Published in final edited form as:

Nat Lang Process Inf Syst. 2014 June ; 8455: 33–38. doi:10.1007/978-3-319-07983-7_4.

Ontology translation: A case study on translating the Gene Ontology from English to German

Negacy D. Hailu[✉], K. Bretonnel Cohen[✉], and Lawrence E. Hunter

Computational Bioscience Program, University of Colorado School of Medicine, USA

Abstract

For many researchers, the purpose of ontologies is sharing data. This sharing is facilitated when ontologies are available in multiple languages, but inhibited when an ontology is only available in a single language. Ontologies should be accessible to people in multiple languages, since multilingualism is inevitable in any scientific work. Due to resource scarcity, most ontologies of the biomedical domain are available only in English at present. We present techniques to translate Gene Ontology terms from English to German using DBPedia, the Google Translate API for isolated terms, and the Google Translate API for terms in sentential context. Average fluency scores for the three methods were 4.0, 4.4, and 4.5, respectively. Average adequacy scores were 4.0, 4.9, and 4.9.

1 Introduction

For many researchers, the purpose of ontologies is sharing data. This sharing is facilitated when ontologies are available in multiple languages, but inhibited when an ontology is only available in a single language. There are public health consequences for the lack of availability of ontologies in multiple languages, as well—Rebholz-Schuhmann has posited that multilinguality will be the key to transferring the results of genomic research to actual patient care (personal communication). There would thus be considerable significance to the development of the ability to translate biomedical ontologies from one language to another. However, manual translation of biomedical ontologies is unlikely to scale, for a number of reasons. Translation of biomedical linguistic resources may require considerable domain expertise in addition to strong language skills. Biomedical ontologies change frequently, making manual updating difficult.

A number of hypotheses are evaluated in the work reported here. The most basic hypothesis is that machine translation of biomedical ontologies at an acceptable level of fluency and adequacy is possible. A further hypothesis is that the problems of ambiguity, domain specificity, and lack of context of terms that come from an ontology can be overcome by providing appropriate linguistic context. The final hypothesis is that context can be provided automatically by retrieval of sentences from a domain-specific document collection.

[✉]negacy.hailu@ucdenver.edu
[✉]kevin.cohen@gmail.com
larry.hunter@ucdenver.edu

German was chosen as the target language for the machine translation effort. There were two motivations for this. The first is that German is the second-most represented language in PubMed/MEDLINE, after English [14]. Thus, there are many opportunities for mapping English-language scientific publications to German-language publications if ontology terms can be translated. The second is that Germany is a Tier 1 market for electronic medical records, with a projected hospital-based electronic medical record adoption rate of 35% in 2013 [8]. Thus, there is a large opportunity for penetrance of basic research into the German health care system if the ontology translation problem can be solved.

As input for the machine translation task, the Gene Ontology [9] was selected. The Gene Ontology is the flagship ontology in the biomedical domain. It contains about 38,000 concepts describing the broad categories of molecular function, biological process, and cellular component. It is essentially the lingua franca for comparing gene functions across and within organisms. It is used by many large biological database efforts for the annotation and classification of genes in multiple organisms. Gene Ontology annotations (mappings between genes and Gene Ontology concepts) are maintained and distributed by some of the largest genomic databases in the world—millions of them are available through the Gene Ontology Annotation [9] database.

2 Research strategy

To understand the nature of the data, two foundational questions were investigated: the distribution of GO terms with respect to word length, and the coverage of GO terms in DBPedia.

Both the full Gene Ontology and the species-neutral GO slim were sampled for evaluation. GO “slims” are subsets of GO terms that are cleaner and give a broader overview of the ontology content, as opposed to the full Gene Ontology.

This was motivated by the desire to test the naive hypothesis that if a GO term is single word, then we can translate it using a dictionary—we evaluated the popular English-German online dictionaries www.dict.cc and www.leo.org. For terms with multiple words (which are not accessible in dictionaries), we hypothesized that Wikipedia and the Google Translate API can be used to do the translation.

The null hypothesis that there is no difference within the two groups of terms from the full Gene Ontology and from the species-neutral GO slim in frequency in terms of word length was tested using the non-parametric *Wilcoxon Rank Sum Test*. At a $p\text{-value} < 0.5$, we failed to reject the null hypothesis. Since the statistical test showed us there is no difference between the GO terms and GO slims with respect to word length, we focused testing and evaluation on terms sampled from the GO slims.

The naive hypothesis would be that single-word terms could be translated using dictionaries, but only 13% of the GO slim terms are a single word. This leaves a large portion that cannot be translated using dictionaries. There are also other issues that remained critical while using dictionaries to translate single word terms. Besides the obvious issue of ambiguity, the main issue is *domain specificity*—for example, www.dict.cc gives us 15 translations for the word

cell—being the second translation for *cell phone*. We concluded that general online dictionaries are not suitable for domain-specific translation. In Section 3, we discuss how we use Wikipedia and Google Translate API to translate multiple-word terms.

