Experiments with Term Translation

Mihael ARCAN[1] *Christian FEDERMANN*[2] *Paul BUITELAAR*[1]
(1) Unit for Natural Language Processing, Digital Enterprise Research Institute
National University of Ireland Galway
Galway, Ireland
(2) Language Technology Lab, German Research Center for AI
Saarbrücken, Germany
mihael.arcan@deri.org, paul.buitelaer@deri.org,
cfedermann@dfki.de

Abstract

In this article we investigate the translation of financial terms from English into German in the isolation of an ontology vocabulary. For this study we automatically built new domain-specific resources from the translation search engine Linguee and from the online encyclopaedia Wikipedia. Due to the fact that we performed the translation approach on a monolingual ontology, we ran several sub-experiments to find the most appropriate model to translate the financial vocabulary. The findings from these experiments lead 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. Finally we undertook a manual cross-lingual evaluation on the monolingual ontology to get a better understanding on this specific short text translation task.

Keywords: Ontologies and terminology, Empirical machine translation.

Proceedings of COLING 2012: Technical Papers, pages 67–82,
COLING 2012, Mumbai, December 2012.

1 Introduction

Our research on the 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 may be used to improve the translation. For our experiment we started by translating financial terms with the baseline systems trained on the JRC-Acquis (Steinberger et al., 2006) corpus and the European Central Bank Corpus (Tiedemann, 2009). Although both resources contain a large amount of parallel data, the translations were not satisfactory. To improve the translations of the financial ontology vocabulary we built a new parallel resource, which was generated using Linguee, an online translation query service. With this data, we could train a small model, 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 the 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), where the authors use the ontology structure to calculate the similarity between the labels. They combine the semantic, terminological and linguistic information for monolingual ontology matching, which can be extended to the multilingual scenario. We split the financial terms into n-grams and queried for financial sub-terms in Wikipedia, which we used to query Wikipedia.

The remainder of the paper is organised as follows: In Section 2 we give an overview on the related work. In Section 3 we describe the ontology and the existing parallel resources, which were used for generating the translation and language model. Section 4 presents the new resources which were used for improving the term translation. Furthermore we discuss the results of exploiting the different resources. We conclude with a summary and give an outlook on future work.

2 Related Work

The related research focusses on different aspects relevant to our work: domain-specific term translation. Firstly we have to understand the structure of these specific terms and the variations which come when dealing with these terms. Kerremans (2010) discusses in detail 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 the target language. As a result, he proposes that specialised translation dictionaries should store different translation possibilities or term variants. In addition to that, Weller et al. (2011) describe methods for terminology extraction and bilingual term alignment from comparable corpora. In their compound translation task, they use a dictionary to avoid out-of-domain translation. In contrast, to address this problem, which frequently arises in domain-specific translation we decided to generate our own customised lexicon; which we constructed from the multilingual Wikipedia and its dense inter-article link structure.

Erdmann et al. (2008) also extracted terms from Wikipedia articles; however, they assumed that two articles connected by an Interlanguage link are likely to have the same content and thus an equivalent title. We likewise build a lexicon from Wikipedia, but instead of collecting all of the titles from Wikipedia, we target only the domain-specific titles and their translated equivalents. 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 selected a set of seed words for a domain, which were then used to find the corresponding nodes in this resource. For cleaning their collected data, they used thresholds to avoid storing undesirable categories. Müller and Gurevych (2008) used

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. Zesch et al. (2008) address the 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.

3 Experimental Data

We are investigating the problem of translating a domain-specific vocabulary, therefore our experiments started with an analysis of the financial terms stored in the investigated ontology. With these extracted terms we built different multilingual resources, which were used for financial term translation. Firstly, we used the encyclopaedia Wikipedia, where we extracted the titles from domain-specific Wikipedia articles. Secondly, we used the same financial labels to build a parallel resource for the financial domain. For this approach we used the Linguee Web service.

In this section, we present several types of data. Section 3.1 gives an overview of the data that was used in translation. In Sections 3.2 and 3.3 we describe existing multilingual resources, which were used to train the translation and language model. For our current research we used JRC-Acquis and the European Central Bank (ECB) corpus, respectively. In the end we describe the procedure to obtain domains-specific resources by Linguee 3.4 and Wikipedia 3.5.

