Bilingual Termbank Creation via Log-Likelihood Comparison and Phrase-Based Statistical Machine Translation

Rejwanul Haque, Sergio Penkale, Andy Way[†]
Lingo24, Edinburgh, UK
{rejwanul.haque, sergio.penkale}@lingo24.com
[†]CNGL, Centre for Global Intelligent Content
School of Computing, Dublin City University
Dublin 9, Ireland
away@computing.dcu.ie

Abstract

Bilingual termbanks are important for many natural language processing (NLP) applications, especially in translation workflows in industrial settings. In this paper, we apply a log-likelihood comparison method to extract monolingual terminology from the source and target sides of a parallel corpus. Then, using a Phrase-Based Statistical Machine Translation model, we create a bilingual terminology with the extracted monolingual term lists. We manually evaluate our novel terminology extraction model on English-to-Spanish and English-to-Hindi data sets, and observe excellent performance for all domains. Furthermore, we report the performance of our monolingual terminology extraction model comparing with a number of the state-of-the-art terminology extraction models on the English-to-Hindi datasets.

1 Introduction

Terminology plays an important role in various NLP tasks including Machine Translation (MT) and Information Retrieval. It is also exploited in human translation workflows, where it plays a key role in ensuring translation consistency and reducing ambiguity across large translation projects involving multiple files and translators over a long period of time. The creation of monolingual and bilingual terminological resources using human experts are, however, expensive and time-consuming tasks. In contrast, automatic terminology extraction is much faster and less expensive, but cannot be guaranteed to be error-free. Accordingly, in real NLP applications, a manual inspection is required to amend or discard anomalous items from an automatically extracted terminology list.

The automatic terminology extraction task starts with selecting candidate terms from the input domain corpus, usually in two different ways: (i) linguistic processors are used to identify noun phrases that are regarded as candidate terms (Kupiec, 1993; Frantzi et al., 2000), and (ii) non-linguistic *n*-gram word sequences are regarded as candidate terms (Deane, 2005).

Various statistical measures have been used to rank candidate terms, such as C-Value (Ananiadou et al., 1994), NC-Value (Frantzi et al., 2000), log-likelihood comparison (Rayson and Garside, 2000), and TF-IDF (Basili et al., 2001). In this paper, we present our bilingual terminology extraction model, which is composed of two consecutive and independent processes:

1. A log-likelihood comparison method is employed to rank candidate terms (*n*-gram word sequences) independently from the source and target sides of a parallel corpus,

2. The extracted source terms are aligned to one or more extracted target terms using a Phrase-Based Statistical Machine Translation (PB-SMT) model (Koehn et al., 2003).

We then evaluate our novel bilingual terminology extraction model on various domain corpora considering English-to-Spanish and low-resourced and less-explored English-to-Hindi language-pairs and see excellent performance for all data sets.

Proceedings of the 4th International Workshop on Computational Terminology, pages 42–51,
Dublin, Ireland, August 23 2014.

The remainder of the paper is organized as follows. In Section 2, we discuss related work. In Section 3, we describe our two-stage terminology extraction model. Section 4 presents the results and analyses of our experiments, while Section 5 concludes, and provides avenues for further work.

2 Related Work

Several algorithms have been proposed to extract terminology from a domain-specific corpus, which can be divided into three broad categories: linguistic, statistical and hybrid. Statistical or hybrid approaches dominate this field, with some of the leading work including the use of frequency-based filtering (Daille et al., 1994), NC-Value (Frantzi et al., 2000), log-likelihood and mutual information (Rayson and Garside, 2000; Pantel and Lin, 2001), TF-IDF (Basili et al., 2001; Kim et al., 2009), weirdness algorithm (Ahmad et al., 1999), Glossex (Kozakov et al., 2004) and Termex (Sclano and Velardi, 2007).

