Using Domain-specific and Collaborative Resources for Term Translation

Mihael Arcan, Paul Buitelaar
Unit for Natural Language Processing
Digital Enterprise Research Institute
Galway, Ireland
firstname.lastname@deri.org

Christian Federmann
Language Technology Lab
German Research Center for AI
Saarbrücken, Germany
cfedermann@dfki.de

Abstract

In this article we investigate the translation of terms from English into German and vice versa in the isolation of an ontology vocabulary. For this study we built new domain-specific resources from the translation search engine Linguee and from the online encyclopedia Wikipedia. We learned that a domain-specific resource produces better results than a bigger, but more general one. The first finding of our research is that the vocabulary and the structure of the parallel corpus are important. By integrating the multilingual knowledge base Wikipedia, we further improved the translation wrt. the domain-specific resources, whereby some translation evaluation metrics outperformed the results of Google Translate. This finding leads us to the conclusion that a hybrid translation system, a combination of bilingual terminological resources and statistical machine translation can help to improve translation of domain-specific terms.

1 Introduction

Our research on translation of ontology vocabularies is motivated by the challenge of translating domain-specific terms with restricted or no additional textual context that in other cases can be used for translation improvement. For our experiment we started by translating financial terms with baseline systems trained on the EuroParl (Koehn, 2005) corpus and the JRC-Acquis (Steinberger et al., 2006) corpus. Although both resources contain a large amount of parallel data, the translations were not satisfying. To improve the translations of the financial ontology vocabulary we built a new parallel resource, which was generated using Linguee[1], an online translation query service. With this data, we could train a small system, which produced better translations than the baseline model using only general resources.

Since the manual development of terminological resources is a time intensive and expensive task, we used Wikipedia as a background knowledge base and examined articles, tagged with domain-specific categories. With this extracted domain-specific data we built a specialised English-German lexicon to store translations of domain-specific terms. These terms were then used in a pre-processing method in the decoding approach. This approach incorporates the work by Aggarwal et al. (2011), which suggests a sub-term analysis. We split the financial terms into n-grams and search for financial sub-terms in Wikipedia.

The remainder of the paper is organised like this. In Section 2 we describe related work while in Section 3 the ontology data, the training data that we used in training the language model, and the translation decoder are discussed. Section 4 presents the new resources which were used for improving the term translation. In Section 5 we discuss the results of exploiting the different resources. We conclude with a summary and give an outlook on future work in Section 6.

2 Related Work

Kerremans (2010) presents the issue of terminological variation in the context of specialised translation on a parallel corpus of biodiversity texts. He shows that a term often cannot be aligned to any term in

[1] See www.linguee.com

Proceedings of SSST-6, Sixth Workshop on Syntax, Semantics and Structure in Statistical Translation, pages 86–94,
Jeju, Republic of Korea, 12 July 2012. ©2012 Association for Computational Linguistics

the target language. As a result, he proposes that specialised translation dictionaries should store different translation possibilities or term variants.

Weller et al. (2011) describe methods for terminology extraction and bilingual term alignment from comparable corpora. In their compound translation task, they are using a dictionary to avoid out-of-domain translation.

Zesch et al. (2008) address issues in accessing the largest collaborative resources: Wikipedia and Wiktionary. They describe several modules and APIs for converting a Wikipedia XML Dump into a more suitable format. Instead of parsing the large Wikipedia XML Dump, they suggest to store the Dump into a database, which significantly increases the performance in retrieval time of queries.

Wikipedia has not only a dense link structure between articles, it has also inter-language links between articles in different languages, which was the main reason to use this invaluable collaborative resource. Erdmann et al. (2008) regarded the titles of Wikipedia articles as terminology. They assumed that two articles connected by an Interlanguage link are likely to have the same content and thus an equivalent title.

