Translating the FINREP taxonomy using a domain-specific corpus

Mihael Arcan[1] **Susan Marie Thomas**[2] **Derek de Brandt**[3] **Paul Buitelaar**[1]

[1] Unit for Natural Language Processing, DERI, NUI Galway, Ireland
{mihael.arcan , paul.buitelaar}@deri.org
[2] SAP Research CEC Karlsruhe, Germany
susan.marie.thomas@sap.com
[3]XBRL Europe, Brussels, Belgium
derek.debrandt@xbrl-eu.org

Abstract

Our research investigates the use of statistical machine translation (SMT) to translate the labels of concepts in an XBRL taxonomy. Often taxonomy concepts are given labels in only one language. To enable knowledge access across languages, such monolingual taxonomies need to be translated into other languages. The primary challenge in label translation is the highly domain-specific vocabulary. To meet this challenge we adopted an approach based on the creation of domain-specific resources. Application of this approach to the translation of the FINREP taxonomy, translating from English to German, showed that it significantly outperforms SMT trained on general resources.

1 Introduction

XBRL[1] is an XML-based format developed for the electronic exchange and automated processing of financial information. Schemas and attached XML linkbases for reports to be exchanged are called taxonomies in XBRL. There has been much work on converting XBRL taxonomies to ontologies (Bao et al., 2010), so that the problem of taxonomy translation can be thought of as being the same as the problem of ontology translation. The challenge of translating ontologies is interesting because, on the one hand, ontologies represent an important linguistic resource for many applications, e.g. Ontology-based Information Extraction. With multilingual information these applications can be elevated to a cross-lingual level, i.e., in Cross-Lingual Ontology-based Information Extraction (CLOBIE) Systems. On the other hand, translating ontologies with an SMT approach is challenging due to the domain-specific vocabulary to be translated, the lack of contextual information, and the scarcity of multilingual ontologies for building translation models.

In this paper we present results from an SMT system to aid the translation of labels from one European language into others. Our experiments and evaluations were performed on labels from the FINREP[2] (FINancial REPorting) taxonomy, which represents the format to be used by European financial institutions to report their financial performance to their respective national supervisors. This reporting effort is coordinated by the European Banking Authority[3] (EBA), and has very important goals, namely, to maintain the stability of the financial system, and to protect investors and depositors. Although supervision of financial institutions is mandatory in all European member states, the XBRL taxonomy labels for FINREP are English only. To make the targeted transparency of financial information possible, and to be able to leverage labels for CLOBIE, these English labels have to be translated into the other European languages; see also Declerck et al. (2010). The challenge here lies in translating domain-specific labels, e.g. *Equity instruments*, *Interest cost* or *Tangible assets*.

The remainder of the paper is organised as follows: In Section 2 we give an overview of the related work. In Section 3 we describe the FINREP

[1]XBRL - eXtensible Business Reporting Language

[2]http://eba.europa.eu/Supervisory-Reporting/FINER.aspx
[3]http://eba.europa.eu/Aboutus.aspx

Sima'an, K., Forcada, M.L., Grasmick, D., Depraetere, H., Way, A. (eds.)
Proceedings of the XIV Machine Translation Summit (Nice, September 2–6, 2013), p. 199–206.

taxonomy and the parallel resources, which were used for label translation. In Section 4 we discuss the results of exploiting the different resources. We conclude with a summary and give an outlook on future work in Section 5.

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 their variants. 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 task of translating compound terms, 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. They evaluate their models, text semantic relatedness (for Wikipedia) and word semantic relatedness (for Wiktionary), by comparing their performance to a statistical model as implemented by 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 Experimental Data

We are investigating the problem of translating a domain-specific vocabulary, therefore our experiment started with an analysis of the financial labels stored in the FINREP taxonomy (Section 3.1).

In Sections 3.2 and 3.3 we describe existing parallel corpora, which were used to train a translation and language model. For our current research we used the JRC-Acquis, Europarl and the European Central Bank (ECB) corpora.

