidate is an acceptable translation of part of the source candidate, but neither candidate is fully translated by a sub-part of the other.
  - ***None of the above***: Please tell us something about the nature of the relation between the candidates:
- **Not related**:
  There is no translation equivalence relation.

Figure 4: Task 4: Assessing term alignment

3.2. Participants

We recruited experienced translators to participate in the evaluation tasks. For each of the six evaluation languages, three assessors carried out each of the evaluation tasks, with the exception of term alignment. In total our study involved 17 assessors – one assessor took part in DE only, EN-DE and EN only tasks. All assessors had excellent backgrounds 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), two were native

speakers whilst one was a fluent speaker, with over 54, 15 and 12 years experience (respectively) in using these languages as second languages.

3.3. Data

3.3.1. Domain classification

For the domain classification task, we selected a set of documents to be evaluated as follows. First, we extracted plaintext versions of all articles from the August 2013 Wikipedia dump in each of the six assessment languages, using our Wikipedia corpus collection tool (Section 2.1). 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, assigning to it the top domain proposed by the classifier, i.e. the domain with the highest score according to our vector space approach (Section 2.3.2). During processing we filtered out documents whose top domain scores were below a previously set minimum threshold and those whose length was below a minimum. 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 randomly selected one document from each bin. Since we were classifying documents into the 11 level-1 TaaS domains, this resulted in 110 documents for each language[5].

3.3.2. Term extraction

For the term-in-domain assessment task, we focussed the task 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 (Section 2.4). Next, we selected 200 terms from both domains, choosing terms of different word

[5] The Latvian set is slightly smaller (106 documents) due to fewer than 10 documents being found in one domain (only 6 documents in the "Energy" domains).

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[6]. This process was repeated for each of our six languages.

3.3.3. Term alignment

For the term alignment assessment data, we selected 150 terms pairs from each language pair from a set of documents that had previously been categorised into one of the two domains: Politics and Administration and Industries and Technology. These term pairs were randomly selected from the list of term pairs produced by the bilingual term alignment tool (Section 2.4). We used the distribution discussed in Section 3.3.2 to select candidate terms with lengths varying between 1 and 4. However, as expected, alignment of terms with lengths 3 and 4 is very rare; if insufficient terms of these lengths were found, we used what was available and made up the rest of the sample using terms of shorter lengths. Note that this evaluation set contains different terms to those used in the monolingual term extraction task, because not all the extracted terms may have been aligned by the term aligner.

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) in 78% of the cases. Considering cases where at least 2 assessors agreed on at least one domain, agreement increases to 98%.

---

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

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 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

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% were assessed to be

candidate terms in the given domain, 5% 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 that 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% of candidates 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.

Table 2: Terms with different lengths

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

Table 3: Terms in different languages

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

For the 43% of the term candidates that genuinely were terms (457 terms), all three assessors agreed about the domain of the term in 45% of the cases, i.e. they either accepted the domain

proposed by the system for the term or they agreed on an alternative(s). 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.

3.4.3. Term boundary in context assessment

Out of the 1,200 assessed terms, 134 were assessed not to be terms in Task 2 (term in domain), yet were specified to be maximal extent term occurrences in Task 3 (term in context). Some examples of these include common words or phrases (e.g. dams, fertility rate), named entities (e.g. Earl of Wessex) and non-terms (e.g. natural and social science). Due to this inconsistency, these terms were filtered out from the evaluation results.

The results in Figure 5 show that overall 40% of term occurrences are assessed as being "maximal extent" occurrences and 36% overlap with terms. This means that around 76% automatically extracted candidate terms are either correct terms (in the given domain or a different domain) or overlap with terms (i.e. these candidates would have been judged correct if the boundary had been identified correctly). Meanwhile, the remaining 24% are identified to neither be terms nor overlap with terms.