In this work, we focus on extracting bilingual terminology from a parallel corpus. He et al. (2006) demonstrate that using log-likelihood for term discovery performs better than TF-IDF. Accordingly, similarly to Rayson and Garside (2000) and Gelbukh et al. (2010), we extract terms independently from both sides of a parallel corpus using log-likelihood comparisons with a generic reference corpus. Some of the most influential research on bilingual terminology extraction includes Kupiec (1993), Gaussier (1998), Ha et al. (2008) and Lefever et al. (2009). Lefever et al. (2009) proposed a sub-sentential alignment-based terminology extraction module that links linguistically motivated phrases in parallel texts. Unlike our approach, theirs relies on linguistic analysis tools such as PoS taggers or lemmatizers, which might be unavailable for under-resourced languages (e.g., Hindi). Gaussier (1998) and Ha et al. (2008) applied statistical approaches to acquire parallel term-pairs directly from a sentence-aligned corpus, with the latter focusing on improving monolingual term extraction, rather than on obtaining a bilingual term list. In contrast, we build a PB-SMT model (Koehn et al., 2003) from the input parallel corpus, which we use to align a source term to one or more target terms. While Rayson and Garside (2000) and Gelbukh et al. (2010) only allowed the extraction of single-word terms, we focus on extraction of up to 3-gram terms.

3 Methodology

In this section, we describe our two-stage bilingual terminology extraction model. In the first stage, we extract monolingual terms independently from either side of a sentence-aligned domain-specific parallel corpus. In the second stage, the extracted source terms are aligned to one or more extracted target terms using a PB-SMT model.

3.1 Monolingual Terminology Extraction

The monolingual term extraction task involves the identification of terms from a list of candidate terms formed from all *n*-gram word sequences from the monolingual domain corpus (i.e. in our case, each side of the domain parallel corpus, cf. Section 4.1). On both source and target sides, we used lists of language-specific stop-words and punctuation marks in order to filter out anomalous items from the candidate termlists. In order to rank the candidate terms in those lists, we used a log-likelihood comparison method that compares the frequencies of each candidate term in both the domain corpus and the large general corpus used as a reference.[1]

The log-likelihood (LL) value of a candidate term ($\mathrm{C_n}$) is calculated using equation (1) from Gelbukh et al. (2010).

$$\mathrm{LL} = 2 * ((\mathrm{F_d} * \log(\mathrm{F_d}/\mathrm{E_d})) + (\mathrm{F_g} * \log(\mathrm{F_g}/\mathrm{E_g}))) \quad (1)$$

where $\mathrm{F_d}$ and $\mathrm{F_g}$ are the frequencies of $\mathrm{C_n}$ in the domain corpus and the generic reference corpus, respectively. $\mathrm{E_d}$ and $\mathrm{E_g}$ are the expected frequencies of $\mathrm{C_n}$, which are calculated using (2) and (3).

$$\mathrm{E_d} = \mathrm{N_d^n} * (\mathrm{F_d} + \mathrm{F_g})/(\mathrm{N_d^n} + \mathrm{N_g^n}) \quad (2)$$

$$\mathrm{E_g} = \mathrm{N_g^n} * (\mathrm{F_d} + \mathrm{F_g})/(\mathrm{N_d^n} + \mathrm{N_g^n}) \quad (3)$$

[1] Before the term-extraction process begins, we apply a number of preprocessing methods including tokenisation to the input domain corpus and the generic reference corpus.

where $\mathrm{N}_{\mathrm{d}}^{\mathrm{n}}$ and $\mathrm{N}_{\mathrm{g}}^{\mathrm{n}}$ are the numbers of *n*-grams in the domain corpus and reference corpus, respectively. Thus, each candidate term is associated with a weight (LL value) which is used to sort the candidate terms: those candidates with the highest weights have the most significant differences in frequency in the two corpora. However, we are interested in those candidate terms that are likely to be terms in the domain corpus. Gelbukh et al. (2010) used the condition in (4) in order to filter out those candidate terms whose relative frequencies are bigger in the domain corpus than in the reference corpus, and we do likewise.

$$\mathrm{F_d}/\mathrm{N_d^n} > \mathrm{F_g}/\mathrm{N_g^n} \tag{4}$$

In contrast with Gelbukh et al. (2010), we extract multi-word terms up to 3-grams, whereas they focused solely on extracting single word terms.

3.2 Creating a Bilingual Termbank

We obtained source and target termlists from the bilingual domain corpus using the approach described in Section 3.1. We use a PB-SMT model (Koehn et al., 2003) to create a bilingual termbank from the extracted source and target termlists.

