Corpus-based Translation of Ontologies for Improved Multilingual Semantic Annotation

Claudia Bretschneider[1,2], **Heiner Oberkampf**[1,3], **Sonja Zillner**[1], **Bernhard Bauer**[3], **Matthias Hammon**[4]
[1]Siemens AG, Corporate Technology, Munich, Germany
[2]Center for Information and Language Processing, University Munich, Germany
[3]Software Methodologies for Distributed Systems, University Augsburg, Germany
[4]Department of Radiology, University Hospital Erlangen, Germany
{claudia.bretschneider.ext,heiner.oberkampf.ext,sonja.zillner}@siemens.com,
bernhard.bauer@informatik.uni-augsburg.de, matthias.hammon@uk-erlangen.de

Abstract

Ontologies have proven to be useful to enhance NLP-based applications such as information extraction. In the biomedical domain rich ontologies are available and used for semantic annotation of texts. However, most of them have either no or only few non-English concept labels and cannot be used to annotate non-English texts. Since translations need expert review, a full translation of large ontologies is often not feasible. For semantic annotation purpose, we propose to use the corpus to be annotated to identify high occurrence terms and their translations to extend respective ontology concepts. Using our approach, the translation of a subset of ontology concepts is sufficient to significantly enhance annotation coverage. For evaluation, we automatically translated RadLex ontology concepts from English into German. We show that by translating a rather small set of concepts (in our case 433), which were identified by corpus analysis, we are able to enhance the amount of annotated words from 27.36 % to 42.65 %.

1 Introduction

Ontologies offer a powerful way to represent a shared understanding of a conceptualization of a domain (Gruber, 1993a). They define concepts and relations between them. Further linguistic information, such as labels, synonyms, abbreviations or definitions, can be attached. This is how ontologies provide a controlled vocabulary for the respective domain. In Information Extraction (IE), the controlled vocabulary of ontologies is used to recognize ontology concepts in text (also referred to as *semantic annotation*) and combine the textual information and the ontological knowledge to allow a deeper understanding of the text's semantics.

The problem, however, is that most of the available ontologies are not multilingual, i.e., they have either no or only few non-English concept labels. To make ontologies applicable for IE-based applications dealing with non-English texts, one has to translate at least some of the concept labels. Since high quality translations need expert review, a full translation of big ontologies is often not feasible. In the biomedical domain, ontologies have a long tradition and many well designed, large and semantically rich ontologies exist. At the time of writing, the BioPortal (Noy et al., 2008), an ontology repository for the biomedical domain, contains 370 ontologies, where 49 have more than 10,000 concepts. Their complete translation would be very costly.

In many application scenarios, only a subset of ontology concepts is of relevance. This is especially true for IE: If we consider, e.g., the semantic annotation of medical records in the context of a specific disease, the translation of a subset of ontology concept labels can be sufficient to increase the number of ontology concepts found. Thus, the translation of a small set of labels, which is relevant for the application scenario, is sufficient to increase the ontology's applicability for IE from non-English texts.

That is why we propose a translation approach that identifies the *most relevant concepts* for the application scenario and adds their translations to the ontology. The application scenario is represented by the *corpus*, a 'large set of domain-specific text'. In the context of IE, the main goal is to achieve a high annotation coverage, i.e., a high amount of words are semantically annotated with the correlating ontology concepts. Therefore, we define the terms with *high frequency* in the corpus as *most relevant* for translation, as the translation of high frequency terms increases the annotation coverage significantly. To demonstrate the feasibility of our approach, we use the RadLex ontology (Langlotz, 2006) and a corpus of German radiology reports of lymphoma patients.

2 Related Work

Ontology-based IE is a commonly used technique in the biomedical domain. (Meystre et al., 2008) give a detailed overview of recent research activities. However, most projects focus on English texts. The ontology translation problem was first described by (Gruber, 1993b) and further formalized by (Espinoza et al., 2009b). The subproblem we are dealing with is ontology localization, which (Suárez-figueroa and Gómez-Pérez, 2008) refers to as 'the adaptation of an ontology to a particular language and culture'. The challenges of ontology localization are analyzed in (Espinoza et al., 2009b) and a general methodology for guiding the localization process is presented. By (Cimiano et al., 2010), ontology localization can affect two different layers: the lexical layer (labels, definitions and accompanying documentation in natural language) and the conceptualization itself. Thus, the translation of concept labels we conduct can be seen as a subtask of ontology localization targeting only the lexical layer. The focus of our work does not lie in the machine translation task itself but in the intelligent use of existing resources for multilingual extension of ontologies with the aim to enhance the annotation coverage for a certain corpus. (Espinoza et al., 2009a) focus on sense disambiguation as major problem in ontology localization, while we investigate how to increase the efficiency by incorporating a corpus.