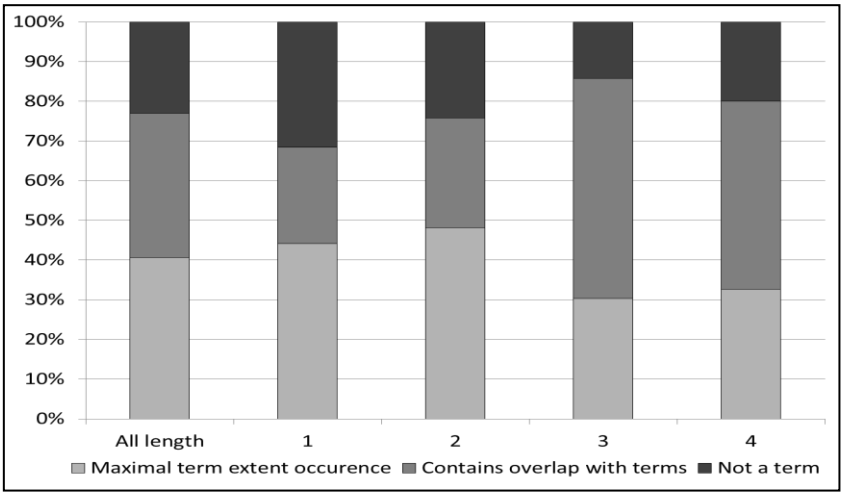


Figure 5: Accuracy of term boundaries

We analysed whether the results of the "maximal extent" assessment vary between short and long terms and found that terms with one or two words obtain more than 10% better scores than those containing 3 or 4 words, i.e., the latter are less likely to be maximal extent term occurrences. However, the proportion of candidates containing 3 or 4 words deemed not to be terms is much lower than candidates containing 1 or 2 words. On average, 25% of the candidates are judged not to be terms. This figure compared to the term-in-domain evaluation (Task 2) shows also that the context (here sentence) has an impact on the assessors' decision making.

For the first question, we obtained 57% (679 terms) full agreement between annotators and 43% partial agreement. For the second question we have 58% full agreement, 35% partial and 7% no agreement.

We also linked these findings to results from the Task 2 evaluation in order to identify characteristics of candidate terms assessed not to be terms. Of candidates assessed not to be terms approximately 58% were found to contain overlap with a term (i.e., their boundaries were incorrectly identified) and around 42% were found to be neither terms nor to contain overlap with terms. These findings also suggest that if TWSC's performance can be improved to correctly identify these term boundaries, its precision would increase to 76%.

3.4.4. Term alignment assessment

For term alignment assessment a total of 750 term pairs were assessed by two assessors, who identified the semantic relation between each candidate pair by selecting one of the options shown in Table 4. They agreed in 88% of cases.

We measured the precision over all languages as shown in Figure 6. These results indicate that MPAligner aligns terms with 94% precision, i.e. 94% of the aligned terms were assessed to be translation equivalences. Only 2% of aligned terms are assessed to be unrelated, whilst 4% were assessed to be partial equivalences.

Table 4: Possible semantic relations between aligned term pairs

Category	Semantic Relation
TE	**T**ranslation **E**quivalence
PE-SinT	**P**artial **E**quivalence: Containment (**S**ource **in** **T**arget)
PE-TinS	**P**artial **E**quivalence: Containment (**T**arget **in** **S**ource)
PE-Over	**P**artial **E**quivalence: **Over**lap
PE-Other	**P**artial **E**quivalence: None of the above (**Other**)
N/U	**N**ot Related/**U**plicable

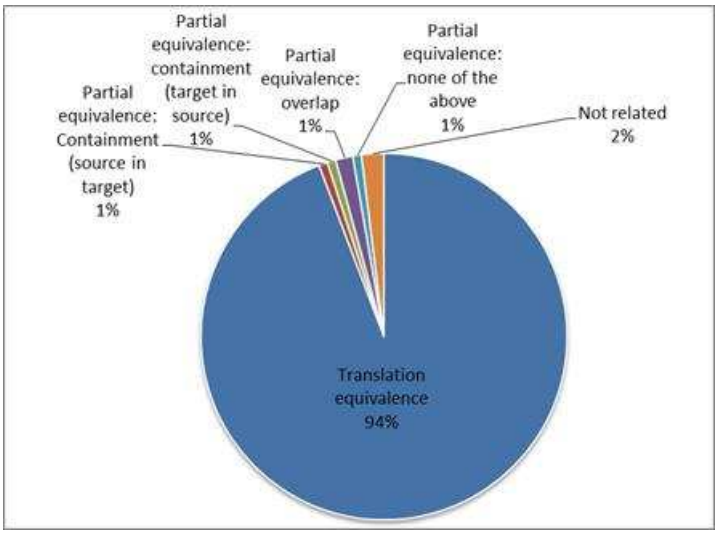


Figure 6: Term alignment results.

We further investigated the performance of the term aligner by language, as shown in Table 5.

Table 5: Term alignment results for each language pair. The numbers are percentages.

Language Pair	TE	PE-SinT	PE-TinS	PE-Over	PE-Other	N/U
CS-EN	90	3	0	3	0	4
DE-EN	90	1	2	3	3	2
ES-EN	97	1	1	1	0	1
LT-EN	98	0	0	0	0	2
LV-EN	95	0	2	2	1	2

The results show high precision (90% or above) for terms aligned for all languages, the highest accuracy being in LT (98%).