Finally, in Section 3.4 we describe the procedure to obtain a domain-specific corpus from Linguee and Wikipedia/DBpedia. The results of the translations produced by an SMT trained on these domain-specific resources were compared to SMT results from a system trained on more general resources.

Although previous research showed that a translation model built by using a general parallel corpus cannot be used for domain-specific vocabulary translation (Wu et al., 2008), we decided to train a baseline translation model on this existing corpora to illustrate any improvements gained by modelling a new domain-specific corpus for the financial domain.

3.1 The Financial taxonomy - FINREP

Under EU law financial institutions such as banks, credit institutions and investment firms must submit periodic reports to national supervisory bodies. The content of these reports is guided by the European Banking Authority[4] (EBA) by means of two complementary reporting frameworks: financial reporting (FINREP) and COREP[5] (COmmon solvency ratio REPorting) common reporting. These frameworks have been articulated into XBRL taxonomies to enable electronic submission and automated processing of their contents.

For our experiment we used the latest draft version of the FINREP taxonomy,[6] of which the tables

[4] www.eba.europa.eu/

[5] http://eba.europa.eu/Supervisory-Reporting/COREP.aspx

[6] http://www.eurofiling.info/

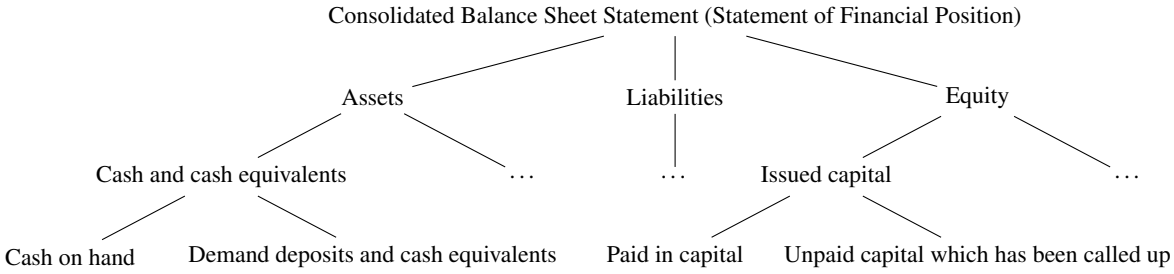

Figure 1: The financial label *Demand deposits and cash equivalents* and its ancestors in the financial taxonomy FINREP

Length	Count	Examples
30	1	*Financial assets pledged as collateral, financial assets pledged as collateral for which the transferre has the right to sell or repledge in the absence of default by the reporting institution*
	...	
2	110	*Repurchase agreements*, *Guarantees given*, *Equity instruments* ...
1	36	*Provisions*, *Securities*, *Assets* ...

Table 1: Examples of the longest and shortest financial labels in FINREP taxonomy

hold 569 monolingual labels in English.

FINREP labels are mostly noun phrases, many of which are quite long as can be seen in Figure 1. The longer labels are the product of nominalizing and condensing descriptions of the meaning of the corresponding reporting concept. Each reporting concept has, in addition to its labels, a unique cluster of XBRL identifiers, which are used to tag instances of the concept, e.g., particular monetary values. By means of the tag clusters computers can process reports automatically, whereas humans need labels to understand the report contents. Therefore it is important to translate these financial labels with exact meaning preservation.

The length of the financial labels varies, e.g. the longest financial label considered for translation has a length of 30 tokens, while others may consist of 1 or 2 (Table 1).

From the distributional aspect, the taxonomy consists of 36 unigrams, which represents 6% of all labels in the taxonomy. Further the taxonomy holds 110 bigrams (20%), 74 3-grams (13%) etc. This shows another property of the labels which are rather short. Unigrams, bigrams, ... 5-grams together represent more than 50% of the vocabulary in the taxonomy.

finrepTaxonomy/finrep-beta-20121210_rend.zip

3.2 JRC-Acquis and Europarl

The parallel corpus JRC-Acquis[7] (version 3.0) is a collection of legislative texts of the European Union written between 1950 and now and is available in more than twenty official European languages (Steinberger et al., 2006). The English-German parallel corpus consists of 1.2 million aligned sentences, and 32 million English and 30 million German tokens.