3.1 The Financial Ontology

For our study we used the UK GAAP[1] financial ontology, prepared by the XBRL[2] European Business Registers (xEBR) Working Group. 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 142 concepts and is partially aligned into German, Dutch, Spanish, French and Italian. We identified only 16 English financial terms and their German equivalents, which were used as reference translations for automatic evaluation.

The financial terms are not really terms from a linguistic point of view, but they are used in financial or accounting reports as unique financial 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 any determiner. 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 (Figure 1).

[1] GAAP - Generally Accepted Accounting Practice

[2] XBRL - eXtensible Business Reporting Language, `http://www.xbrl.org/`

Term Length	Term Examples
11	Taxes Remuneration And Social Security Payable After More Than One Year
10	Amounts Owed To Credit Institutions After More Than One Year . . .
	. . .
2	Net Turnover, Liquid Assets, Income Taxes, Financial Charges . . .
1	Assets, Capital, Equity, Securities, Charges, Balance, Capital, Reserves . . .

Table 1: Examples for financial labels in the UK GAAP

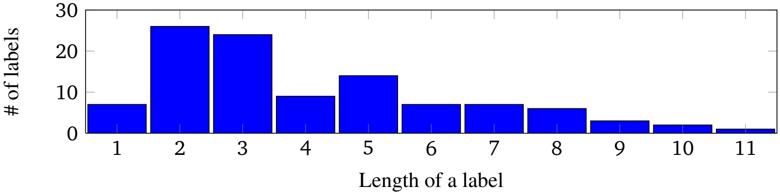

Figure 1: Label length of the UK GAAP ontology

3.2 JRC-Acquis

The general parallel corpus JRC-Acquis[3] was used as baseline training data. This corpus is available in almost every EU official language (except Irish), and is a collection of legislative texts written between 1950 and now.

Although previous research showed, 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 evaluate the translations on these resources to illustrate any improvement steps from a general resource to specialised domain resources.

3.3 European Central Bank Corpus

For comparison with JRC-Acquis, we also did experiments using the European Central Bank Corpus[4], which contains a financial vocabulary. The multilingual corpus is generated by extracting the website and documentation from the European Central Bank and is aligned among 19 European languages. For our research we used the English-German language pair, which consists of 113,171 sentence pairs or 2.8 million English and 2.5 million German tokens.

3.4 Linguee - Dictionary and Translation Search Engine

Alongside these existing resources, we built a new parallel resource based on the ontology vocabulary that we want to translate. Therefore we used Linguee,[5] a combination of a dictionary and a search engine, which indexes around 100 million bilingual texts on words and expressions. The search results show example sentences that depict how the searched expression has been translated in context. The bilingual dataset was gathered from the web, particularly from multilingual websites of companies, organisations or universities. Other sources include EU documents and patent

[3] http://langtech.jrc.it/JRC-Acquis.html
[4] http://opus.lingfil.uu.se/ECB.php
[5] http://www.linguee.com/

specifications. Since Linguee includes EU documents, they also use parallel sentences from JRC-Acquis, whereby the proportion of sentences returned by Linguee is very low, only 131 sentences or 0.54% overlap with the corpus.

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 was translated manually.

Domain-specific parallel corpus generation

To build a new training model that is specialised for 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 of our financial terms and parsed these files to extract plain parallel text. From this, we built a financial parallel corpus with 24,247 translation pairs, including single words, multi-word expressions and sentences (Table 2). The English part of the parallel resource contained 1,032,676 tokens and the German part 865,460.

Single terms	Enterprise, share, reserve, debtor, expenses, ...
Multi-words	at a specific amount, credit institute, in the amount of, doubled over the last year
Sentences	Finally, the European Parliament called for social and cultural aspects of immigration to receive equal treatment than economic and security aspects of the issue.

Table 2: Examples of extracted text from the translation search engine Linguee

3.5 Wikipedia

Wikipedia[6] is a multilingual, freely available encyclopaedia that was built by a collaborative effort of voluntary contributors. All combined Wikipedias hold approximately 19 million articles or more than 8 billion words in more than 270 languages, making it the largest collection of freely available knowledge.[7]

With the heavily interlinked information base, Wikipedia forms a rich lexical and semantic resource. Besides a large number 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.