This section provides a mathematical derivation of the PB-SMT model to show how we scored candidate term-pairs using the PB-SMT model. We built a source-to-target PB-SMT model from the bilingual domain corpus using the Moses toolkit (Koehn et al., 2007). In PB-SMT, the posterior probability $\mathrm{P}(e_1^I|f_1^J)$ is directly modelled as a (log-linear) combination of features (Och and Ney, 2002), that usually comprise M translational features, and the language model, as in (5):

$$\log \mathrm{P}(e_1^I|f_1^J) = \sum_{m=1}^{M} \lambda_m h_m(f_1^J, e_1^I, s_1^K) + \lambda_{\mathrm{LM}} \log \mathrm{P}(e_1^I) \tag{5}$$

where $e_1^I = e_1, ..., e_I$ is the probable candidate translation for the given input sentence $f_1^J = f_1, ..., f_J$ and $s_1^K = s_1, ..., s_k$ denotes a segmentation of the source and target sentences respectively into the sequences of phrases $(\hat{f}_1, ..., \hat{f}_k)$ and $(\hat{e}_1, ..., \hat{e}_k)$ such that (we set $i_0 := 0$):

$$\begin{aligned} \forall k \in [1, K] \quad & s_k := (i_k; b_k, j_k), (b_k \text{ corresponds to starting index of } f_k) \\ & \hat{e}_k := \hat{e}_{i_{k-1}+1}, ..., \hat{e}_{i_k}, \\ & \hat{f}_k := \hat{f}_{b_k}, ..., \hat{f}_{j_k} \end{aligned}$$

Each feature h_m in (5) can be rewritten as in (6):

$$h_m(f_1^J, e_1^I, s_1^K) = \sum_{k=1}^{K} \hat{h}_m(\hat{f}_k, \hat{e}_k, s_k) \tag{6}$$

Therefore, the translational features in (5) can be rewritten as in (7):

$$\sum_{m=1}^{M} \lambda_m h_m(f_1^J, e_1^I, s_1^K) = \sum_{m=1}^{M} \lambda_m \sum_{k=1}^{K} \hat{h}_m(\hat{f}_k, \hat{e}_k, s_k) \tag{7}$$

In equation (7), $\hat{h}_m$ is a feature defined on phrase-pairs $(\hat{f}_k, \hat{e}_k)$, and λ_m is the feature weight of $\hat{h}_m$. These weights (λ_m) are optimized using minimum error-rate training (MERT) (Och, 2003) on a held-out 500 sentence-pair development set for each of the experiments.

We create a list of probable source–target term-pairs by taking each source and target term from the source and target termlists, respectively, provided that those source–target term-pairs are present in the PB-SMT phrase-table. We calculate a weight (w) for each source–target term-pair (essentially, a phrase-pair, i.e. $(\hat{e}_k, \hat{f}_k)$) using (8):[2]

$$w(\hat{e}_k, \hat{f}_k) = \sum_{m=1}^{M} \lambda_m \hat{h}_m(\hat{f}_k, \hat{e}_k) \tag{8}$$

[2]Equation (8) is derived from the right-hand side of equation (7) for a single source–target phrase-pair.

In order to calculate w, we used the four standard PB-SMT translational features ($\hat{h}_m$), namely forward phrase translation log-probability ($\log \mathrm{P}(\hat{e}_k|\hat{f}_k)$), its inverse ($\log \mathrm{P}(\hat{f}_k|\hat{e}_k)$), the lexical log-probability ($\log \mathrm{P}_{\mathrm{lex}}(\hat{e}_k|\hat{f}_k)$), and its inverse ($\log \mathrm{P}_{\mathrm{lex}}(\hat{f}_k|\hat{e}_k)$). We considered a higher threshold value for weights and considered those term-pairs whose weights exceeded this threshold. For each source term, we considered a maximum of the four highest-weighted target terms.

Domain Parallel Corpus		
Domain	Sentences	Words (English)
English-to-Spanish		
Banking, Finance and Economics	50,112	548,594
Engineering	91,896	1,165,384
IT	33,148	367,046
Tourism and Travel	50,042	723,088
Science	79,858	1,910,482
Arts and Culture	9,124	100,620
English-to-Hindi		
EILMT	7,096	173,770
EMILLE	9,907	159,024
Launchpad	67,663	380,546
KDE4	84,089	324,289
Reference Corpus		
Language	Sentences	Words
English	4,000,000	82,048,154
Spanish	4,132,386	128,005,190
Hindi	10,000,000	182,066,982

Table 1: Corpus Statistics.