A similar corpus to JRC-Acquis is the Europarl parallel corpus (version 7),[8] which holds proceedings of the European Parliament in 21 European languages. We used the English-German parallel corpus with around 1.9 million aligned sentences and 47 million English and 45 million German tokens (Koehn, 2005).

3.3 European Central Bank Corpus

For comparison with the large Acquis and Europarl corpora, we also did experiments using the European Central Bank Corpus[9], which contains 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 (Tiedemann, 2009).

For our research we used the English-German language pair, which consists of around 113,000 English-German sentence pairs or 2.8 million English and 2.5 million German tokens.

3.4 Domain-Specific Corpus

Besides the approach to translate specific vocabulary with large corpora, i.e. Europarl or Acquis, we modelled a new domain-specific financial corpus. Sections 3.4.2 and 3.4.1 describe the resources that

[7] http://ipsc.jrc.ec.europa.eu/index.php?id=198

[8] http://www.statmt.org/europarl/

[9] http://opus.lingfil.uu.se/ECB.php

we used to build the new domain-specific resource (Section 3.4.3).

3.4.1 Wikipedia and DBpedia

Wikipedia[10] is a multilingual, freely available encyclopaedia that was built by a collaborative effort of voluntary contributors. It stores approximately 22 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.[11]

With its 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.

To avoid the scalability problem which comes from parsing the whole Wikipedia XML dump, we used the datasets, generated by the DBpedia project (Bizer et al., 2009). The DBpedia project is a community effort to extract structured information from Wikipedia and to make this information accessible on the Web.

We used the DBpedia datasets (3.8)[12] to extract relevant Wikipedia article titles, the variants and the translations of article titles, and the categories associated with these articles. The article titles were used to build a domain-specific multilingual lexicon.

3.4.2 Linguee - Dictionary and Translation Search Engine

To create another domain-specific resource, we built a new parallel corpus based on the taxonomy vocabulary that we want to translate. For this we used Linguee,[13] a combination of a dictionary and a search engine, which indexes words and expressions from around 100 million bilingual texts. Linguee search results display example sentences that show how the expression searched for has been translated in context. In contrast to translation engines like Google Translate[14] or Bing Translator,[15] which are automatic translation engines based on statistical methods 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.

Since Linguee includes EU documents, it also contains parallel sentences from JRC-Acquis and Europarl. These sentences were excluded from our domain-specific parallel corpus.

3.4.3 Financial Corpus Generation

From the Wikipedia knowledge represented in the DBpedia datasets we built a multilingual financial lexicon. We started with the labels extracted from a similar financial taxonomy, German GAAP (GAAP - Generally Accepted Accounting Practice),[16] ontology, which has 2794 concepts, each with labels in German and English. We collected Wikipedia article titles which matched labels, or sub-parts of labels, in the German GAAP ontology. An example of collected titles is shown in Figure 2, part 1. Each collected title was also annotated with the categories associated with the title (part 2). These categories were used for word sense disambiguation (WSD) to avoid collecting out-of-domain translation pairs such as the English *Stocks* and its German translation *Stock*.[17] WSD was performed by ranking the categories by frequency and filtering the title collection down to those which were annotated with the top categories (part 3).

Wikipedia also relates titles to variant or similar titles. i.e. Wikipedia redirection function. We also gathered these, i.e., the Wikipedia title *Profit and loss statement* and *Operating statement* are redirected to the Wikipedia article *Income statement*. This information allowed us to align *Profit and loss statement* with the translation of the Wikipedia article *Income statement*. In addition, we also aligned multiple German expressions with one English Wikipedia article title, e.g. the Wikipedia article title *Financial crisis* was aligned with the German Wikipedia article titles *Finanzkrise* and *Finanzmarktkrise*.