Domain-specific lexicon generation

To improve translations, based on the domain-specific parallel corpus, we built a cross-lingual terminological lexicon. From the Wikipedia articles we used different information units: the title, 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. The domain-specific lexicon was generated by two approaches:

a) domain detection of the ontology (bottom-up approach);

b) extraction of cross-lingual terminology (top-down approach).

[6] http://www.wikipedia.org

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

In our first approach, we used Wikipedia to determine the domain (or several domains) of the ontology. The bottom-up approach (a) is to represent this domain by the most frequent categories associated with the vocabulary we want to translate. For this approach, the financial terms, which were extracted from the ontology, were used to query the Wikipedia knowledge base.[8] Initially a Wikipedia article was considered for further examination if its title is equivalent to our financial terms. In this first step, 7 terms from our ontology were identified in the Wikipedia knowledge base, i.e.:

Income tax, Earnings before interest and taxes, Asset, Stocks, Debtor, Gross profit, Income

We then collected the categories of the articles associated with these titles. Since a category can appear with different financial term, we also stored the frequency of these categories.[9] In a second round, we split our financial terms into all possible n-grams and repeated the query again to find additional categories based on the split n-grams. Table 3 shows the collected categories of the first approach and how often they appeared with respect to the extracted financial terms.

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

Table 3: Collected Wikipedia Categories based on the extracted financial terms

After storing all categories, the only categories considered were the ones that had a frequency value more than the calculated arithmetic mean of all the frequencies (> 3.15). For the calculation of the arithmetic mean only the categories that had a frequency larger than 1 were considered, since 2,262 of 3,615 collected categories (62.6%) had a frequency of 1. Using this threshold we avoided the 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 a category associated with the financial term *Stocks*.

In the next step, we further extended the list of the previous collected categories with the 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 3. For example, we collected all categories of the Wikipedia article *Balance sheet*.[10] In addition to that, we examined all Interwiki links of the article *Balance sheet* and also stored the categories of articles which have an incoming link from this article.[11] For example, we stored all categories of the 106 articles which are linked with the article *Balance sheet*. The frequencies of these categories were summed up again to re-calculate the geometric mean. Finally a new category was added to the final category list, if the new category frequency exceeds the arithmetic mean threshold (> 18.40).

[8] For the Wikipedia Query we used the Wikipedia XML dump; `enwiki-20120702-pages-articles`

[9] The Wikipedia titles *Operating Income*, *Income*, *Gross profit*, *Income statement*, *Debtor* ... are tagged with the category *Generally Accepted Accounting Principles*

[10] *Financial statements*, *Accounting terminology*

[11] *Balance sheet*

Final Category List	
Frequency	Name
95	Economics terminology
62	Generally Accepted Accounting Principles
61	Macroeconomics
	...

Table 4: Most frequent Categories based on the xEBR terms and their Interwiki links

The final category list contained 33 financial Wikipedia categories (Table 4), which were used to extract the financial terms and their translations.

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 5,228 Wikipedia articles that were tagged with one or more financial categories. From this set of articles we were able to generate a terminological lexicon with 3,228 English-German entities. The difference between the number of examined titles and the lexicon items is attributed to the fact that not all English Wikipedia titles are linked to a German one. These translation pairs were used to suggest the SMT system to choose the extracted translations by annotating the decoder input using the XML input markup scheme.

4 Experiments and Evaluation

Since the UK GAAP is a monolingual ontology, it holds no reference translation needed for automatic evaluation. Therefore we performed several experiments to find the best approach to translate this financial ontology. For decoding, we used the Moses Toolkit, with its standard settings (Section 4.1). If reference translations were available, we undertook an automatic evaluation using the BLEU (Papineni et al., 2002), NIST (Doddington, 2002), TER (Snover et al., 2006), and Meteor[12] (Denkowski and Lavie, 2011) algorithms.

With the first evaluation experiment we translated 16 aligned English-German labels with different translation models (Section 4.2). Furthermore, we translated the bilingual German GAAP to see which translation model performs best regarding the 2794 financial labels that are stored in this ontology (Section 4.3). We also compared the perplexity between several language models and the vocabulary stored in the UK GAAP ontology (Section 4.4). Finally we applied the best translation model to the monolingual ontology and undertook a manual, cross-lingual evaluation with six annotators (Section 4.5).