Vivaldi and Rodriguez (2010) proposed a methodology for term extraction in the biomedical domain with the help of Wikipedia. As a starting point, they manually select a set of seed words for a domain, which is used to find corresponding nodes in this resource. For cleaning their collected data, they use thresholds to avoid storing undesirable categories.

Müller and Gurevych (2008) use Wikipedia and Wiktionary as knowledge bases to integrate semantic knowledge into Information retrieval. Their models, text semantic relatedness (for Wikipedia) and word semantic relatedness (for Wiktionary), are compared to a statistical model implemented in Lucene. In their approach to Bilingual Retrieval, they use the cross-language links in Wikipedia, which improved the retrieval performance in their experiment, especially when the machine translation system generated incorrect translations.

3 Experiments

Our experiment started with an analysis of the terms in the ontology to be translated, which was stored in RDF[2] data model. These terms were used to automatically extract any corresponding Wikipedia Categories, which helped us to define more exactly the domain(s) of the ontology to be translated. The collected Categories were further used to build a domain-specific lexicon to be used for improving term translation. At the same time a new parallel corpus was built, which was also generated with the help of the ontology terms. This new data was then used to pre-process the input data for the decoder and to build a specialised training model which yielded to a translation improvement.

In this section, several types of data will be presented and furthermore the translation decoder, which has to access this data to build the training models. Section 3.1 gives an overview of the data that was used in translation. In Sections 3.2 and 3.3 we describe the data that is used to train the translation and language model. We used different parallel corpora, JRC-Acquis, EuroParl and a domain-specific corpus built from Linguee. In Section 3.4, we discuss a domain-specific lexicon, extracted from Wikipedia. In the last Section 3.5 we describe the phrase-based machine translation decoder Moses that we used for our experiments.

3.1 xEBR Dataset

For the translation dataset a financial ontology developed by the XBRL European Business Registers[3] (xEBR) Working Group was used. This financial ontology is a framework for describing financial accounting and profile information of business entities across Europe, see also Declerck et al. (2010). The ontology holds 263 concepts and is partially translated into German, Dutch, Spanish, French and Italian. The terms in each language are aligned via the SKOS[4] Exact Match mechanism to the xEBR core taxonomy. In this partially translated taxonomy, we identified 63 English financial terms and their German equivalents, which were used as reference translations in evaluating the different experiment steps.

The xEBR financial terms are not really terms from a linguistic point of view, but they are used in financial or accounting reports as unique finan-

[2] RDF: Resource Description Framework

[3] XBRL: eXtensible Business Reporting Language

[4] SKOS: Simple Knowledge Organization System

Length	Count	Examples
11	1	Taxes Remuneration And Social Security Payable After More Than One Year
10	2	Amounts Owed To Credit Institutions After More Than One Year, Variation In Stocks Of Finished Goods And Work In Progress
		. . .
2	57	Net Turnover, Liquid Assets, . . .
1	10	Assets, Capital, Equity, . . .

Table 1: Examples of xEBR terms

cial expressions or tags to organize and retrieve automatically reported information. Therefore it is important to translate these financial terms exactly.

Table 1 illustrates the structure of xEBR terms. It is obvious that they are not comparable to general language, but instead are more like headlines in newspapers, which are often short, very informative and written in a telegraphic style. xEBR terms are often only noun phrases without determiners. The length of the financial terms varies, e.g. the longest financial term considered for translation has a length of 11 tokens, while others may consist of 1 or 2.

3.2 General Resources: EuroParl and JRC-Acquis

As a baseline, the largest available parallel corpora were used: EuroParl and the JRC-Acquis parallel corpus. The EuroParl parallel corpus holds the proceedings of the European Parliament in 11 European languages. The JRC-Acquis corpus is available in almost all EU official languages (except Irish) and is a collection of legislative texts written between 1950 and today.

Although research work proved, that a training model built by using a general resource cannot be used to translate domain-specific terms (Wu et al., 2008), we decided to train a baseline model on these resources to illustrate any improvement steps from a general resource to specialised domain resources.