[10] http://en.wikipedia.org/wiki/Main_Page
[11] http://en.wikipedia.org/wiki/Wikipedia:About
[12] http://wiki.dbpedia.org/Downloads38
[13] http://www.linguee.com/
[14] http://translate.google.com/
[15] http://www.bing.com/translator
[16] http://www.xbrl.de/
[17] en. "devices used as a form of physical punishment"

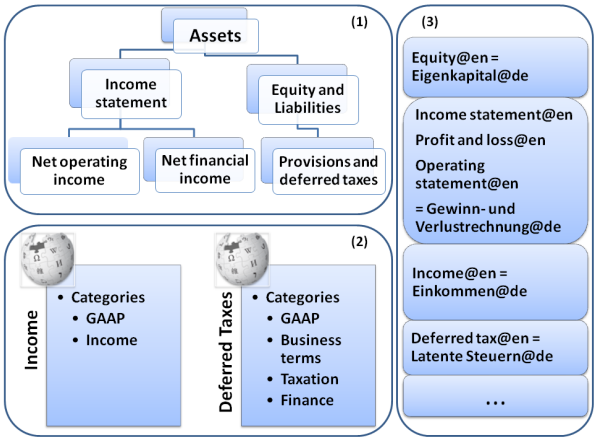

Figure 2: Steps of extraction Wikipedia titles and its translations

In summary, we created a rich multiligual lexicon from Wikipedia article titles, and used the categories associated with articles to perform WSD in order to create a domain-specific lexicon. With this approach we extracted more than 7000 aligned financial expressions from the DBpedia dataset, examples of which can be seen in part 3.

In a similar fashion, the Linguee search engine was queried for unigrams extracted from the German GAAP ontology. For each query, we extracted the parallel output and stored the source and target sentences. With this approach we built a parallel corpus with around 193,000 aligned sentences.

Finally we merged both resources together, which resulted in more than 200,000 aligned entities, i.e. financial sentences, phrases and unigram expressions. Thanks to the extensive multilingual data of Wikipedia and Linguee we were actually able to generate translation models for several languages, i.e. English-German, English-Spanish and English-French (see Figure 3).

4 Experiments and Evaluation

Since the FINREP taxonomy is monolingual a straightforward automatic evaluation is not possible. Therefore we randomly chose 100 labels, which were translated into German by an expert. For this experiment we concentrated only on translations from English to German and vice versa.

For the automatic evaluation we used the BLEU (Papineni et al., 2002), NIST (Doddington, 2002), TER (Snover et al., 2006), and Meteor[18] (Denkowski and Lavie, 2011) algorithms.

[18]METEOR configuration: exact, stem, paraphrase

4.1 Moses Toolkit and Graphical User Interface

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 alignments were built with the GIZA++ toolkit (Och and Ney, 2003), where the 5-gram language model was built by SRILM with Kneser-Ney smoothing (Stolcke, 2002).

In combination with the Moses Toolkit we built a freely accessible graphical user interface (GUI), which uses the domain-specific translation models described.[19]

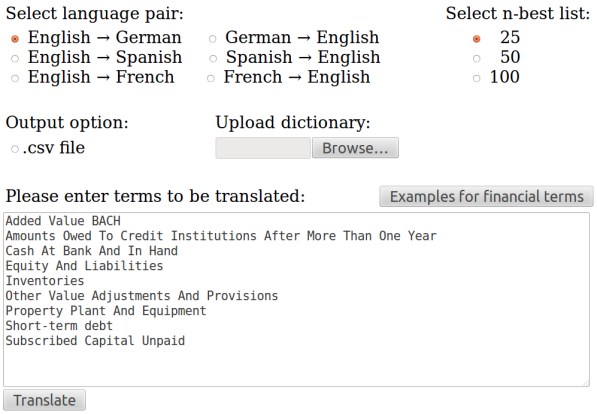

Figure 3: Translation GUI for the financial domain

Figure 3 illustrates the options of the GUI. The interface allows different language pairs and different size n-best lists when the translations are generated. Further an output option is available, which generates a downloadable .csv file. The "Upload dictionary" option allows the user to upload a dictionary. The translation pairs in the uploaded dictionary are used to guide the SMT system to choose the dictionary entries over other options. This is achieved by annotating the decoder input using the Moses XML input markup scheme.