4.1 Translation System: Moses Toolkit

For generating the translations from English into German, we used the statistical translation toolkit Moses (Koehn et al., 2007). Furthermore, we aimed to improve 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 the GIZA++ toolkit (Och and Ney, 2003), where the 5-gram language

[12] Meteor configuration: exact, stem, paraphrase

model was built by SRILM with Kneser-Ney smoothing (Stolcke, 2002).

4.2 Translating aligned UK – German GAAP labels

The UK GAAP is a monolingual ontology which holds 142 financial labels. With the help of the German equivalent, i.e. German GAAP, we aligned 16 German labels with the English ones, stored in the UK GAAP. This allowed us to do a small automatic evaluation, regardless of the low number of labels to be translated.

		Scoring Metric				
Source	# correct	BLEU-2	BLEU-4	NIST	TER	Meteor
JRC-Acquis	3	0.2629	0.2747	1.8112	0.6969	0.1579
ECB	3	0.2572	0.2725	1.5282	0.7878	0.1707
Linguee+Wikipedia	5	0.3623	0.2922	2.3259	0.6363	0.4085

Table 5: Evaluation scores for aligned UK–German GAAP translations

Despite the small amount of translations Table 5 shows the Linguee + Wikipedia resource produces the best BLEU score.

#	Source Label	Linguee+Wikipedia Model	Reference Translation
1	Fixed assets	Anlagevermögen	Anlagevermögen
2	Tangible fixed assets	Sachanlagen	Sachanlagen
3	Other tangible fixed assets	sonstige Sachanlagen	sonstige Sachanlagen
4	Equity	Eigenkapital	Eigenkapital
5	Income statement	Gewinn- und Verlustrechnung	Gewinn- und Verlustrechnung
6	Intangible fixed assets	immaterielle Vermögenswerte	Immaterielle Vermögensgegenstände
7	Other intangible fixed	sonstige immaterielle Vermögenswerte	sonstige immaterielle Vermögensgegenstände
8	Social security cost	Sozialbeiträge	soziale Abgaben
9	Other provisions	die sonstigen Rückstellungen	sonstige Rückstellungen
10	Other operating income	die sonstigen betrieblichen Erträge	sonstige betriebliche Erträge
11	Wages and salaries	die Löhne und Gehälter	Löhne und Gehälter
12	Current assets	kurzfristige Vermögenswerte	Umlaufvermögen
13	Work in progress	angefangene Arbeiten	unfertige Erzeugnisse
14	Work in progress	angefangene Arbeiten	unfertige Leistungen
15	Extraordinary income	das außerordentliche Ergebnis	außerordentliche Erträge
16	Equity and Liabilities	Eigenkapital und Zur	Bilanzsumme, Summe Passiva

Table 6: Results of financial translations generated by Linguee+Wikipedia translation model

Table 6 shows the translations of the 16 financial labels which were aligned between the UK and the German GAAP. The first part of the table, examples 1 to 5, represents the correct translations, which match exactly with the reference provided by the xEBR Working Group.

The next block represents translations which do not match completely with the reference translations. Examples 6 and 7 illustrate the problem of translating the label *fixed assets*[13] that can be translated into near synonyms *Vermögenswerte* or *Vermögensgegenstände*. Example 8 shows where the translation model generated a compound, but the reference translation consists of two separate tokens. If we de-compound the translation *Sozialbeträge* into *soziale Beträge*, we get a synonym to

[13] *Fixed assets* and *Other fixed assets*

the reference translation. Examples 9 to 11 represent translations with over-specification, since the ontology labels do not require the German article[14] at the beginning of a label.

The last part of the table illustrates incorrect translations. Examples 12 to 14 are translated into idiomatic expressions, whereby example 15 shows a wrong lexical choice. The word *Income* was translated into*Ergebnis*, whereas it should have been translated into *Erträge*. In example 16 a part of the source label, i.e. *Liabilities* is missed in the target translation.

4.3 Translating the German GAAP with different models

Since we built a financial parallel resource (see Section 3.4 and 3.5) and generated a translation model based on this financial vocabulary, we tested how well the model performs on a similar ontology. Therefore we translated the aforementioned German GAAP ontology, which holds 2,794 labels[15].