3.3 Domain Resource: Linguee

Linguee is a combination of a dictionary and a search engine, which indexes around 100 Million bilingual texts on words and expressions. Linguee search results show example sentences that depict how the searched expression has been translated in context.

In contrast to translation engines like Google Translate and Bing Translator, which give you the most probable translation of a source text, every entry in the Linguee database has been translated by humans. The bilingual dataset was gathered from the web, particularly from multilingual websites of companies, organisations or universities. Other sources include EU documents and patent specifications.

The language pairs available for querying are English↔German, English↔Spanish, English↔French and English↔Portuguese.

Since Linguee includes EU documents, they also use parallel sentences from EuroParl and JRC-Acquis. We investigated the proportion of sentences returned by Linguee which are contained in EuroParl or JRC-Acquis. The outcome is that the number of sentences is very low, where 131 sentences (0.54%) are gathered from JRC-Acquis corpus and 466 (1.92%) from EuroParl.

3.4 Collaborative Resource: Wikipedia

Wikipedia is a multilingual, freely available encyclopedia that was built by a collaborative effort of voluntary contributors. All combined Wikipedias hold approximately 20 million articles or more than 8 billion words in more than 280 languages. With these facts it is the largest collection of freely available knowledge[5].

With the heavily interlinked information base, Wikipedia forms a rich lexical and semantic resource. Besides a large amount of articles, it also holds a hierarchy of Categories that Wikipedia Articles are tagged with. It includes knowledge about named entities, domain-specific terms and word senses. Furthermore, the redirect system of Wikipedia articles can be used as a dictionary for synonyms, spelling variations and abbreviations.

3.5 Translation System: Moses

For generating translations from English into German and vice versa, the statistical translation toolkit Moses (Koehn et al., 2007) was used to build the training model and for decoding. For this approach, a phrase-based approach was taken instead of a tree based model. Further, we aimed at improving the translations only on the surface level, and therefore no part-of-speech information was taken into account. Word and phrase alignments were built with

[5] http://en.wikipedia.org/wiki/Wikipedia:Size_comparison

the GIZA++ toolkit (Och and Ney, 2003), whereby the 5-gram language model was built by SRILM (Stolcke, 2002).

4 Domain-specific Resource Generation

In this section, two different types of data and the approach of building them will be presented. Section 4.1 gives an overview of generating a parallel resource from Linguee, which was used in generating a new domain-specific training model. In Section 4.2 a detailed description is given how we extracted terms from Wikipedia for generating a domain-specific lexicon.

4.1 Domain-specific parallel corpus generation

To build a new training model that is specialised on our xEBR ontology, we used the Linguee search engine. This resource can be queried on single words and on word expressions with or without quotation marks. We stored the HTML output of the Linguee queries on our financial terms and parsed these files to extract plain parallel text. From this, we built a financial parallel corpus with 13,289 translation pairs, including single words, multi-word expressions and sentences. The English part of the parallel resource contained 410,649 tokens, the German part 347,246.

4.2 Domain-specific lexicon generation

To improve translation based on the domain-specific parallel corpus, we built a cross-lingual terminological lexicon extracted from Wikipedia. From the Wikipedia Articles we used different information units, i.e. the Title of a Wikipedia Article, the Category (or Categories) of the Title and the internal Interwiki
Interlanguage links of the Title. The concept of Interwiki links can be used to make links to other Wikipedia Articles in the same language or to another Wikipedia language i.e. Interlanguage links.

In our first approach, we used Wikipedia to determine the domain (or several domains) of the ontology. This approach (a) is to understand as the identification of the domain through the vocabulary of the ontology. For this approach, the financial terms, which were extracted from the ontology, were used to query the Wikipedia knowledge base[6]. The Wikipedia Article was considered for further examination, if its Title is equivalent to our financial terms. In this first step, 7 terms of our ontology were identified in the Wikipedia knowledge base. With this step, we collected the Categories of these Titles, which was the main goal of this approach. In a second round, we split all financial terms into all possible n-grams and repeated the query again to find additional Categories based on the split n-grams. Table 2 shows the collected Categories of the first approach and how often they appeared in respect to the extracted financial terms.