In addition to the GUI, a RESTfull service was implemented, which returns a .json format, and also provides translation probabilities for further programmatic processing.

4.2 Automatic Evaluation

For the automatic evaluation we randomly selected 100 FINREP labels, which were translated manu-

[19]http://monnet01.sindice.net/monnet-translation/

ally by an expert. The longest label in the evaluation set is *Impairment or (-) reversal of impairment on financial assets not measured at fair value through profit or loss*. On the other hand, the test set contained 17 unigram labels, i.e., *Subsidiaries*, *Deposits*, *Restructuring* ...

Table 2 illustrates the automatic metrics used for evaluating the translations of the 100 FINREP labels. The best BLEU result (0.3154) was produced by the financial translation model. We can deduce from this experiment that even though JRC-Acquis has a larger number of tokens than the financial parallel corpus, it does not generate better translations of financial labels (0.1230 BLEU). The ECB corpus also does not generate better translations (0.1186) than the financial model, although it is considered to be a domain-specific corpus. Comparing the ECB model with the JRC-Acquis model, the ECB model performs even worse that the JRC-Acquis. The worst scores were generated by the Europarl translation model, i.e. 0.0494 BLEU points.

Besides the comparison with SMT trained parallel corpora, we also compared our approach with translations generated by Google Translate and Bing Translator.[20] Regarding the BLEU evaluation metric, the financial corpus generates better translations of financial labels than Google Translate or Bing Translator.

4.3 Manual error analysis of FINREP labels

In addition to the automatic evaluation of the 100 FINREP labels, which assessed the different resources (Europarl, JRC-Acquis, ECB) and services (Google Translate and Bing Translator), we also performed a deeper manual evaluation of the 100 labels.

Analysing the translation manually we recognised several error classes. The first issue concerns ambiguous labels which are translated into the general domain. This phenomenon is caused by the general translation model, i.e. Europarl and JRC-Acquis. If we use the translation models from Europarl, the segment *Table* is translated into *Tisch*, which names a piece of furniture, but not a diagram with rows and columns. Furthermore, the segment *equity* was translated by the Europarl model into *Gerechtigkeit*.[21] The financial model generated the reference translation *Eigenkapital*. Similarly the ambiguous segment *balances* was translated into *Gleichgewichte*.[22] Building the financial model with domain-specific data provided the correct financial expression *Guthaben*.

Although Google Translate and Bing Translator often translate the financial labels correctly, they make mistakes if the label consists of only one token. Without any contextual information they translate a label into the most probable meaning, which is usually the general one. Therefore they translate the label *Capital* into *Hauptstadt*.[23]

Another group of mistranslations were caused by the fact that the general model often translates specific vocabulary separately, token by token, and not as a phrase. In this case the label *Other demand deposits* was translated by JRC-Acquis and Europarl models as *andere fordern Einlagen* and *Sonstige Nachfrage Einlagen*, where each token on the source side was translated in isolation from the other tokens. On the other hand, the translation provided by the domain-specific model matches the reference translation, i.e. *sonstige Sichteinlagen*. The same pattern is seen when translating the label *Foreign currency translation*, which is mistranslated by the Europarl model as *ausländische Währung übersetzung*. When using the JRC-Acquis model we get an acceptable translation *Umrechnung von Fremdwährungen*, whereas the domain-specific model provides the preferred German compound, *Währungsumrechnung*, which matches the reference translation. Similarly, the segment *parent* in the label segment *Profit or loss attributable to owners of the parent* is translated wrongly as *der Eltern*, although it should be, as shown by the domain-specific corpus, translated as i.e. des Mutterunternehmens.

A final observation concerns the translation of non-alphabetic characters, e.g., not only the domain-specific model, but also Google Translate, translates the segment *Table 10.0.* as *Tabelle 10,0*, where the period is wrongly changed into a comma. In this case some re-normalisation has to be done.