Performance figures for the term aligner when aligning terms of different lengths are shown in Table 6. Note, we categorise the data based on the length of the source language term (i.e. the non-English data). The results show that there is not much variance between the accuracy for terms with different lengths, with the accuracy ranging from 91% to 96%.

Table 6: Term alignment results for different term lengths. The numbers are percentages.

Length	Total	TE	PE-SinT	PE-TinS	PE-Over	PE-Other	N/U
1	244	91	0	1	0	1	6
2	298	96	2	0	2	0	0
3	179	95	0	1	3	1	0
4	29	95	0	0	0	5	0

4. Analysis and Discussion

Here we summarise and discuss the results above in relation to six research questions and then discuss the application of BiTES to data originating from sources other than Wikipedia.

(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?

First, we should 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 needs 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 assignments, 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 if accurate domain classification is important. However, if 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 difference in scores between the highest scoring domains is small. This would make the comparison with the human figures fairer, as now the system can only propose one domain per document while the humans can propose several. We conclude that the vector space classifier utilizing domain representations derived from a pre-existing, multi-lingual thesaurus has much to recommend it: it is simple,

needs no training data, is straightforwardly applicable to multiple (25 in our case) different languages and its performance is adequate, if it is suitably constrained.

(2) 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.

(3) 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 across languages and term lengths – 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.

(4) To what extent do humans agree on the boundaries of terms when assessing them in context?

Our results show that all three assessors agree in identifying whether a term proposed by our automatic term extractors is a maximal extent term occurrence in context in 57% of cases. For cases where the term boundaries were incorrect, assessors were asked to provide the correct maximal term occurrence and in 93% of these cases we have an agreement between at least two assessors.

(5) How accurately can our automatic term extractor identify correct term boundaries?

The results in Figure 5 show that TWSC is able to correctly identify term boundaries for 40% of the term candidates it proposes, whilst 36% of term candidates have one or both of their

boundaries incorrectly identified yet still to overlap with a genuine term. Less than a quarter are not terms and do not contain any overlap with a term.

(6) What is the accuracy of system-proposed bilingual term alignments?

The precision of MPAligner in aligning terms extracted from Wikipedia documents is above 90%, indicating highly accurate bilingual term extraction. Such term pairs are very important for machine translation and their injection into existing parallel data can significantly increase the performance of SMT systems (Aker et al. 2012a).

BiTES incorporates four different workflows – one that uses inter-language linked Wikipedia articles as comparable documents, another that uses comparable news articles, a third that uses general web documents and a fourth that extracts bilingual term pairs from parallel data. An interesting question is what the quality of bilingual term pairs is when noisier data are used (news and generic Web data) or when the data used is parallel. To investigate this we conducted a manual evaluation on a sample of bilingual terms (150 term pairs for each language pair) resulting from each workflow, following the same evaluation protocol as described in Section 3.4.4 for Wikipedia data. Table 7 reports the proportion of system outputs assessed as "translation equivalents".

Table 7: Term alignment results for different workflows. The numbers are percentages.

Language pairs	Wikipedia	FMC	News	Parallel data
All language pairs	94	87	90	98
CS-EN	90	91	93	-
DE-EN	90	97	95	97
ES-EN	97	90	94	98
LT-EN	98	82	83	98
LV-EN	95	85	84	98

The news data was collected using the news gathering tool reported in Aker et al. (2012b). The generic Web data was obtained using the FMC crawler (Mastropavlos and Papavassiliou, 2011). The parallel data was obtained using a STRAND-like approach (Resnik and Smith 2003). The results suggest that the quality of the extracted term pairs across different workflows varies with the likely comparability of the data sources. The parallel data workflow, which aligns terms

contained in parallel segments, produces the highest quality bilingual terms (up to 98% translation equivalence), which was expected as the parallel nature of the data significantly reduces the likelihood of term candidates being incorrectly aligned. The Wikipedia workflow, which we believe yields more highly comparable document pairs than the news article workflow or the generic Web document workflow, produces aligned terms with 94% translation equivalence, followed by the news workflow (90% translation equivalence) and FMC (87% translation equivalence).

We also investigated the number of terms resulting from the different data sources to get an idea of the yield of the workflow. Results are summarised in Table 8, which reports average number of term pairs found in each document pair in each workflow.

Table 8: Term pairs per document/parallel segment pair.