Collected Wikipedia Categories	
Frequency	Name
8	Generally Accepted Accounting Principles
4	Debt
4	Accounting terminology
	…
1	Political science terms
1	Physical punishments

Table 2: Collected Wikipedia Categories based on the extracted financial terms

After storing all Categories, only such Categories were considered, which frequency had a value more than the calculated arithmetic mean of all frequencies (> 3.15). For the calculation of the arithmetic mean only Categories were considered, which had a frequency more than 1, since 2,262 of 3,615 collected Categories (62.6%) had a frequency equals 1. With this threshold we avoided extraction of a vocabulary that is not related to the ontology. Without this threshold, out-of-domain Categories would be stored, which would extend the lexicon with vocabulary that would not benefit the ontology translation, e.g. *Physical punishments*, which was access by the financial term *Stocks*.

In the next step, we further extended the list of Categories collected previously by use of full and split terms. This was done by storing new Categories based on the Wikipedia Interwiki links of each Article which was tagged with a Category from Table 2. For example, we collected all Categories wherewith the Article *Balance sheet*[7] is tagged and the Categories of the 106 Interwiki links of the Article *Balance sheet*. The frequencies of these Categories were summed up for all Interwiki links. Finally a

[6] For the Wikipedia Query we used the Wikipedia XML dump; `enwiki-20120104-pages-articles.xml`

[7] Financial statements, Accounting terminology

Final Category List	
Frequency	Name
95	Economics terminology
62	Generally Accepted Accounting Principles
61	Macroeconomics
55	Accounting terminology
47	Finance
44	Economic theories
	...

Table 3: Most frequent Categories based on the xEBR terms and their Interwiki links

new Category was added to the final Category list, if the new Category frequency exceeds the arithmetic mean threshold (> 18.40).

The final Category list contained 33 financial Wikipedia Categories (Table 3), which was in the next step used for financial term extraction.

With the final list of Categories, we started an investigation of all Wikipedia articles tagged with these financial Categories. Each Wikipedia Title was considered as a useful domain-specific term and was stored in our lexicon if a German title in the Wikipedia knowledge base also existed. As an example, we examined the Category Accounting terminology and stored the English Wikipedia Title *Balance sheet* with the German equivalent Wikipedia Title *Bilanz*.

At the end of the lexicon generation we examined 5228 Wikipedia Articles, which were tagged with one or more financial Categories. From this set of Articles we were able to generate a terminological lexicon with 3228 English-German entities.

5 Evaluation

Tables 4 to 5 illustrate the final results for our experiments on translating xEBR ontology terms, using the NIST (Doddington, 2002), BLEU (Papineni et al., 2002), and Meteor (Lavie and Agarwal, 2005) algorithms. To further study any translation improvements of our experiment, we also used Google Translate[8] in translating 63 financial xEBR terms (cf. Section 3.1) from English into German and from German into English.

		Scoring Metric		
Source	# correct	BLEU	NIST	Meteor
Google Translate	18	0.264	4.382	0.369
JRC-Acquis	12	0.167	3.598	0.323
EuroParl	4	0.113	2.630	0.326
Linguee	25	0.347	4.567	0.408
Lexical substitution	4	0.006	0.223	0.233
Linguee+Wiki	25	0.324	4.744	0.432

Table 4: Evaluation scores for German term translations

		Scoring Metric		
Source	# correct	BLEU	NIST	Meteor
Google Translate	21	0.452	4.830	0.641
JRC-Acquis	9	0.127	2.458	0.480
EuroParl	5	0.021	1.307	0.412
Linguee	15	0.364	3.938	0.631
Lexical substitution	4	0.006	0.243	0.260
Linguee+Wiki	22	0.348	3.993	0.644