4 Experiments and Discussion

4.1 Data Used

We conducted experiments on several data domains for two different language-pairs, English-to-Spanish and English-to-Hindi. For English-to-Spanish, we worked with client-provided data taken from six different domains in the form of translation memories. For English-to-Hindi, we used three parallel corpora from three different sources (EILMT, EMILLE and Launchpad) taken from HindEnCorp[3] (Bojar et al., 2014) released for the WMT14 shared translation task,[4] and a parallel corpus of KDE4 localization files[5] (Tiedemann, 2009). The EMILLE corpus contains leaflets from the UK Government and various local authorities. The domain of the EILMT[6] corpus is tourism.

We used data from a collection of translated documents from the United Nations (MultiUN)[7] (Tiedemann, 2009) and the European Parliament (Koehn et al., 2005) as the monolingual English and Spanish reference corpora. We used the HindEnCorp monolingual corpus (Bojar et al., 2014) as the monolingual Hindi reference corpus. The statistics of the data used in our experiments are shown in Table 1.

4.2 Runtime Performance

Our terminology extraction model is composed of two main processes: (i) Moses training and tuning (restricting the number of iterations of MERT to a maximum of 6), and (ii) terminology extraction. In Table 2, we report the actual runtimes of these two processes on the six domain corpora. As Table

[3]http://ufallab.ms.mff.cuni.cz/ bojar/hindencorp/

[4]http://www.statmt.org/wmt14/

[5]http://opus.lingfil.uu.se/KDE4.php

[6]English-to-Indian Language Machine Translation (EILMT) is a Ministry of IT, Govt. of India sponsored project.

[7]http://opus.lingfil.uu.se/MultiUN.php

2 demonstrates, both MT system-building (training *and* tuning combined) and terminology extraction processes are very short on each corpus. Given the crucial influence of bilingual terminology on quality in translation workflows, we believe that the creation of such assets from scratch in less than 30 minutes may prove to be a significant breakthrough for translators.

	MT System Building	Terminology Extraction
English-to-Spanish		
Banking, Finance and Economics	05:49	04:23
Engineering	06:47	04:33
IT	04:10	04:31
Tourism and Travel	05:34	04:24
Science	15:26	04:52
Arts and Culture	03:20	04:16
English-to-Hindi		
EILMT	12:41	15:47
EMILLE	05:41	17.18
Launchpad	04:37	24.11
KDE4	04:05	16:50

Table 2: Runtimes (minutes:seconds) for MT system-building and bilingual terminology extraction on the different domain data sets.

4.3 Human Evaluation

Of course, it is one thing to rapidly create translation assets such as bilingual termbanks, and another entirely to ensure the quality of such resources. Accordingly, we evaluated the performance of our bilingual terminology extraction model on each English-to-Spanish and English-to-Hindi domain corpus reported in Table 1, with the evaluation goals being twofold: (i) measuring the accuracy of the monolingual terminology extraction process, and (ii) measuring the accuracy of our novel bilingual terminology creation model.

As mentioned in Section 3.2, a source term may be aligned with up to four target terms. For evaluation purposes, we considered the top-100 source terms based on the LL values (cf. (1)) and their target counterparts (i.e. one to four target terms). The quality of the extracted terms was judged by native Spanish and Hindi speakers, both with excellent English skills, and the evaluation results are reported in Table 3. Note that we were not able to measure recall of the term extraction model on the domain corpora due to the unavailability of a reference terminology set. The evaluator counted the number of valid terms in the source term list for the domain in question, and the percentage of valid terms with respect to the total number of terms (i.e. 100) is reported in the second column in Table 3. We refer to this as VST (Valid Source Terms). For each valid source term there are one to four target terms that are ranked according to the weights in (8). In theory, therefore, the top-ranked target term is the most suitable target translation of the aligned source term. The evaluator counted the number of instances where the top-ranked target term was a suitable target translation of the source term; the percentage with respect to the number of valid source terms is shown in the third column in Table 3, and denoted as VTT (Valid Target Terms). The evaluator also reported the number of cases where any of the four target terms was a suitable translation of the source term; the percentage with respect to the number of valid source terms is given in the fourth column in Table 3. Furthermore, the evaluator counted the number of instances where any of the four target terms with minor editing can be regarded as suitable target translation; the percentage with respect to the number of valid source terms is reported in the last column of Table 3. In Table 4, we show three English–Spanish term-pairs extracted by our automatic term extractor where the target terms (Spanish) are slightly incorrect. In all these examples the edit distance between the correct term and the one proposed by our automatic extraction method is quite low, meaning that just a few keystrokes can transform