3 Overview of the approach

As explained, our main goal is to enhance the annotation coverage of a given non-English corpus by ontology translation. Using the corpus to be annotated within the translation process has three advantages:

- The translation is conducted more efficiently, since we reduce the number of translations that require a review. This is because only concepts that actually occur in the corpus are proposed as translations.
- The process results in high quality translations, because the corpus can be used to disambiguate the correct (target) translation candidate for a concept automatically.
- By facilitating a corpus, we make sure that the terms extracted as (target) translation candidates result in semantic text annotations in the end.

Figure 1 illustrates the approach: Based on the corpus information, "Läsion" is added as German translation to the ontology concept with RID38780. Now, the corpus term can be annotated, which was not possible before.

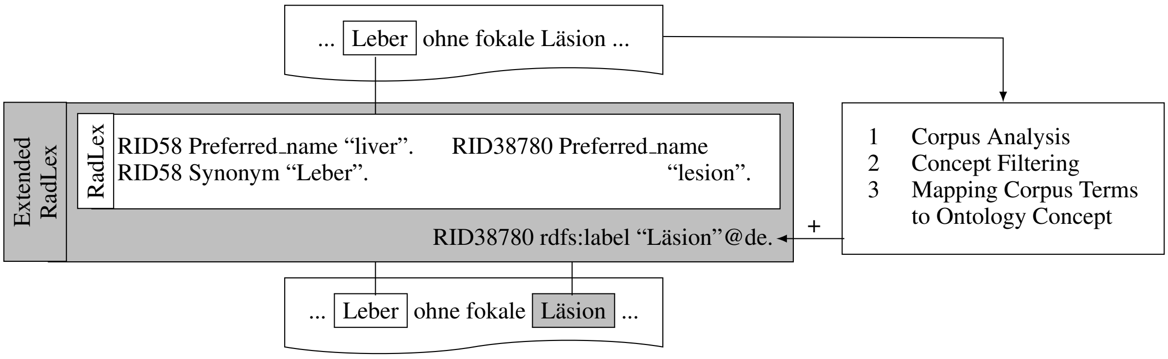
... Leber ohne fokale Läsion ...
Extended RadLex
RadLex
RID58 Preferred_name "liver".
RID58 Synonym "Leber".
RID38780 Preferred_name "lesion".
RID38780 rdfs:label "Läsion"@de.
+
1 Corpus Analysis
2 Concept Filtering
3 Mapping Corpus Terms to Ontology Concept
... Leber ohne fokale Läsion ...

Figure 1: The text *Leber ohne fokale Läsion* "Liver without focal lesion" from a large medical corpus is processed and a new translation is added to the ontology to increase the number of semantic annotations.

The system designed makes use of this rationale and implements an approach that operates in three steps (as illustrated in Figure 2) for translating the ontology vocabulary:

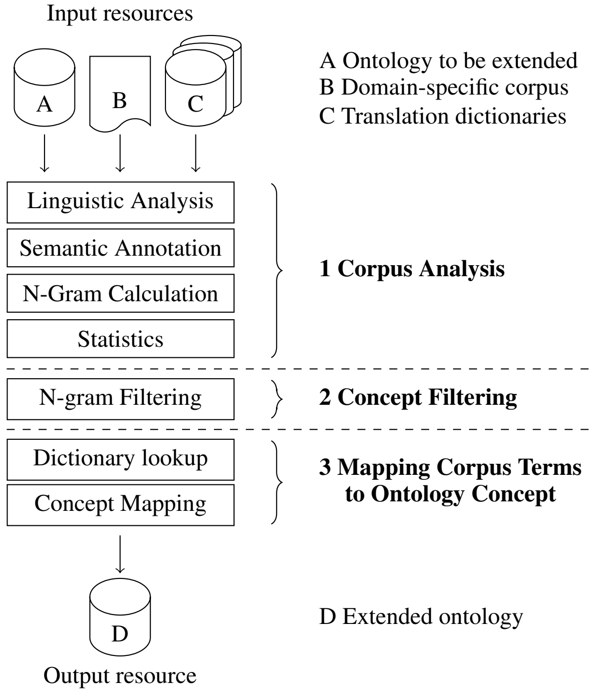

Figure 2: Processing steps in text analysis system

```
@prefix rdfs:
  <http://www.w3.org/2000/01/rdf-schema#> .
@prefix radlex:
  <http://www.owl-ontologies.com/
  Ontology1375951364.owl#> .

radlex:RID58
  rdfs:subClassOf radlex:RID13419 ;
  radlex:Preferred_name "liver"^^xsd:string ;
  radlex:Synonym "Leber"^^xsd:string ;

radlex:RID38780
  rdfs:subClassOf radlex:RID34300 ;
  radlex:Preferred_name "lesion"^^xsd:string ;
  rdfs:label "Läsion"@de.
```