		Scoring Metric				
Source	# correct	BLEU-2	BLEU-4	NIST	TER	Meteor
JRC-Acquis	47	0.2276	0.1122	2.7022	0.9337	0.1761
ECB	24	0.1715	0.0596	2.1921	0.9834	0.1321
Linguee+Wikipedia	79	0.3397	0.2292	3.9383	0.8291	0.2917

Table 7: Evaluation scores for German GAAP term translations

Table 7 illustrates the automatic metrics used to evaluate the translation of the German GAAP, where the best BLEU results are generated by the Linguee+Wikipedia translation model. We can deduce from this experiment that even though JRC-Acquis has a larger number of tokens than the Linguee+Wikipedia corpus, it does not generate better translations of financial labels. The ECB corpus also does not generate better translations, although it is considered a domain-specific corpus.

4.4 Perplexity of different language models

The automatic evaluation with the small amount of translation and their references cannot demonstrate the quality of the translation model with regard to the whole UK GAAP ontology. Therefore we compared the perplexity[16] of different language models and the vocabulary of the UK GAAP ontology. Since a better language model should assign a higher probability to its test set, we tested which generated language model gives the highest probability on the UK GAAP vocabulary.

The perplexity (1) is a reformulation of cross-entropy (2).

$$PP = 2^{H(p_{LM})} \tag{1}$$

$$H(p_{LM}) = -\frac{1}{n}\sum_{i=1}^{n} log p_{LM}(w_i|, \ldots, w_{i-1}) \tag{2}$$

Table 8 illustrates that the ECB language model generates the worst perplexity on the UK GAAP vocabulary. On the other hand, the best probability is calculated by the Linguee+Wikipedia language

[14] German articles: die, der, das

[15] For comparison, the monolingual UK GAAP holds only 142 financial labels

[16] The perplexity was calculated with the SRILM ngram tool

model, which is not a surprise, since the resource is generated from the same vocabulary. Besides that, the best perplexity is generated by the German GAAP language model, which indicates that the vocabulary is most similar to the UK GAAP in comparison to other languages models.

	logprob	Perplexity
JRC-Acquis LM	-1,656.39	243.625
ECB LM	-1,871.33	497.098
German GAAP LM	-1,528.92	159.608
Linguee + Wikipedia LM	-1,277.15	69.226

Table 8: Perplexity of the language models

4.5 Manual Evaluation of Translation Quality - UK GAAP

We have undertaken a manual evaluation campaign to assess the translation quality of our terminology translation system, which was performed with the Appraise Toolkit.(Federmann, 2012)

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 Linguee+Wikipedia approach, and c) present inter-annotator agreement scores that allow us to judge the reliability of the human rankings.

4.5.1 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.

4.5.2 Annotation Data

We set up an evaluation task containing 142 term translations and the corresponding source term. The set was then given to a total of six human annotators who classified the observed translation output according to the classification scheme described above. The human annotators were lay users without in-depth knowledge of the terms' domain.

In total, we collected 852 classification items from six annotators. Table 9 shows the results from the manual evaluation for term translations into German. We report the distribution of classes per evaluation task which are displayed in *best-to-worst* order.

	Classes		
System	A	C	N
Linguee+Wikipedia Model	59.15%	29.34%	11.50%

Table 9: Results from the manual evaluation for German

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

- S (Bennett et al., 1954);
- π (Scott, 1955);
- κ (Fleiss, 1971);
- α (Krippendorff, 2004).

Table 10 presents the aforementioned metrics' scores for German term translations.

	Agreement Metric			
System	S	π	κ	α
Linguee+Wikipedia Model	0.467	0.355	0.357	0.355

Table 10: Annotator agreement scores for German

Overall, we achieve an average κ score of 0.357, which can be interpreted as *fair agreement* following (Landis and Koch, 1977). Given the observed inter-annotator agreement, we expect the reported ranking results to be meaningful. The inclusion of domain experts into the manual evaluation campaign will be an interesting extension of the work presented.

4.6 Manual error analysis of UK GAAP

In addition to the manual evaluation we performed with six annotators on the UK GAAP ontology monolingual (Section 4.5), we also performed a closer analysis of each label.