the candidate term into the correct one. In these cases editing the candidate term is much cheaper (in terms of time) than creating the translations from scratch.

	VST (%)	VTT1 (%)	VTT4 (%)	VTTME4 (%)
English-to-Spanish				
Banking, Finance and Economics	76	92.1	93.4	94.7
Engineering	84	90.5	91.7	94.1
IT	89	90.0	97.8	97.8
Tourism and Travel	72	86.1	93.1	93.1
Science	94	93.6	93.6	93.6
Arts and Culture	89	91.9	96.5	96.5
English-to-Hindi				
EILMT	91	81.3	83.5	96.7
EMILLE	79	62.1	83.5	98.7
Launchpad	88	95.4	98.8	98.8
KDE4	79	88.6	89.8	94.9

Table 3: Manual evaluation results obtained on the top-100 term pairs. VST: Valid Source Terms, VTT1: Valid Target Terms (1-best), VTT4: Valid Target Terms (4-best), VTTME4: Valid Target Terms with Minor Editing (4-best).

Source Terms (using Bilingual Term Extractor)	Target Terms	Target Terms corrected with Minor Editing	Edit Distance
Shutter	Obturación	Obturador	5
comment: wrong choice of inflection is likely caused by the term being most frequently used as 'shutter speed'			
Lenses	Objetivos EF	Objetivos	3
comment: The qualifier 'EF' should not be present in the target, as it is not in the source			
Leave Cancel	Cancelación Vacaciones	Cancelación de Vacaciones	3
comment: The preposition 'de' is missing in the target term			

Table 4: Slightly wrong target terms corrected with minor editing.

In Table 3, we see that the accuracy of the monolingual term extraction model varies from 72% to 94% for both English-to-Spanish and English-to-Hindi. For English-to-Spanish, the accuracy of our bilingual terminology creation model ranges from 86.1% to 93.6%, 91.7% to 97.8% and 93.1% to 97.8% when the 1-best, 4-best and 4-best with slightly edited target terms are considered, respectively. For English-to-Hindi, the accuracy of our bilingual terminology creation model ranges from 62.1% to 95.4%, 83.5% to 98.8% and 94.9% to 98.8% when the 1-best, 4-best and 4-best with slightly edited target terms are considered, respectively.

We are greatly encouraged by these results, as they demonstrate that our novel bilingual termbank creation method is robust in the face of the somewhat noisy monolingual term-extraction results; as a consequence, if better methods for suggesting monolingual term candidates are proposed, we expect the performance of our bilingual term-creation model to improve accordingly.

We calculated the distributions of unigram, bigram and trigram in the valid source terms (cf. Table 3) and reported in Table 5. We also calculated the percentages of their distributions in the valid source terms averaged over all 10 data sets. As can be seen from Table 3, the percentage of the average distribution of the trigram terms is quite low (i.e. 2.5%). This result justifies our decision for extraction of up to 3-gram terms.

	Unigram	Bigram	Trigram
English-to-Spanish			
Banking, Finance and Economics	55	20	1
Engineering	64	18	2
IT	75	12	2
Tourism and Travel	49	18	5
Science	91	3	0
Arts and Culture	76	10	3
English-to-Hindi			
EILMT	73	17	1
EMILLE	35	37	7
Launchpad	85	3	0
KDE4	74	5	0
Average	80.4%	17.0 %	2.5%

Table 5: Distributions of unigram, bigram and trigram in the valid source term pairs (cf. second column in Table 3).