5 Conclusion and Future Work

We presented our work on the translation of a monolingual FINREP financial taxonomy. Our ex-

[20] Both queries were made on April 14th, 2013

[21] en, justice, righteousness, fairness

[22] en. counterbalance, equilibrium, equipoise

[23] en. capital of a country

		Scoring Metric				
Source	# correct	BLEU-2	BLEU-4	NIST	TER	METEOR
Europarl	7	0.1239	0.0494	1.6021	1.0746	0.0851
JRC-Acquis	16	0.2619	0.1272	2.8062	0.8567	0.2112
ECB	7	0.2283	0.1186	2.4208	0.9791	0.1671
domain-specific model	17	0.4119	0.3154	3.7615	0.8328	0.3171
Google Translate	15	0.3586	0.2225	3.5733	0.7641	0.2761
Bing Translator	15	0.3566	0.2337	3.6062	0.7462	0.2831

Table 2: Evaluation scores for the FINREP label translations, English to German (# correct = exact translations / perfect matches)

periment proved that the approach of building a new, domain-specific corpus showed a large impact on the translation quality.

On the one hand, building appropriate translation models is important, but our experiment also highlighted the importance of additional non-parallel resources, like Wikipedia and DBpedia. In addition to Wikipedia article titles with their multilingual equivalents, Wikipedia holds much more information in the articles themselves. Therefore, exploiting these non-parallel resources in future, as shown by Fišer et al. (2011), would clearly help to improve the performance of the translation system.

Besides Wikipedia/DBpedia, which can be used for lexicon generation and WSD, the Web itself stores an enormous amount of data, which is often represented in a multilingual way. Therefore a major part of the future work needs to be focused on extraction and alignment on multilingual websites and documents (Resnik and Smith, 2003).

In addition to exploiting new resources for statistical machine translation, the manual evaluation for translated labels needs to become the focus of our future work. Although such manual evaluation is time consuming, it provides a closer look into the translation errors. Even through the small manual evaluation of 100 FINREP labels we learned that fine-grained translation error classes have to be formulated. We observed that we have to 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* (en. *factory supplies*). During the evaluation we also observed over-specification, where the translation into German *die Bilanzsumme*,[24] does not require the German definite article *die* at the beginning. Specifically, in the German language we further observed some compound errors, e.g. *Wert Anpassungen* should be merged into a compound expression *Wertberichtigungen* (en. *Value adjustments*). Another major issue was errors of omission, where we miss some information from the source side and change the original meaning, e.g. the translation of the segment *Non-current assets* omits the negative particle, i.e. *die derzeitigen Guthaben* (en. *the current assets*). Beyond 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 can be used as a variant expression *Anfangsbestand* vs. *Eröffnungsbilanz*, (en. *Opening balance*) or whether it is too broad or too narrow. Therefore discussions are ongoing to have the German labels verified by experts appointed by the German Bundesbank (en. German Federal Bank). National differences are especially important, as financial concepts heavily depend on the legal system of the country.

In summary, the work presented in this paper described and evaluated an approach to financial taxonomy translation. The evaluation indicates that external resources, such as the data stored in the web are useful for overcoming the sparsity of domain-specific training data for SMT.

Acknowledgments

This publication has emanated from research supported in part by a research grant from Science Foundation Ireland (SFI) under Grant Number

[24] generated from *Total assets*

SFI/12/RC/2289 and by the European Union under Grant No. 248458 for the Monnet project.

References

Bao, J., Rong, G., Li, X., and Ding, L. 2010. Representing financial reports on the semantic web: - a faithful translation from xbrl to owl. In *RuleML*, pages 144–152.

Bizer, C., Lehmann, J., Kobilarov, G., Auer, S., Becker, C., Cyganiak, R., and Hellmann, S. 2009. Dbpedia - a crystallization point for the web of data. volume 7, pages 154–165.

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*, pages 67–76. Internal Financial Control Assessment Applying Multilingual Ontology Framework. HVG Press Kft.

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.

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.

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

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.

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.

Resnik, P. and Smith, N. A. 2003. The web as a parallel corpus. volume 29, pages 349–380.

Snover, M., Dorr, B., Schwartz, R., Micciulla, L., and Makhoul, J. 2006. A study of translation edit rate with targeted human annotation. In *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.