Figure 3: (Incomplete) RDF representation of the RadLex concept radlex:RID58 with German translation ‘Leber’ as currently maintained as radlex:Synonym and concept radlex:RID38780 with translation ‘Läsion’ and proposed representation using rdfs:label and language tags

1 Corpus Analysis The initial processing step is designed to make use of the corpus to find the high frequency terms. Using this resource allows us to customize our approach for the required application scenario. Its content is used to digest the most relevant concepts for translation and determine the correct translation option. The processing incorporates linguistic and statistical NLP techniques to extract terms in target language with high frequency from the corpus.

2 Concept Filtering As the list of extracted terms still includes terms without semantic importance, we introduce this step in order to reduce the list. This includes the removal of terms with certain technical characters but also those with special linguistic structures, which makes the approach more efficient.

3 Mapping Corpus Terms to Ontology Concepts Our approach is targeted to translate only existing ontology concepts. Thus, we need a mechanism to map the terms of the corpus to the ontology concepts. We do this by employing state-of-the-art dictionary lookups: The English dictionary equivalences of the German corpus terms are used to find ontology concepts with the same English labels. Then, the (corpus) term is added as translation to the matching ontology concept as non-English label. The resulting translated ontology can be used in subsequent NLP-based applications and is able to serve the need for non-English texts.

In the end, the ontology will be extended with translations. In our case, the RadLex ontology currently maintains translations as synonyms, but we propose the usage of rdfs:label and language tags as shown in Figure 3. The introduced steps are described in detail in the following sections.

4 Corpus-Based Analysis and Concept Filtering

4.1 Corpus Description

One of the core resources for the approach is a domain-specific corpus. Combined with the ontology to be translated it serves several purposes: On the one hand, based on IE techniques we find and extract translations from the corpus in order to extend the ontology's vocabulary. Further, we use the corpus as semantic annotation target, which is annotated with ontology terms. The language-specific translations used for semantic annotation were found before with the help of the corpus itself. For the study, we use a corpus of 2,713 radiology reports (from 27 different readers[1]) of lymphoma patients containing the findings and evaluation sections.

4.2 Linguistic Analysis

This initial analysis includes several steps that enable a statistical analysis of the textual context. Each of the processing steps is implemented as a single UIMA annotator and integrated into an overall pipeline.

First, semantic information units such as dates and measurements are recognized using regular expressions. Medical language is rich in abbreviations. Particularly radiologists make use of them, because they allow an efficient reporting. Therefore, as second step, we build an abbreviation recognition and extension algorithm on a simple dictionary. The third linguistic task is the determination of the basic processing units: (1) tokens and (2) sentences. Tokens are split employing the spaces and '-' in the text, hence no compound splitting is conducted. While token splitting is a rather simple task, sentence splitting requires disambiguation facilities. Indicators like '?','!',';','.' are used to determine sentence ends. However, the full stop determines sentence ends only if they are not part of a measurement, date or abbreviation. As a fourth step, stopwords are removed from the documents to reduce the content to only relevant tokens. Available language-dependent stopword lists are employed. Finally, each of the tokens in the text is stemmed with the German version of the Porter stemmer. (Porter, 1997)

4.3 Semantic Annotation

Since most ontologies are already *partially* translated, we make use of this fact and semantically annotate concepts and exclude them in the subsequent filter process (Section 4.6). The annotator implementation is based on the UIMA ConceptMapper (Tanenblatt et al., 2010). The annotation dictionary is built from the preferred names and synonyms in the RadLex ontology (as shown in Figure 3). Our concept mapper combines the stems of the dictionary terminology and the stems of the text tokens and annotates the matches with the ontology information. If a dictionary term consists of more than one token, an annotation is created if all of its stems are contained in a single sentence of the corpus. That is also how single tokens can be assigned more than one annotation.

4.4 N-Gram Calculation

After the linguistic processing of the preceding steps, the actual term extraction can be performed. In this initial work, we limit the length of n-grams to three because of performance reasons. Furthermore, we define that the individual tokens of an n-gram have to co-occur within the same sentence. The output of this step is a list of terms in target language that are candidates for ontology translation.