Table 5: Evaluation scores for English term translations

5.1 Interpretation of Evaluation Metrics

In our experiments translation models built from a general resource performed worst. These results show that building resources from general language does not improve the translation of terms. The Linguee financial corpus, which is built from 13,289 sentences and holds 304K English and German 250K words, however demonstrates the benefit of domain-specific resources. Its size is less than two percent of that of the JRC-Acquis corpus (1,131,922 sentences, 21M English words, 19M German words), but evaluation scores are more than double than those for JRC-Acquis. This is clear evidence that such a resource benefits the translation of terms in a specific domain.

The models produced by the Linguee search engine are generating better translations than those produced by general resources. This approach outperforms Google Translate translations from German into English for all used evaluation metrics.

The table further shows results for our approach in using extracted Wikipedia terms as an example-based approach. For this we used the terms extracted from Wikipedia and exchanged English terms with German translations and vice versa. The evaluation metrics are very low in this case; only for Correct Translation we generate four positive findings.

Finally, the table gives results for our approach in using a combination of domain-specific parallel financial corpus with the lexicon extracted from Wikipedia. The domain-specific lexicon contains 3228 English-German translations, which were extracted from 18 different financial Categories. This

[8] Translations were generated on February 2012.

combination of highly specialised resources gives the best results in our experiment. Translating financial terms into German, we get more Correct Translations as well as the Meteor metric shows better results compared to Google Translate. For translations into English, all used evaluation metrics show better results than those of Google Translate. As a final observation, we learned that translations made by domain-specific resources are on the same quality level, either if we translate from English into German or vice versa. In comparison, we see that Google Translate has a larger discrepancy when translating into German or English respectively. Our research showed that translations from English into German built by specialised resources were slightly better, which goes along with Google Translate that also produces better translations into German.

5.2 Manual Evaluation of Translation Quality

In addition to the automatic evaluation with BLEU, NIST, and Meteor scores, we have also undertaken a manual evaluation campaign to assess the translation quality of the different systems. In this section, we will a) describe the annotation setup and task presented to the human annotators, b) report on the translation quality achieved by the different systems, and c) present inter-annotator agreement scores that allow to judge the reliability of the human rankings.

5.2.1 Annotation Setup

In order to manually assess the translation quality of the different systems under investigation, we designed a simple classification scheme consisting of three distinct classes:

1. *Acceptable (A)*: terms classified as acceptable are either fully identical to the reference term or semantically equivalent;
2. *Can easily be fixed (C)*: terms in this class require some minor correction (such as fixing of typos, removal of punctuation, etc.) but are nearly acceptable. The general semantics of the reference term are correctly conveyed to the reader.
3. *None of both (N)*: the translation of the term does not match the intended semantics or it is plain wrong. Items in this class are considered severe errors which cannot easily be fixed and hence should be avoided wherever possible.

	Classes		
System	A	C	N
Linguee+Wiki	58%	27%	15%
Google Translate	55%	31%	14%
Linguee	51%	37%	12%
JRC-Acquis	32%	28%	40%
EuroParl	5%	25%	70%

Table 6: Results from the manual evaluation into German

	Classes		
System	A	C	N
Linguee+Wiki	56%	32%	12%
Linguee	56%	31%	13%
Google Translate	39%	40%	21%
JRC-Acquis	39%	31%	30%
EuroParl	15%	30%	55%

Table 7: Results from the manual evaluation into English

5.2.2 Annotation Data

We setup ten evaluation tasks, five for translations into English, five for translations into German. Each of these sets was comprised of 63 term translations and the corresponding reference. Every set was given to at least three human annotators who then classified the observed translation output according to the classification scheme described above. The human annotators included both domain experts and lay users without knowledge of the terms domain.