Workflow	Term pairs per document pair
Wikipedia workflow	5.1
FMC workflow	0.22
News workflow	3.5
Parallel data workflow	0.5

Table 8 show that both news and Wikipedia workflows are good resources for retrieving bilingual term pairs. On the other hand, the FMC workflow produces significantly fewer term pairs per document pair (approximately 1 term pair found in 5 document pairs). This is likely due to the comparability of document pairs produced in this workflow being low or the web documents found containing few terms. The parallel data workflow does not produce document pairs but rather parallel segments and in this case we report the average number of term pairs per parallel text segment (sentence). This workflow produces 0.5 term pairs per parallel text segment, which is a much higher rate than the comparable corpora workflows.

5. Related Work

5.1. Component technologies

There has been extensive previous work in all the component technology areas of BiTES: corpus collection from the Web, document classification, monolingual term extraction and bilingual term alignment. We cannot possibly hope to position BiTES in relation to all this work and, besides, the contribution of BiTES is not so much in the specifics of the individual components we use (though there are novelties in some of these as indicated in relevant citations above), but in how they have been brought together to produce an end-to-end bilingual term extraction system for 24 languages pairs with modest effort.

For example, document classification has been exhaustively studied, particularly using machine learning methods (Sebastiani 2002; Manning et al. 2008). State-of-the-art results for standardized tasks, such as the ModApte split of the Reuters-21758 corpus, are over 90% F1 measure for the top 10 classes. These figures are well beyond what we achieve but are for one language only where substantial numbers of labelled training documents exist. The novelty in BiTES is in exploiting EuroVoc to assign documents to a common set of domains across 25 languages with reasonable accuracy and without any labelled documents.

Monolingual terminology extraction has also been widely studied (see, e.g. Pazienza et al., 2005). Like many term extraction approaches, the BiTES TWSC component uses a combination of both linguistic and statistical information. It is distinctive because it runs on 25 languages, something it achieves via two features: (1) The linguistic POS tag sequence patterns it uses are induced from occurrences of EuroVoc terms matched in POS-tagged Wikipedia sentences (2) Many term extraction approaches exploit statistical contrasts between domain-specific and general reference corpora, where collections of domain-specific documents (for each domain and language) are either presupposed (Chung 2003; Drouin 2004; Kim et al. 2009; Marciniak and Mykowiecka 2013; Kilgariff 2014) or gathered from the Web using existing domain-specific term lists or seed terms (Kida et al. 2007; De Benedictis et al. 2013). By

contrast, TWSC is given a document already classified into a domain, extracts terms from it and assigns them the domain of the document.

Bilingual term alignment too has been well studied. Much work has focused on aligning terms in parallel corpora (Kupiec 1993; Daille et al. 1994; Fan et al. 2009; Okita et al. 2010; Bouamor et al. 2012), however parallel data is insufficiently available for minority languages and specialized domains. Instead we need to exploit comparable corpora for which we need techniques that do not depend on alignment information. Such techniques can be based on one or a combination of:

- Cognate information, typically computed by some sort of a transliteration measure (e.g. Al-Onaizan and Knight 2002; Knight and Graehl 1998; Udupa et. al. 2008; Aswani and Gaizauskas 2010).
- Context congruence – a measure of the extent to which the words that the source term co-occurs with have the same sort of distribution and co-occur with words with the same sort distribution as do those words that co-occur with the candidate target term (e.g. Rapp 1995; Fung and McKeown 1997; Morin et al. 2007; Cao and Li 2002; Ismail and Manandhar 2010);
- Translation of component words in terms and/or in context words, where some limited dictionary exists (e.g. Cao and Li 2002; Aker et al. 2013).

To ensure that MPAligner supports cross-lingual term mapping between all 25 TaaS languages and works with documents that are relatively short (e.g., possibly single sentences), MPAligner uses a context independent method that performs cognate and translation-based term mapping without the need of training supervised models.

5.2. Evaluation

Much of the work on the evaluation of monolingual term extraction and bilingual term alignment takes the form of manual review of outputs from implemented systems or comparison against pre-existing term resources and is for a small number of languages (e.g. Kim et al. 2009;

Daille et al. 2004; Drouin 2004). Our work contrasts with this in that we have used multiple assessors, have avoided limiting comparisons with existing resources and have evaluated across six languages.

Directly assessing system outputs stands in contrast to the approach taken in related areas, such as named entity extraction, where the norm is to create a gold standard annotated corpus independently from any particular system and then evaluate system outputs against this gold standard (e.g. Grishman and Sundheim 1996; Sang and De Meulder 2003). While such an approach has advantages, such as enabling system developers to evaluate system variants whenever they please with no additional human effort, it also has problems: (1) it requires the creation of explicit, detailed guidelines for annotating terms, which are extremely difficult to produce and gain agreement on (by contrast it is straightforward to get experienced terminologists and translators to judge system-proposed candidates); (2) annotating all terms in running text is wasteful in that technical documents tend to have many occurrences of the same terms or variants of them and annotator time is wasted redundantly annotating the same term (by contrast it is far less effortful to select and review a sample from system-annotated documents than to choose and then fully annotate a set of complete documents).