3 Methods

Machine translation is traditionally done in one of two ways—Statistical Machine Translation (SMT), or Rule-based Machine Translation (RBMT). The strength of RBMT is that it does not require huge computational resources, and as pointed out in [1], it is good for languages with limited resources. However, RBMT is based on sometimes controversial linguistic theories, and it can be costly to get human experts with the appropriate expertise. Also, inconsistency among the experts is a big challenge. The strength of SMT is that it works well and can be done without domain or linguistic expertise. However, it requires large collections of parallel texts. Neither of these approaches can be applied directly to the problem of ontology translation, as there are currently no rule-bases for translations of the domain-specific terminology, and there is no extant parallel corpus for translation of technical noun phrases.

We used DBpedia—structured Wikipedia [2] to translate GO terms from English to German. We performed keyword-based search for each term in the DBpedia dataset. If a term has both an English and a German Wikipage, we pick the title of that term in both pages as a translation. It turns out that 25% of the terms have a wiki page in both languages.

Studies show that the Google Translate API is widely used by many users in day to day life, usually on a sentence-level basis, though its accuracy varies widely from language to language [3]. Although it is very popular for general-domain translation, Steedman has shown that it is easy to reveal systematic sources of problems with it even in the general domain [4]. Domain-specific translation using Google Translate has not, to our knowledge, been evaluated. The Google Translate API is a statistical machine translation model which is trained on a huge data set collected from different sources, but mainly from online news. It was not primarily trained on biomedical literature. We hypothesize that the Google Translate API gives better translation for terms in the context of full sentences than for terms in isolation. In order to test this hypothesis we need to prepare two datasets—GO terms in their isolated form, and within a sentence. We searched PubMed/Medline by full text match to collect sentences that have GO terms from our sample in them. Note that there is no guarantee that this solves, nor does it attempt to solve, the issue of within-domain word sense ambiguity (as opposed to general language ambiguity—see [16]). We collected more than 1000 sentences in this way. We translated both the GO terms in their isolated form and in sentences from English to German using the Google Translate API.

4 Results

Evaluation was done by six human experts in the field who are fluent in both languages. Five are native speakers of German who speak and write in English routinely in their professional lives. One is bilingual in English and German (i.e., speaks both languages as a mother tongue). Four are biomedical scientists. Two are computer scientists. The characteristics of

the random sample of terms and the translation techniques applied to them are as shown in Table 1a. The evaluators were asked to evaluate each translated term for *adequacy* and *fluency* on a Likert scale of 1 through 5, where 1 is poor and 5 is excellent. Adequacy addresses how much information is preserved between the original and translated text. Fluency addresses how good the German output is.

The results of the evaluation for the three methods are shown in Table 1b. The scores for the last two methods, i.e. Google Translate API with and without context, are similar. However, Wiki translation performed relatively low.

The goal of the third method—constructing sentences that include Gene Ontology terms and using Google Translate API to translate them—was meant to improve the adequacy of the translation. From the evaluation results in Table 1b, there appears to be little difference in the results for the two methods. Even though previous works using Statistical Machine Translation to translate ontologies [5] and [6] concluded that context plays a large role in ontology translation, our results suggest that the magnitude of the difference is small.

Terms like *bleb*—an irregular bulge in the plasma membrane of a cell—were not properly translated by any of the methods. However, it was correctly translated as *zeiose* with the help of its English synonym *zeiosis*. This suggests that synonyms could play a large role in translation.

We noted large differences in performance between DBPedia and the Google Translate API, but not between the performance of the Google Translate API with and without additional context. There is neither an improvement nor a degradation of performance when additional context is provided. From this we conclude that the extra work of providing context may not be required for translating biomedical ontologies, although further work is required to see if this result holds for other ontologies.

5 Conclusion

Ontology terms can be a single word or as long as a complex phrase. Translating the short terms—terms which have word length less than or equal to 3—is difficult, since they are ambiguous. In this work, we constructed sentences using the ontologies so that the terms would have some contextual cues as to word sense. The sentence construction was done using the terms as key words in a PubMed/MEDLINE search. Evaluation was done by human experts. The evaluators were given a Likert scale ranging from 1 through 5, where 1 is poorly translated and 5 is fluent. They were asked to assess adequacy and fluency. On average, the adequacy Likert scale scores were 4.0, 4.9, and 4.9 for DBPedia, GTA without context, and GTA with context. Average fluency scores for the three methods were 4.0, 4.4, and 4.5. Our results showed that performance for translation of short terms was not improved when the terms are in a sentential context rather than in isolation.

The Google Translate API performed well, both with and without context. Not surprisingly, the wiki-based approach did not work as well. Part of the reason is the structure of entries in DBPedia. We used keyword-based exact match query on the DBPedia dataset. This fails with acronyms, as in *ADPG pyrophosphorylase complex*, which are quite common in the

Gene Ontology, and therefore cause problems with adequacy. What is surprising is that the fluency was rated low for this approach, as well.