4.4 Comparison: Monolingual Terminology Extraction

In this section we report the performance of our monolingual terminology extraction model (cf. Section 3.1) comparing with the performance of several state-of-the-art terminology extraction algorithms capable of recognising multiword terms. In order to extract monolingual multiword terms we used the JATE toolkit[8] (Zhang et al., 2008). This toolkit first extracts candidate terms from a corpus using linguistic tools and then applies term extraction algorithms to recognise terms specific to the domain corpus. The JATE toolkit is currently available only for the English language. For evaluation purposes, we considered the source-side of the English-to-Hindi domain corpora.

Algorithm	Reference	EILMT	EMILLE	Launchpad	KDE4
LLC (Bilingual)	cf. VST in Table 3	91	79	88	79
LLC		77	53	80	71
STF		46	04	54	44
ACTF		42	15	62	48
TF-IDF		50	36	45	17
Glossex	Kozakov et al. (2004)	76	43	76	71
JK	Justeson & Katz (1995)	42	13	58	42
NC-Value	Frantzi et al. (2000)	46	34	52	25
RIDF	Church & Gale (1995)	27	16	23	21
TermEx	Sclano et al. (2007)	42	08	46	41
C-Value	Ananiadou (1994)	49	44	62	40
Weirdness	Ahmed et al. (1999)	77	57	82	63

Table 6: Monolingual evaluation results. LLC: Log-Likelihood Comparison, STF: Simple Term Frequency, ACTF: Average Corpus Term Frequency, JK: Justeson Katz

For comparison, we considered the top-100 source terms based on the log-likelihood values (cf. (1)). The automatic term extraction algorithms in JATE assign weights (domain representativeness) to the candidate terms giving an indication of the likelihood of being a good domain-specific term. The quality of the extracted terms (top-100 highest weighted) was judged by an evaluator with excellent English skills, and the evaluation results are reported in Table 6. The evaluator counted the number of valid terms

[8]https://code.google.com/p/jatetoolkit/

in the highest weighted 100 terms that were extracted using different state-of-the-art term extraction algorithms.

The third row of Table 6 represents the percentage of the valid source terms extracted by our log-likelihood comparison (LLC) based monolingual term extraction algorithm. The next three rows represent three basic monolingual term extraction algorithms (STF: simple term frequency, ACTF: average corpus term frequency and TF-IDF) available in the JATE toolkit. The last seven rows represent seven state-of-the-art terminology extraction algorithms. As can be seen from Table 6, LLC is the best-performing algorithm with the Weirdness (Ahmad et al., 1999) and the Glossex (Kozakov et al., 2004) algorithms on the EILMT and the KDE4 corpora, respectively. The LLC is also the second-best performing algorithm on the EMILLE and the Lauchpad corpora.

We see in Table 6 that the percentage of valid source terms is quite low on the EMILLE corpus. This might be caused by it containing information leaflets in a variety of domains (consumer, education, housing, health, legal, social), which might bring down the percentage of valid source terms on this corpus.

Note that the percentage of valid source terms (VST) reported in Table 3 is calculated taking the top-100 source terms from the bilingual term-pair list that were extracted using the method described in Section 3.2. For comparison purposes we again report this percentage (VST in Table 3) in the second row in Table 6. Our bilingual term extraction method discards any anomalous pairs from the initial candidate term-pair list (cf. Section 3.2). This essentially removes some of the source entries that are not pertinent to the domain. As a result, the percentage of the valid source terms extracted applying our bilingual terminology extraction method (Table 3) is higher than the percentage of the valid source terms extracted applying our monolingual terminology extraction algorithm (LLC) (Table 6). We clearly see from Tables 3 and 6 that this bilingual approach to term extraction not only achieves remarkable performance on the bilingual task, but that when used in a monolingual context it outperforms most state-of-the-art extraction algorithms, and is comparable with the best ones. We should also note that JATE's implementation of these algorithms (including Weirdness) uses language-dependent modules such as a lemmatizer, unlike our implementation of LLC which is language-independent.