In the first step, we extracted 36 labels from the manual evaluation campaign, where all evaluators annotated the translation as "Acceptable". Examples 1 to 7 (Table 11) depict a small set of the acceptable translations.

#	Source label	Target label
1	Equity	Eigenkapital
2	Stocks	Wertpapiere
3	Key Balance Sheet Figures	Bilanzkennzahlen
4	Revaluation Reserve	Neubewertungsrücklage
5	Interest And Similar Charges	Zinsen und ähnliche Aufwendungen
6	Debenture Loans After More Than One Year	Schuldscheindarlehen nach mehr als einem Jahr
7	Profit Or Loss On Ordinary Activities Before Taxes	Gewinn oder Verlust aus der gewöhnlichen Geschäftstätigkeit vor Steuern
8	Net Operating Income	Ergebnis aus der
9	Equity And Liabilities	Und Passiva
10	Profit Loss For The Period	Ergebnis der

Table 11: Translations which all annotators considered as "Acceptable" (1-7) and "None of both" (8-10)

We also extracted financial labels where all evaluators annotated the translations of the labels as "None of both", which indicates a low quality of the translations. These labels are shown in the last part of Table 11, examples 8 to 10. The reason for the low quality of the translations is that the target label omits part of the source label. In example 8 we miss the translation for the segment *Net operating*, in 9 *Equity* is not translated and in example 10 *Loss for the period* is missing.

4.7 Interpretation of the evaluation time and the quality of translations

In addition to the evaluation of the quality of financial label translation, we also measured the evaluation time regarding different criteria, i.e. regarding the length of the label, the quality of the translation and the evaluation time for all labels.

Evaluation time regarding the length of the source labels

Figure 2 illustrates the evaluation time regarding the length of a source label. We learned that, on average, the evaluation time increased with the length of the source label, e.g. the evaluators spent more than 9 seconds to evaluate unigram label.[17] On the other hand, it took more than 26 seconds to evaluate the longest financial label.[18]

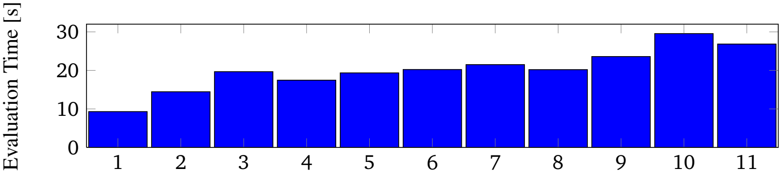

Figure 2: Evaluation time per length of the source labels

Evaluation time with respect to the quality of the translations

The evaluation task asked the evaluators to evaluate the translation quality based on three classes, "Acceptable", "Can easily be fixed" and "None of both" (cf. Section 4.5). To get a more fine-grained classification with a broader span of data, we gave each label a numeric value regarding the translation quality set by the six evaluators, e.g. the financial label *Charges* and its translation *Kosten* was annotated by all evaluators as "Acceptable"; analogously, the financial label *Financial Charges* and its translation *finanziellen Belastungen* was annotated by four evaluators with "Acceptable", whereas two evaluators annotated it as "Can easily be fixed". Since we know how each evaluator annotated a translation, we interpret the three evaluation classes into a numerical value evaluation score, i.e. if an translation was annotated with "Acceptable" we add the value 3 to the evaluation score, if it was annotated with "Can easily be fixed" we add 2, and if it was annotated with "None of both", we do not add any value to the score. With this reformulation, the financial label *Charges-Kosten* gets an evaluation score of 18,[19] and the *Financial Charges-finanziellen Belastungen* gets an evaluation score of 16.[20] With this additional classification we get a broader variety with 18 different quality classes, compared to the three classes set by the evaluators.