The work reported here describes experiments and analyses that add to our understanding of the challenges and opportunities for automatic translation of biomedical ontologies. Overall, the results are quite encouraging, with high adequacy and fluency achievable for the terms associated with most concepts. Automatic translation overcomes the issues of scale and updating in biomedical ontologies, and enables research into the fundamental issues concerned with translated ontologies, such as preservation of structure and relationships in the second language. Do the relationships between concepts in an ontology remain constant when the terms in the ontology are translated into arbitrary languages? In theory, they should, since the concepts are specified by their definitions, not by the associated terms. However, whether or not this is the case has not been evaluated. The availability of complete translated ontologies, possibly produced by methods like the ones described here, would enable investigation of this question.

Acknowledgments

The authors thank Carsten Goerg, Barbara Grimpe, Daniel Pape, Christophe Roeder, Kai Stuehler, and Daniel Waldera-Lupa for their work on the evaluation. The work reported here was supported by grants 5R01 LM009254-07, 5R01 LM008111-08, and 2T15LM009451 to Lawrence E. Hunter.

References

1. Costa-Jussà, Marta R., Farrús, Mireia, Mariño, Jo e B., Fonollosa, Jo e AR. Study and Comparison of Rule-based and Statistical Catalan-Spanish Machine Translation Systems. *Computing and Informatics*. 2012; 31:245270.
2. Lehmann, Jens, Isele, Robert, Jakob, Max, Jentzsch, Anja, Kontokostas, Dimitris, Mendes, Pablo N., Hellmann, Sebastian, Morsey, Mohamed, van Kleef, Patrick, Auer, Sören, Bizer, Christian. DBpedia A Large-scale, Multilingual Knowledge Base Extracted from Wikipedia. *Semantic Web*. 2012
3. Aiken, Milam, Balan, Shilpa. An Analysis of Google Translate Accuracy. *Translation Journal* 16.2. 2013
4. Steedman, Mark. On becoming a discipline. *Computational Linguistics*. 2008; 34(1):137–144.
5. Michael Arcan and Paul Buitelaar Ontology Label Translation. *HLT-NAACL*. 2013
6. McCrae, John. *ACL HLT 2011. The Association for Computational Linguistics*; 2013. Mauricio Espinoza Combining statistical and semantic approaches to the translation of ontologies and taxonomies *Fifth Workshop on Syntax, Semantics and Structure in Statistical Translation*, p. 40-46.
7. Brown, Peter F., Della Pietra, Vincent J., Della Pietra, Stephen A., Mercer, Robert L. *Computational Linguistics*. Jun.1993 19(2)
8. Accenture 2010. Overview of international EMR/EHR markets: Results from a survey of leading health care companies
9. Michael, Ashburner, Catherine, Ball, Judith, Blake, David, Botstein, Heather, Butler, Cherry, J Michael, Davis, Allan, et al. The Gene Ontology Consortium. Vol. 25. *Nature Genetics*; 2000. Gene ontology: tool for the unification of biology; p. 2529
10. Aggarwal N, Asooja K, Buitelaar P. DERI&UPM: Pushing corpus based relatedness to similarity: Shared task system description. *SemEval-2012*. 2012
11. Denkowski, M., Lavie, A. Proceedings of the Sixth Workshop on Statistical Machine Translation. Edinburgh, Scotland: Association for Computational Linguistics; 2011. Meteor 1.3: Automatic Metric for Reliable Optimization and Evaluation of Machine Translation Systems; p. 85-91.

12. Gabrilovich, E., Markovitch, S. Proceedings of The Twentieth International Joint Conference for Artificial Intelligence. Hyderabad, India: 2007. Computing semantic relatedness using Wikipedia-based explicit semantic analysis; p. 1606-1611.
13. Papineni, K., Roukos, S., Ward, T., Zhu, WJ. Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA, USA: Association for Computational Linguistics; 2002. BLEU: A Method for Automatic Evaluation of Machine Translation; p. 311-318.ACL 02
14. Jimeno-Yepes A, Prieur-Gaston É, Névél A. Combining MEDLINE and publisher data to create parallel corpora for the automatic translation of biomedical text. BMC Bioinformatics. 2013; 14(1):146. [PubMed: 23631733]
15. Doddington G. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. Proceedings of the second international conference on Human Language Technology Research. 2002:138–145. HLT 02.
16. Johnson, Helen L., Cohen, K Bretonnel, Baumgartner, William A., Jr, Lu, Zhiyong, Bada, Michael, Kester, Todd, Kim, Hyunmin, Hunter, Lawrence. Evaluation of lexical methods for detecting relationships between concepts from multiple ontologies. Pacific Symposium on Biocomputing. 2006; 11:28–39.

Table 1
Dataset size and score

(a) Evaluated dataset size		
Methods	Total terms	Evaluated terms
Wikipedia translation	75	75
Google Translate API without context	1000	100
Google Translate API with context	1000	100
Total terms	2075	275

(b) Average likert scale score				
Methods	Group-1		Group-2	
	Adequacy	Fluency	Adequacy	Fluency
Wiki translation	4.3	4.3	3.9	3.7
Google Translate API without context	4.9	4.6	4.8	4.2
Google Translate API with context	4.9	4.5	4.9	4.4