6. Conclusion

In this paper we have described an approach to automatic extraction of bilingual term pairs from web sources, implemented in a system called BiTES, and have reported an evaluation of the system's major components. The system is embedded in the TaaS on-line system terminology platform and the terms gathered by BiTES form a significant part of the TaaS term-base, which is in daily use by translators and terminologists.

The major contributions of our work are two-fold:

(1) A multi-component approach for the automatic acquisition of domain-classified bilingual term pairs from web sources. Our system comprises four major software components, for gathering sets of comparable document pairs or parallel fragments

from the web, classifying documents into domains, automatically extracting terms from monolingual documents and aligning extracted terms from comparable documents or parallel fragments. A major strength of the approach is that our techniques have been readily extensible to work on 24 language pairs without the need for labour-intensive, language specific resource development, though we can take advantage of language-specific resources, when available.

(2) A set of task definitions and protocols for intrinsic evaluation of various components of our bilingual term extraction pipeline. These task definitions and protocols may be reused to evaluate other automatic term extraction systems. Their strength is that they do not require the creation of a gold standard corpus of term-annotated documents in advance of the evaluation, with all the overheads that entails, while they do afford significant insight at reasonable cost and result in materials that can be used as an approximation to a gold standard in subsequent work.

Some of the key results of our evaluations show:

- Our simple domain classification method, which is straightforward to implement and needs no training data for the 25 languages we address, achieves 77% agreement in domain assignment with at least one assessor for the most confidently classified documents.
- Humans generally agree about domain classification of documents and terms – in 99% of cases at least 2 of 3 assessors agree on at least one domain for a document.
- Terms are generally (88% of the time on average) likely to be of the same domain as the document in which they occur.
- Three assessors agree in identifying whether a term is a maximal extent term occurrence in a given context in 57% of cases. For cases where system-proposed term boundaries

are incorrect at least two of three assessors agree about the maximal term occurrence 93% of the time.

- Our monolingual term extractor TWSC correctly identifies term boundaries in 40% of the candidate terms it proposes, whilst a further 36% of its proposed candidates have imperfect boundaries yet still overlap with genuine terms.
- Our bilingual term aligner correctly identifies bilingual term equivalents in Wikipedia comparable corpora with accuracy of over 90%, with similar accuracy for data from other workflows.

There are various directions to pursue in future work. Each of the individual BiTES components can be improved in various ways. The module most needing performance improvement is the monolingual term extractor. More detailed failure analysis needs to be carried out to determine the best way to improve it. Perhaps the biggest challenge is to move BiTES beyond European languages (i.e. those represented in EuroVoc). This requires an equivalent data resource to EuroVoc for new languages or a new approach to training our document classifier and to inducing term grammars. Another, more open-ended challenge is to investigate how feedback from end users of the TaaS platform could be used to adapt BiTES components.

Acknowledgements

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 all their work in making the project a success and the anonymous reviewers whose comments have helped us improve the paper.

References

Agarwal, B., and N. Mittal. 2014. “Text Classification Using Machine Learning Methods - A Survey.” In Proceedings of the 2nd International Conference on Soft Computing for Problem Solving (SocProS 2012), 701-709. New Delhi: Springer.

Aker, A., Y. Feng, and R.J. Gaizauskas. 2012a. “Automatic Bilingual Phrase Extraction from Comparable Corpora.” In Proceedings of the 24th International Conference on Computational Linguistics (Posters) (COLING 2012), 23-32. Bombay: The COLING 2012 Organizing Committee.

Aker, A., E. Kanoulas, and R.J. Gaizauskas, R. J. 2012b. “A Light Way to Collect Comparable Corpora from the Web.” In Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012), 15-20. Istanbul: European Language Resources Association (ELRA).

Aker, A., M.L. Paramita, E. Barker, and R, Gaizauskas. 2014. “Bootstrapping Term Extractors for Multiple Languages.” In Proceedings of the 9th International Conference on Language Resources and Evaluation Conference (LREC 2014), 483-489. Reykjavik: European Language Resources Association.

Aker, A., M. Paramita, and R. Gaizauskas. 2013. “Extracting Bilingual Terminologies from Comparable Corpora.” In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL 2013), 402-411. Sofia: Association for Computational Linguistics.