Figure 3 depicts the evaluation time regarding the translation quality of the financial labels. For

[17] *Assets, Reserves, Equity, Stocks ...*

[18] *Taxes Remuneration And Social Security Payable After More Than One Year*

[19] 3+3+3+3+3+3 = 18

[20] 3+3+3+3+2+2 = 16

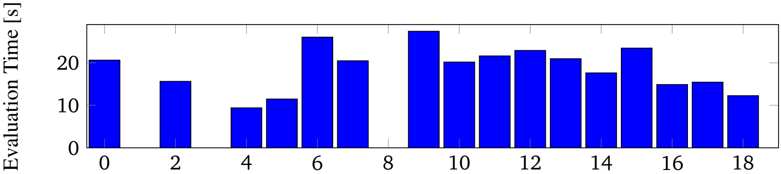

Figure 3: Evaluation time per quality of the translation

labels, which have an evaluation score of zero[21] the evaluation time is more than 20 seconds. The evaluation time decreases for labels with an evaluation score between two and five, but starts to increase when the evaluation score is equal six or more. For labels that have an evaluation score between six and thirteen, the evaluation time is higher than for labels with a lower or higher score. At the end the evaluation time decreases again. We can deduce from this experiment that it is easier to evaluate good and weak translations, but on the other hand it is harder to evaluate translations that do not belong to these two evaluation classes.

Evaluation time for the financial 142 labels

Figure 4 shows the evaluation time for all 142 labels stored in the UK GAAP ontology. We can see that the longest evaluation time to evaluate one term was more than 62 seconds, namely for the label *Operating Bach Ratios*. On the other hand, the fastest time to evaluate a label was less than 3 second for the label *Staff Costs* which was translated into *Personalkosten*.

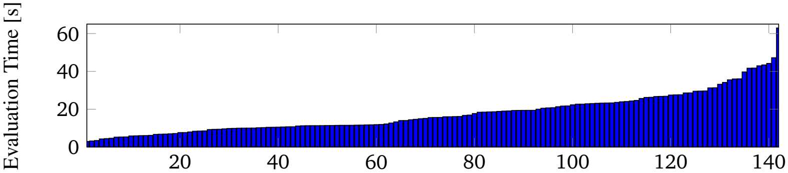

Figure 4: Evaluation time for the financial 142 labels

#	Source label	Time [s]
1	Staff Costs	2.946
2	Capital	3.177
3	Extraordinary Charges	3.421
⋮	⋮	⋮
140	Deferred Charges And Accrued Income	44.213
141	Depreciation On Intangible And Tangible Fixed Assets	47.211
142	Operating Bach Ratios	62.943

Table 12: Financial labels with the fastest (above) and the slowest (below) evaluation time

Table 12 shows the five fastest and slowest evaluation for the financial labels.

[21] *Net Operating Income*, *Equity And Liabilities*, *Profit Loss For The Period*

Conclusion and Future Work

We presented our work on the translation of a monolingual financial ontology. We performed smaller sub-experiments to determine the most appropriate translation model to translate financial labels in isolation. Hence we evaluated the translations on a small subset of aligned labels between different financial ontologies. Furthermore, we evaluated different translation models on a comparable ontology from the financial domain and compared the perplexity of the ontology to be translated with different resources. All these sub-experiments proved that the approach of building new, specific resources showed a large impact on the translation quality. Therefore, generating specialised resources for different specific domains will be the focus of our future work. On the one hand, building appropriate translation models is important, but our experiment also highlighted the importance of additional non-parallel resources, like Wikipedia, Wiktionary,[22] and DBpedia.[23] In addition to extracting Wikipedia articles with their multilingual equivalents, Wikipedia holds much more information in the articles themselves. Therefore, exploiting these non-parallel resources, as shown by (Fišer et al., 2011), would clearly help to improve the performance of the translation system. Future work needs to include the Wikipedia redirect system, which would allow a better understanding of the synonymy and spelling variations of specific terms.

In addition to exploiting new resources for statistical machine translation, the manual evaluation for monolingual resources needs to become the focus of our future work. The manual evaluation campaign was time consuming, but provided a closer look into the translation errors. It indicates that the evaluation classes for manual evaluation have to be reformulated into more fine-grained decisions. We learned that we may distinguish between translations with "one grammatical error" or "several grammatical errors". It might also be interesting to classify the types of grammatical error, e.g. number, gender or case, e.g. *Betriebsstoffen* vs. *Betriebsstoffe*. During the evaluation we also observed over-specification, where the translation into German *Die Forderungen ...*,[24] does not require the German article *die* at the beginning. Specifically to the German language we further observed some compound errors, e.g. *Ergebnis Verlust* should be merged into a compound expression. Another major issue were errors of omissions, where we miss some information from the source side, e.g. the translation *Und Passiva* omits the source part *Equity*. Further to the linguistic error classification, the type of the translation mismatch might be interesting to investigate, i.e. cultural, linguistic or domain-specific. Also it is important to know if a translation is too broad or too narrow. Especially for GAAP national differences are important as financial concepts largely depend on the legal system of the country.