In total, we collected 2,520 classification items from six annotators. Tables 6, 7 show the results from the manual evaluation for term translations into German and English, respectively. We report the distribution of classes per evaluation task which are displayed in *best-to-worst* order.

In order to better be able to interpret these rankings, we computed the inter-annotator agreement between human annotators. We report scores generated with the following agreement metrics:

- S (Bennet et al., 1954);
- π *(averaged across annotators)* (Scott, 1955);
- κ (Fleiss and others, 1971);
- α (Krippendorff, 1980).

Tables 8, 9 present the aforementioned metrics scores for German and English term translations.

Overall, we achieve an average κ score of 0.463, which can be interpreted as *moderate agreement* following (Landis and Koch, 1977). Notably, we also reach *substantial* agreement for one of the annotation tasks with a κ score of 0.657. Given the

	Agreement Metric			
System	S	π	κ	α
Linguee+Wiki	0.599	0.528	0.533	0.530
Google Translate	0.698	0.655	0.657	0.657
Linguee	0.484	0.416	0.437	0.419
JRC-Acquis	0.412	0.406	0.413	0.408
EuroParl	0.515	0.270	0.269	0.273

Table 8: Annotator agreement scores for German

	Agreement Metric			
System	S	π	κ	α
Linguee+Wiki	0.532	0.452	0.457	0.454
Linguee	0.599	0.537	0.540	0.539
Google Translate	0.480	0.460	0.465	0.463
JRC-Acquis	0.363	0.359	0.366	0.360
EuroParl	0.552	0.493	0.499	0.495

Table 9: Annotator agreement scores for English

observed inter-annotator agreement, we expect the reported ranking results to be meaningful. Our Linguee+Wiki system performs best for both translation directions while out-of-domain systems such as JRC-Acquis and EuroParl perform badly.

5.3 Manual error analysis

Table 10 provides a manual analysis of the provided translations from Google Translate and the combined Linguee and Wikipedia Lexicon approach. Example Ex. 1 shows the results for [*Other intangible*] *fixed assets*. Since both translating systems translate it the same, namely *Vermögenswerte*, they could be considered as term variants.

A similar example is [*Receivables and other*] *assets* in Ex. 4. Google Translate translates the segment *asset* into *Vermögensgegenstände*, whereby the domain-specific approach translates it into *Vermögenswerte*. These examples prove the research by Kerremans (2010) that one term does not necessarily have only one translation on the target side. As term variants can further be considered *Aufwendungen* and *Kosten*, which were translated from *Costs* [*of old age pensions*] (Ex. 5).

In contrast, the German term in [*sonstige betriebliche*] *Aufwendungen* (Ex. 8) is according to the xEBR translated into [*Other operating*] *expenses*, which was translated correctly by both systems.

A deeper terminological analysis has to be done in the translation of the English term [*Cost of*] *old age pensions* (Ex. 5). In general it can be translated into *Altersversorgung* (provided by Google Translate and xEBR) or *Altersrente* (generated by the domain-specific model). Doing a compound analysis, the translation of [*Alters*]*versorgung* is *supply* or *maintenance*. On the other side, the translation of [*Alters*]*rente* is pension, which has a stronger connection to the financial term in this domain.

Ex. 6 shows an improvement of domain specific translation model in comparison to a general resource. Both general resources translated *Securities* as *Sicherheiten*, which is correct but not in the financial domain. The domain-specific trained model translates the ambiguous term correctly, namely *Wertpapiere*. Google Translate generates the same term as on the source site, *Securities*. Further, the term *Equity* (Ex. 7) is translated by Google Translate as *Gerechtigkeit*, the domain-specific model translates it as *Eigenkapital*, which is the correct translation. Finally, Ex. 2 and Ex. 3 open the issue of accurateness of the references for translation evaluation. The translations of these terms are correct if we consider the source language. On the other hand, if we compare them with the proposed references, they are not the same. In Ex. 2 they are truncated or extended in Ex. 3, which opens up problems in translation evaluation.