Al-Onaizan, Y., and K. Knight. 2002. “Machine Transliteration of Names in Arabic Text.” In Proceedings of the ACL-02 Workshop on Computational Approaches to Semitic Languages, 1–13. Stroudsburg: Association for Computational Linguistics.

Aswani, N., and R. Gaizauskas. 2010. “English-Hindi Transliteration Using Multiple Similarity Metrics.” In Proceedings of the Seventh International Conference on Language

Resources and Evaluation (LREC 2010), 1786-1793. Valetta: European Language Resources Association (ELRA).

Bouamor, D., N. Semmar, and P. Zweigenbaum. 2012. "Identifying Bilingual Multi-Word Expressions for Statistical Machine Translation." In Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012), 674-679. Istanbul: European Language Resources Association (ELRA).

Cao, Y., and H. Li. 2002. "Base Noun Phrase Translation Using Web Data and the EM Algorithm." In Proceedings of the 19th International Conference on Computational Linguistics - Volume 1, 1-7. Stroudsburg: Association for Computational Linguistics.

Chung, T. M. 2003. "A Corpus Comparison Approach for Terminology Extraction." Terminology, 9 (2): 221-246.

Daille, B., E. Gaussier, and J. Lange. 1994. "Towards Automatic Extraction of Monolingual and Bilingual Terminology." In Proceedings of the 15th Conference on Computational Linguistics - Volume 1, 515-521. Stroudsburg: Association for Computational Linguistics.

De Benedictis, F., S. Faralli, and R. Navigli. 2013. "Glossboot: Bootstrapping Multilingual Domain Glossaries from the Web." In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL 2013), 528-538. Sofia: Association for Computational Linguistics.

De Bessé, B., B. Nkwenti-Azeh, and J.C. Sager. 1997. "Glossary of Terms Used in Terminology." Terminology: International Journal of Theoretical and Applied Issues in Specialized Communication, 4 (1): 117-156.

Drouin, P. 2004. "Detection of Domain Specific Terminology Using Corpora Comparison." In Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 2004), 79-82. Lisbon: European Language Resources Association (ELRA).

EuroTermBank. 2015. EuroTermBank. Accessed September 15. http://www.eurotermbank.com.

EuroVoc. 2015. EuroVoc, the EU's Multilingual Thesaurus. Thesaurus Eurovoc - Volume 2: Subject-Oriented Version. Ed. 3/English Language. Annex to the index of the Official

Journal of the EC. Luxembourg, Office for Official Publications of the European Communities. http://eurovoc.europa.eu/.

Fan, X., N. Shimizu, and H. Nakagawa. 2009. "Automatic Extraction of Bilingual Terms from a Chinese-Japanese Parallel Corpus." In Proceedings of the 3rd International Universal Communication Symposium (IUCS '09), 41-45. New York: Association for Computing Machinery (ACM).

Fung, P., and K. McKeown. 1997. "Finding Terminology Translations from Non-Parallel Corpora." In Proceedings of the 5th Annual Workshop on Very Large Corpora, 192-202. Hong Kong: Association for Computational Linguistics.

Gaizauskas, R., E. Barker, M.L. Paramita, and A. Aker. 2014. "Assigning Terms to Domains by Document Classification." In Proceedings of the 4th International Workshop on Computational Terminology (Computerm), 11-21. Dublin: Association for Computational Linguistics and Dublin City University.

Gornostay, T., and A. Vasiljevs. 2014. "Terminology Resources and Terminology Work Benefit from Cloud Services." In Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 2014), 1943-1948. Reykjavik: European Language Resources Association (ELRA).

Grishman, R., and B. Sundheim. 1996. "Message Understanding Conference - 6: A Brief History." In Proceedings of the 16th International Conference on Computational Linguistics, 466-471. Copenhagen: Association for Computational Linguistics.

Halcsy, P., A. Kornai, and C. Oravecz. 2007. "HunPos: an Open Source Trigram Tagger." In Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions, 209-212. Prague: Association for Computational Linguistics.

IATE. 2015. InterActive Terminology for Europe. Accessed September 15. http://iate.europa.eu.

Ismail, A., and S. Manandhar. 2010. "Bilingual Lexicon Extraction from Comparable Corpora Using In-Domain Terms." In Proceedings of the 23rd International Conference on

Computational Linguistics: Poster (COLING 2010), 481–489. Beijing: COLING 2010 Organizing Committee.

Justeson, J.S., and S.M. Katz. 1995. “Technical Terminology: Some Linguistic Properties and an Algorithm for Identification in Text.” Natural Language Engineering, 1 (1): 9-27.

Kida, M., M. Tonoike, T. Utsuro, and S. Sato. 2007. “Domain Classification of Technical Terms Using the Web.” Systems and Computers in Japan, 38 (14): 11-19.