5 Conclusions and Future Work

In this paper we presented a bilingual multi-word terminology extraction model based on two independent consecutive processes. Firstly, we employed a log-likelihood comparison method to extract source and target terms independently from both sides of a parallel domain corpus. Secondly, we used a PB-SMT model to align source terms to one or more target terms. The manual evaluation results on ten different domain corpora of two syntactically divergent language-pairs showed the accuracy of our bilingual terminology extraction model to be very high, especially in the light of the rather noisier monolingual candidate terms presented to it. Given the reported high levels of performance – minimum levels of 93.1% and 94.9% in the 4-best set-up across all six domains for English-to-Spanish and all four domains for English-to-Hindi, respectively – we are convinced that the extracted bilingual multiword termbanks are useful 'as is', and with a small amount of post-processing from domain experts would be completely error-free.

The proposed bilingual terminology extraction model has been tested on a highly investigated language-pair, English-to-Spanish, and a less-explored and low-resourced English-to-Indic language-pair, English-to-Hindi. Interestingly, the performance of the bilingual terminology extraction model is excellent for the both language-pairs. We also tested several state-of-the-art monolingual terminology extraction algorithms including our own (log-likelihood comparison) on the source-side of the four English-to-Hindi domain data sets. According to the manual evaluation results, our monolingual multiword term extraction model proves to be the best-performing algorithm on two domain data sets and the second best-performing algorithm on the remaining two domain data sets. Our monolingual multiword terminology extraction method is clearly comparable to the state-of-the-art monolingual terminology extraction algorithms.

In this work, we considered all n-gram word sequences from the domain corpus as candidate terms.

In future work, we would like to incorporate the candidate phrasal term identification model of Deane (2005), which would omit irrelevant multiword units, and help us extend our evaluation beyond the top-100 terms. We also plan to demonstrate the impact of the created termbanks on translator productivity in a number of workflows – different language pairs, domains, and levels of post-editing – in an industrial setting.

Acknowledgements

This work was partially supported by the Science Foundation Ireland (Grant 07/CE/I1142) as part of CNGL at Dublin City University, and by Grant 610879 for the Falcon project funded by the European Commission.

References

S. Ananiadou. 1994. A methodology for automatic term recognition. In *COLING: 15th International Conference on Computational Linguistics*, pages 1034–1038.

K. Ahmad, L. Gillam and L. Tostevin. 1999. University of Surrey Participation in TREC8: Weirdness Indexing for Logical Document Extrapolation and Retrieval (WILDER). In *the Eighth Text REtrieval Conference (TREC-8)*. National Institute of Standards and Technology, Gaithersburg, MD., pp.717–724.

R. Basili, A. Moschitti, M. Pazienza and F. Zanzotto. 2001. A contrastive approach to term extraction. In *Proceedings of the 4th Conference on Terminology and Artificial Intelligence (TIA 2001).* Nancy, France, 10pp.

K. Church and W. Gale. 1995. Inverse Document Frequency (IDF): A Measure of Deviation from Poisson. In *Proceedings of the 3rd Workshop on Very Large Corpora*, pages 121–130. Cambridge, MA.

B. Daille, E. Gaussier and J-M. Langé. 1994. Towards automatic extraction of monolingual and bilingual terminology. In *COLING 94, The 15th International Conference on Computational Linguistics, Proceedings*. Kyoto, Japan, pp.515–521.

P. Deane. 2007. A nonparametric method for extraction of candidate phrasal terms. In *ACL-05: 43rd Annual Meeting of the Association for Computational Linguistics*. Ann Arbor, Michigan, USA, pp.605–613.

K. Frantzi, S. Ananiadou and H. Mima. 2000. Automatic Recognition of Multi-word Terms: the C-value/NC-value Method. *International Journal of Digital Libraries.* **3**(2): 115–130.

E. Gaussier. 1998. Flow Network Models for Word Alignment and Terminology Extraction from Bilingual Corpora. In *COLING-ACL '98, 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Proceedings of the Conference, Volume II*. Montreal, Quebec, Canada, pp.444–450.

A. Gelbukh, G. Sidorov, E. Lavin-Villa and L. Chanona-Hernandez. 2010. Automatic Term Extraction Using Log-Likelihood Based Comparison with General Reference Corpus. In *15th International Conference on Applications of Natural Language to Information Systems, NLDB 2010, Proceedings*. LNCS vol. 6177. Berlin: Springer. pp.248–255.