5.4 Discussion

Our approach shows the differences between improving translations with different resources. It was shown to be necessary to use additional language resources, i.e. specialised parallel corpora and if available, specialised lexica with appropriate translations. Nevertheless, to move further in this direction, translation of specific terms, more research is required in several areas that we identified in our experiment. One is the quality of the translation model. Because the translation model can only translate terms that are in the training model, it is necessary to use a domain-specific resource. Although we got better results with a smaller resource (if we translate into English), comparing those results with Google Translate, we learned that more effort has to be done in the direction of extending the size and quality of domain-specific resources.

Apart from that, with the aid of Wikipedia, which can be easily adapted for other language pairs, we further improved the translations into English to a

	Term		Translations	
#	Source	Reference	Google	Domain-specific
1	Other intangible fixed assets	sonstige immaterielle Vermögensgegenstände	Sonstige immaterielle Vermögenswerte	Sonstige immaterielle Vermögenswerte
2	Long-term financial assets	Finanzanlagen	Langfristige finanzielle Vermögenswerte	Langfristige finanzielle Vermögenswerte
3	Financial result	Finanz- und Beteiligungsergebnis	Finanzergebnis	Finanzergebnis
4	Receivables and other assets	Forderungen und sonstige Vermögensgegenstände	Forderungen und sonstige Vermögensgegenstände	Forderungen und sonstige Vermögenswerte
5	Cost of old age pensions	Aufwendungen für Altersversorgung	Aufwendungen für Altersversorgung	Kosten der Altersrenten
6	Securities	Wertpapiere	Securities	Wertpapiere
7	Equity	Eigenkapital	Gerechtigkeit	Eigenkapital
8	sonstige betriebliche Aufwendungen	Other operating expenses (TC)	other operating expenses	other operating expenses

Table 10: Translations provided by Google Translate and by the domain-specific resource

point where we outperform translations provided by Google Translate. Nevertheless, our experiment showed that the translations into German were better in regard of Google translate only for the Meteor evaluation system, for BLEU and NIST we did not achieve significant improvements. Also here more work has to be done in domain adaptation in a more sophisticated way to avoid building out-of-domain vocabulary.

6 Conclusion

The approach of building new resources showed a large impact on the translation quality. Therefore, generating specialised resources for different domains will be the focus of our future work. On the one hand, building appropriate training models is important, but our experiment also highlighted the importance of additional collaborative resources, like Wikipedia, Wiktionary, and DBpedia. Besides extracting Wikipedia Articles with their multilingual equivalents, as shown in Section 4.2, Wikipedia holds much more information in the articles itself. Therefore exploiting non-parallel resources, shown by Fišer et al. (2011), would clearly help the translation system to improve performance. Future work needs to better include the redirect system, which would allow a better understanding of synonymy and spelling variety of terms.

Focusing on translating ontologies, we will try to better exploit the structure of the ontology itself. Therefore, more work has to be done in the combination of linguistic and semantic information (structure of an ontology) as demonstrated by Aggarwal et al. (2011), which showed first experiments in combining semantic, terminological and linguistic information. They suggest that a deeper semantic analysis of terms, i.e. understanding the relations between terms and analysing sub-terms needs to be considered. Another source of useful information may be found in using existing translations for improving the translation of other related terms in the ontology.

Acknowledgments

This work has been funded under the Seventh Framework Programme for Research and Technological Development of the European Commission through the T4ME contract (grant agreement no.: 249119) and in part by the European Union under Grant No. 248458 for the Monnet project as well as by the Science Foundation Ireland under Grant No. SFI/08/CE/I1380 (Lion-2). The authors would like to thank Susan-Marie Thomas, Tobias Wunner, Nitish Aggarwal and Derek De Brandt for their help with the manual evaluation. We are grateful to the anonymous reviewers for their valuable feedback.