Kilgariff, A., M. Jakubıcek, V. Kovár, P. Rychlý, and V. Suchomel. 2014. “Finding Terms in Corpora for Many Languages with the Sketch Engine.” In Proceedings of the Demonstrations at the 14th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2014), 53-56. Gothenburg: Association for Computational Linguistics.

Kim, S. N., T. Baldwin, and M-Y. Kan. 2009. “An Unsupervised Approach to Domain-Specific Term Extraction.” In Proceedings of the Australasian Language Technology Association Workshop, 94-98. Sydney: Australasian Language Technology Association.

Knight, K., and J. Graehl. 1998. “Machine Transliteration.” Computational Linguistics, 24 (4): 599-612.

Koehn, P., H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst. 2007. “Moses: Open Source Toolkit for Statistical Machine Translation.” In Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions (ACL 2007), 177-180. Prague: Association for Computational Linguistics.

Kupiec, J. 1993. “An Algorithm for Finding Noun Phrase Correspondences in Bilingual Corpora.” In Proceedings of the 31st Annual Meeting on Association for Computational Linguistics (ACL 1993), 17-22. Columbus: Association for Computational Linguistics.

Manning, C.D., P. Raghavan, and H. Schütze. 2008. Introduction to Information Retrieval. Cambridge: Cambridge University Press.

Marciniak, M., and A. Mykowiecka. 2013. “Terminology Extraction from Domain Texts in Polish.” In Intelligent Tools for Building a Scientific Information Platform, 171-185. Berlin, Heidelberg: Springer.

Mastropavlos, N., and V. Papavassiliou. 2011. “Automatic Acquisition of Bilingual Language Resources.” In Proceedings of the 10th International Conference of Greek Linguistics (ICGL 2011). Komotini, Greece.

Morin, E., B. Daille, K. Takeuchi, and K. Kageura. 2007. “Bilingual Terminology Mining Using Brain, not Brawn Comparable Corpora.” In Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics (ACL 2007), 664-671. Prague: Association for Computational Linguistics.

Okita, T., A. Maldonado-Guerra, Y. Graham, and A. Way. 2010. “Multi-Word Expression-Sensitive Word Alignment.” In Proceedings of the 4th International Workshop on Cross Lingual Information Access (CLIA 2010), 26-34. Beijing: COLING 2010 Organizing Committee.

Paramita, M.L., P. Clough, A. Aker, and R.J. Gaizauskas. 2012. “Correlation Between Similarity Measures for Inter-Language Linked Wikipedia Articles.” In Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012), 790–797. Istanbul: European Language Resources Association.

Pazienza, M.T., M. Pennacchiotti, and F.M. Zanzotto. 2005. “Terminology Extraction: an Analysis of Linguistic and Statistical Approaches.” In Knowledge Mining, 255-279. Berlin, Heidelberg: Springer.

Pinnis, M. 2013. “Context Independent Term Mapper for European Languages.” In Proceedings of Recent Advances in Natural Language Processing (RANLP 2013), 562-570. Hissar: Incoma Ltd. Shoumen, Bulgaria.

Pinnis, M. 2014. “Bootstrapping of a Multilingual Transliteration Dictionary for European Languages. ” In Human Language Technologies The Baltic Perspective - Proceedings of the 6th International Conference Baltic (HLT 2014), 132–140. Amsterdam: IOS Press.

Pinnis, M.. and K. Goba. 2011. “Maximum Entropy Model for Disambiguation of Rich Morphological Tags.” In Proceedings of the 2nd International Workshop on Systems and Frameworks for Computational Morphology (SFCM 2011), 14-22. Berlin, Heidelberg: Springer.

Pinnis, M., N. Ljubešic, D. Stefanescu, I. Skadina, M. Tadic, and T. 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), 20-21. Madrid.

Rapp, R. 1995. “Identifying Word Translations in Non-Parallel Texts.” In Proceedings of the 33rd Annual Meeting on Association for Computational Linguistics (ACL 1995), 320-322. Cambridge, Massachusetts: Association for Computational Linguistics.

Resnik, P., and N.A. Smith. 2003. “The Web as a Parallel Corpus.” Computational Linguistics, 29 (3): 349-380.

Sang, E.F.T.K., and F. De Meulder, F. 2003. “Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition.” In Proceedings of the 7th Conference on Natural Language Learning at HLT-NAACL 2003 – Volume 4, 142-147. Edmonton: Association for Computational Linguistics.

Spärck Jones, K. 1972. "A Statistical Interpretation of Term Specificity and Its Application in Retrieval.” Journal of Documentation, 28: 11–21.