In summary, the work presented in this paper outlines an initial approach to domain-specific ontology translation. It provides an indication that external resources are useful for overcoming the sparsity of data, as well as a wealth of challenges to fuel future work on this task.

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 Noëmi Aepli, Tobias Wunner and Thierry Declerck for their help with the manual evaluation. We are grateful to the anonymous reviewers for their valuable feedback.

[22] `en.wiktionary.org/wiki/Wiktionary:Main_Page`

[23] `dbpedia.org/About`

[24] generated from *Trade Debtors*

References

Aggarwal, N., Wunner, T., Arcan, M., Buitelaar, P., and O'Riain, S. (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)*.

Bennett, E. M., Alpert, R., and Goldstein, A. C. (1954). Communications Through Limited-response Questioning. *Public Opinion Quarterly*, 18(3):303–308.

Declerck, T., Krieger, H.-U., Thomas, S. M., Buitelaar, P., O'Riain, S., Wunner, T., Maguet, G., McCrae, J., Spohr, D., and Montiel-Ponsoda, E. (2010). Ontology-based multilingual access to financial reports for sharing business knowledge across europe. In *Internal Financial Control Assessment Applying Multilingual Ontology Framework*.

Denkowski, M. and Lavie, A. (2011). Meteor 1.3: Automatic Metric for Reliable Optimization and Evaluation of Machine Translation Systems. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 85–91, Edinburgh, Scotland. Association for Computational Linguistics.

Doddington, G. (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.

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

Federmann, C. (2012). Appraise: An open-source toolkit for manual evaluation of machine translation output. *The Prague Bulletin of Mathematical Linguistics*, 98:25–35.

Fišer, D., Vintar, v., Ljubešić, N., and Pollak, S. (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.

Fleiss, J. (1971). Measuring Nominal Scale Agreement among Many Raters. *Psychological Bulletin*, 76(5):378–382.

Kerremans, K. (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.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., and Herbst, E. (2007). Moses: Open Source Toolkit for Statistical Machine Translation. In *Proceedings of the 45th Annual Meeting of the ACL*, ACL '07, pages 177–180.

Krippendorff, K. (2004). Reliability in Content Analysis. Some Common Misconceptions and Recommendations. *Human Communication Research*, 30(3):411–433.

Landis, J. and Koch, G. (1977). Measurement of Observer Agreement for Categorical Data. *Biometrics*, 33(1):159–174.

Müller, C. and Gurevych, I. (2008). Using wikipedia and wiktionary in domain-specific information retrieval. In *Working Notes for the CLEF 2008 Workshop*.

Och, F. J. and Ney, H. (2003). A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). BLEU: A Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, ACL '02, pages 311–318, Stroudsburg, PA, USA. Association for Computational Linguistics.

Scott, W. A. (1955). Reliability of Content Analysis: The Case of Nominal Scale Coding. *The Public Opinion Quarterly*, 19(3):321–325.

Snover, M., Dorr, B., Schwartz, R., Micciulla, L., and Makhoul, J. (2006). A study of translation edit rate with targeted human annotation. In *Proceedings of Association for Machine Translation in the Americas*, pages 223–231.

Steinberger, R., Pouliquen, B., Widiger, A., Ignat, C., Erjavec, T., Tufis, D., and Varga, D. (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).*

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

Tiedemann, J. (2009). News from OPUS - A collection of multilingual parallel corpora with tools and interfaces. In Nicolov, N., Bontcheva, K., Angelova, G., and Mitkov, R., editors, *Recent Advances in Natural Language Processing*, volume V, pages 237–248. John Benjamins, Amsterdam/Philadelphia, Borovets, Bulgaria.

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

Weller, M., Gojun, A., Heid, U., Daille, B., and Harastani, R. (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.

Wu, H., Wang, H., and Zong, C. (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.

Zesch, T., Müller, C., and Gurevych, I. (2008). Extracting lexical semantic knowledge from wikipedia and wiktionary. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08).*