L. Ha, G. Fernandez, R. Mitkov and G. Corpas. 2008. Mutual bilingual terminology extraction. In *LREC 2008: 6th Language Resources and Evaluation Conference*. Marrakech, Morocco, pp.1818–1824.

T. He, T., X. Zhang and Y. Xinghuo. 2006. An Approach to Automatically Constructing Domain Ontology. In *Proceedings of the 20th Pacific Asia Conference on Language, Information and Computation, PACLIC 2006*. Wuhan, China, pp.150–157.

J. S. Justeson, and S. M. Katz. 1995. Technical terminology: some linguistic properties and an algorithm for identification in text. Natural language engineering, 1(1) 9–27.

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

P. Koehn. 2005. Europarl: a parallel corpus for statistical machine translation. In *MT Summit X: The Tenth Machine Translation Summit*. Phuket, Thailand, pp.79–86.

P. Koehn, 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 *ACL 2007, Proceedings of the Interactive Poster and Demonstration Sessions*. Prague, Czech Republic, pp.177–180.

P. Koehn, F. Och and H. Ney. 2003. Statistical Phrase-Based Translation. In *HLT-NAACL 2003: conference combining Human Language Technology conference series and the North American Chapter of the Association for Computational Linguistics conference series*. Edmonton, Canada, pp. 48–54.

L. Kozakov, Y. Park, T. H. Fin, Y. Drissi, Y. N. Doganata, and T. Cofino. 2004. Glossary extraction and knowledge in large organisations via semantic web technologies. In *Proceedings of the 6th International Semantic Web Conference and the 2nd Asian Semantic Web Conference*.

J. Kupiec. 1993. An algorithm for finding noun phrase correspondences in bilingual corpora. In *31st Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference*. Columbus, Ohio, USA, pp.17–22.

E. Lefever, L. Macken and V. Hoste. 2009. Language-Independent Bilingual Terminology Extraction from a Multilingual Parallel Corpus. In *EACL '09 Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*. Athens, Greece, pp.496–504.

F. Och. 2003. Minimum Error Rate Training in Statistical Machine Translation. In *41st Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference*. Sapporo, Japan, pp.160–167.

F. Och and H. Ney. 2002. Discriminative Training and Maximum Entropy Models for Statistical Machine Translation. In *40th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference*. Philadelphia, PA, USA, pp.295–302.

O. Bojar, V. Diatka, P. Rychlý, P. Straňák, A. Tamchyna, and D. Zeman. 2014. Hindi-English and Hindi-only Corpus for Machine Translation. In *Proceedings of the Ninth International Language Resources and Evaluation Conference (LREC'14)*. Reykjavik, Iceland.

P. Pantel and D. Lin. 2001. A Statistical Corpus-Based Term Extractor. In E. Stroulia and S. Matwin (eds.) *Advances in Artificial Intelligence, 14th Biennial Conference of the Canadian Society for Computational Studies of Intelligence, AI 2001, Ottawa, Canada, Proceedings*. LNCS vol. 2056. Berlin: Springer, pp.36–46.

P. Rayson and R. Garside. 2000. Comparing corpora using frequency profiling. In *Proceedings of the Workshop on Comparing Corpora, held in conjunction with the 38th Annual Meeting of the Association for Computational Linguistics (ACL 2000)*. Hong Kong, pp.1–6.

F. Sclano and P. Velardi. 2007. TermExtractor: a web application to learn the shared terminology of emergent web communities. In *Proceedings of the 3rd International Conference on Interoperability for Enterprise Software and Applications (I-ESA 2007)*. Funchal, Madeira Island, Portugal, pp.287–290.

J. Tiedemann. 2009. News from OPUS - A Collection of Multilingual Parallel Corpora with Tools and Interfaces. In N. Nicolov and K. Bontcheva and G. Angelova and R. Mitkov (eds.) Recent Advances in Natural Language Processing (vol V), pages 237–248, John Benjamins, Amsterdam/Philadelphia.

Z. Zhang, J. Iria, C. Brewster and F. Ciravegna. 2008. A Comparative Evaluation of Term Recognition Algorithms. In *Proceedings of The sixth international conference on Language Resources and Evaluation, (LREC 2008)*, pages 2108–2113, Marrakech, Morocco.