Steinberger, R., B. Pouliquen, and J. Hagman. 2002. “Cross-Lingual Document Similarity Calculation Using the Multilingual Thesaurus EuroVoc.” Computational Linguistics and Intelligent Text Processing, 415-424. Berlin, Heidelberg: Springer.

Udupa, R., K. Saravanan, A. Kumaran, and J. Jagarlamudi. 2008. “Mining Named Entity Transliteration Equivalents from Comparable Corpora.” In Proceedings of the 17th ACM Conference on Information and Knowledge Management (CIKM 2008), 1423-1424. New York: Association for Computing Machinery.

Wikipedia. 2014. “Hydraulic Fracturing.” Accessed June 23. http://en.wikipedia.org/wiki/Hydraulic_fracturing.

Authors' addresses

Robert Gaizauskas
University of Sheffield
Sheffield, United Kingdom
r.gaizauskas@sheffield.ac.uk

Monica Lestari Paramita
University of Sheffield
Sheffield, United Kingdom
m.paramita@sheffield.ac.uk

Emma Barker
University of Sheffield
Sheffield, United Kingdom
e.barker@sheffield.ac.uk

Mārcis Pinnis
Tilde, Latvia
marcis.pinnis@tilde.lv

Ahmet Aker
University of Sheffield
Sheffield, United Kingdom
ahmet.aker@sheffield.ac.uk

Marta Pahisa Solé
University of Sheffield
Sheffield, United Kingdom
mpahisa@gmail.com

About the authors

Robert Gaizauskas is Professor of Computer Science and Head of the Natural Language Processing (NLP) group within the Department of Computer Science, University of Sheffield, where he has worked since 1993. He obtained a DPhil from the School of Cognitive and Computing Sciences, University of Sussex, in 1992. His research interests lie in applied NLP, especially in its potential to improve information access to large text collections. To that end he has worked on information extraction, information retrieval, automatic question answering, text summarization, text reuse, cross-language word alignment, mono- and bilingual term extraction and use of comparable corpora for machine translation. He has published over 160 papers in peer-reviewed journals and conference proceedings and has served on the programme committees of numerous leading international conferences and workshops in the area of computational linguistics and as a reviewer for all the major journals in this area.

Monica Lestari Paramita is a Research Associate in Natural Language Processing at the University of Sheffield. She obtained her bachelor degree from the Computer Science Department, University of Indonesia in 2006, and her master degree from the Information School, University of Sheffield in 2008. Since 2008, she has worked in various areas in Information Retrieval and Natural Language Processing, specifically in cross-lingual similarity

identification and bilingual term extraction. She is currently a PhD student in the Information School, University of Sheffield, in which she develops language-independent methods to identify similarity in Wikipedia.

Emma Barker is a Research Associate in Natural Language Processing at the University of Sheffield. Her research interests are in user contexts for language processing technologies (e.g. information seeking, translation, reading comprehension and writing tasks), and how knowledge of context (tasks and texts) can inform technology applications and evaluation. She has expertise in text analysis and annotation and has worked on various projects to build evaluation resources and carry out user evaluations, in areas e.g. news and social media summarisation, image captioning, cross-lingual term, and phrase alignment and extraction. She has a PhD in History and Information Studies (The University of Sheffield), an MPhil in History and Computing and an MA Hons. in History (The University of Glasgow).

Mārcis Pinnis is a researcher at Tilde, a private company that is based in Latvia and is developing language technologies. His current research interests are statistical machine translation, term extraction, cross-lingual term mapping, and machine transliteration. Mārcis received his bachelor's degree in Computer Science from the University of Latvia, and his MPhil Degree in Computer Speech, Text and Internet Technology from the St. Edmund's College (University of Cambridge, Cambridge, UK). He is currently finalising his PhD studies in Computer Science at the University of Latvia. The topic of his PhD thesis is about integration of multilingual terminology into statistical machine translation systems.

Ahmet Aker is a Research Fellow at the University of Sheffield in the natural-language processing group. His research interests are in automatic text summarisation, machine learning, statistical machine translation, comparable data acquisition from the Web, multilingual term alignment, short text clustering and semantic linking. Ahmet has a German Diploma in computer science, an MS in advanced software engineering and PhD in natural language processing.

Marta Pahisa Solé is a freelance translator, project manager and trainer and has worked as an associate teacher of IT and Terminology applied to Translation at Universitat Autònoma de Barcelona, and as a Translation Technologies and Localisation Tutor at The University of Sheffield. Her research focuses on the didactics of localisation, terminology and corpus studies.