References

Nitish Aggarwal, Tobias Wunner, Mihael Arcan, Paul Buitelaar, and Seán O'Riain. 2011. A similarity measure based on semantic, terminological and linguistic information. In *The Sixth International Workshop on Ontology Matching collocated with the 10th International Semantic Web Conference (ISWC'11)*.

E. M. Bennet, R. Alpert, and A. C. Goldstein. 1954. Communications through limited response questioning. *Public Opinion Quarterly*, 18:303–308.

Thierry Declerck, Hans-Ulrich Krieger, Susan M. Thomas, Paul Buitelaar, Sean O'Riain, Tobias Wunner, Gilles Maguet, John McCrae, Dennis Spohr, and Elena Montiel-Ponsoda. 2010. Ontology-based multilingual access to financial reports for sharing business knowledge across europe. In *Internal Financial Control Assessment Applying Multilingual Ontology Framework*.

George Doddington. 2002. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proceedings of the second international conference on Human Language Technology Research*, HLT '02, pages 138–145.

M. Erdmann, K. Nakayama, T. Hara, and S. Nishio. 2008. An approach for extracting bilingual terminology from wikipedia. *Lecture Notes in Computer Science*, (4947):380–392. Springer.

Darja Fišer, Špela Vintar, Nikola Ljubešić, and Senja Pollak. 2011. Building and using comparable corpora for domain-specific bilingual lexicon extraction. In *Proceedings of the 4th Workshop on Building and Using Comparable Corpora: Comparable Corpora and the Web*, BUCC '11, pages 19–26.

J.L. Fleiss et al. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382.

Koen Kerremans. 2010. A comparative study of terminological variation in specialised translation. In *Reconceptualizing LSP Online proceedings of the XVII European LSP Symposium 2009*, pages 1–14.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. In *Proceedings of the 45th Annual Meeting of the ACL*, ACL '07, pages 177–180.

Philipp Koehn. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. In *Conference Proceedings: the tenth Machine Translation Summit*, pages 79–86. AAMT.

Klaus Krippendorff. 1980. *Content Analysis: An Introduction to Methodology*. Sage Publications, Inc.

J. R. Landis and G. G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.

Alon Lavie and Abhaya Agarwal. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the EMNLP 2011 Workshop on Statistical Machine Translation*, pages 65–72.

Christof Müller and Iryna Gurevych. 2008. Using wikipedia and wiktionary in domain-specific information retrieval. In *Working Notes for the CLEF 2008 Workshop*.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 311–318.

W. A. Scott. 1955. Reliability of Content Analysis: The Case of Nominal Scale Coding. *Public Opinion Quarterly*, 19:321–325.

Ralf Steinberger, Bruno Pouliquen, Anna Widiger, Camelia Ignat, Tomaz Erjavec, Dan Tufis, and Dniel Varga. 2006. The jrc-acquis: A multilingual aligned parallel corpus with 20+ languages. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'2006)*.

Andreas Stolcke. 2002. Srilm-an extensible language modeling toolkit. In *Proceedings International Conference on Spoken Language Processing*, pages 257–286.

Jorge Vivaldi and Horacio Rodriguez. 2010. Using wikipedia for term extraction in the biomedical domain: first experiences. *Procesamiento del Lenguaje Natural*, 45:251–254.

Marion Weller, Anita Gojun, Ulrich Heid, Béatrice Daille, and Rima Harastani. 2011. Simple methods for dealing with term variation and term alignment. In *Proceedings of the 9th International Conference on Terminology and Artificial Intelligence*, pages 87–93.

Hua Wu, Haifeng Wang, and Chengqing Zong. 2008. Domain adaptation for statistical machine translation with domain dictionary and monolingual corpora. In *Proceedings of the 22nd International Conference on Computational Linguistics - Volume 1*, COLING '08, pages 993–1000.

Torsten Zesch, Christof Müller, and Iryna Gurevych. 2008. Extracting lexical semantic knowledge from wikipedia and wiktionary. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*.