4.5 Statistics

The n-grams relevant for translation are determined by their frequency in the corpus. Based on the stems, the frequency of each n-gram is calculated according to their (co-)occurrence. The individual (co-)occurrence count of the terms is used for ordering of the terms, whereas the most frequent occurring term is ranked top.

4.6 N-Gram Filtering

The list of high frequency terms still contains several terms with tokens representing special characters and sentence ends (like '.', '?', '<', '>', '/') or semantic classes meaningless for ontology extension (like dates, measurements, negation, and image references). Since the overall aim is to identify concepts that should be added as translations to the ontology, we remove occurrences of these information units that are very specific and without ontology importance. Also, if the term contains numbers, this precise and

[1] In the radiology domain, readers are physicians, who read and interpret radiology images and produce the reports analyzed in this work.

rather technical information is removed from the n-gram list. The resulting list contains terms we would like to add as labels to respective ontology concepts if available.

5 Mapping Corpus Terms to Ontology Concepts

Based on the list of terms ranked by their frequency, we identify ontology concepts, whose translations have a high impact on annotation coverage for the respective corpus. We assume that each ontology concept has at least one label in the source language, in our case in English. In the following, we describe our language resources employed in the approach and the mapping procedure.

5.1 Translation dictionaries

For this work, we used German-English translations from Dict.cc[2] and multilingual information from DBpedia to create two dictionaries.

1. **Medical Dictionary: 60,082 different English entries**
   Dict.cc contains specialized dictionaries for 130 different subjects. For our medical dictionary, we collected all entries from the specialized dictionaries with subjects ‘anatomy’, ‘biology’, ‘chemistry’, ‘medicine’, ‘pharmacy’, and ‘medical engineering and imaging’. Additionally, we retrieved all medically relevant concepts from DBpedia that have an English and a German or Latin label (about 9,500 concepts). More precisely, we used the DBpedia ontology (Bizer et al., 2009) to retrieve all concepts of type dbp:AnatomicalStructure[3], dbp:Disease, dbp:Drug, dbp:ChemicalSubstance and subclasses (see SPARQL query in Figure 4).

2. **General Dictionary: 623,294 different English entries**
   The general dictionary is the complete English-German Dict.cc dictionary without restriction to a specific subject.

```
PREFIX rdfs:      <http://www.w3.org/2000/01/rdf-schema#>
PREFIX dbp:       <http://dbpedia.org/ontology/>
PREFIX dbpedia2:  <http://dbpedia.org/property/>
SELECT ?s ?labelEn ?labelDe ?labelLat
WHERE {
    ?s a ?type ;
       rdfs:label ?labelEn .
    FILTER ( ?type = dbp:AnatomicalStructure
        || ?type = dbp:Disease
        || ?type = dbp:Drug
        || ?type = dbp:ChemicalSubstance )
    FILTER ( lang(?labelEn) = "en" )
    OPTIONAL { ?s dbpedia2:latin ?labelLat }
    OPTIONAL { ?s rdfs:label ?labelDe .
        FILTER ( lang(?labelDe) = "de" ) }
    FILTER( bound(?labelDe) || bound(?labelLat) )
}
```

Figure 4: SPARQL query to retrieve English-German and English-Latin translations from DBpedia using the SPARQL endpoint at http://dbpedia.org/sparql.

5.2 Ontology concept translation

The mapping of given corpus terms to corresponding ontology concepts as translations involves two sub steps.

1. **Dictionary Lookup** For all occurrences of a term, we try to find English options in our dictionaries. If no complete lookup option is found for a n-gram, we try to find a lookup option in the dictionary for each single token to combine them into a complete English n-gram. E.g. the corpus term “Läsion” is translated to “lesion” using the medical dictionary.

[2] http://www.dict.cc/

[3] We use the prefix notation dbp for http://dbpedia.org/ontology/AnatomicalStructure

2. **Concept Mapping** The list of English lookup options from the first step is used to find ontology concepts, whose (English) labels match the dictionary lookup. We find that the ontology concept with RID38780 is assigned the given preferred name "lesion". If a match is found, the German n-gram that resulted in the match ("Läsion") is regarded as probable translation. In order to increase the quality of the translation, an expert review is conducted at this time. This is the only manual step in the whole translation process. After the review, the n-gram is inserted as new RDF triple for the respective ontology concept. In RadLex translations are currently maintained as synonyms. However, as this modeling of translations as synonyms does not represent the correct semantics and misses the important language information, we propose to use rdfs:label for translations added by a corresponding language tag. Thus, for the example we insert "Läsion" as additional German label to the ontology concept (see Figure 3).

6 Evaluation

6.1 Resources

The evaluation of our system is based on the RadLex ontology and a corpus of 2,713 radiology reports of lymphoma patients. We use the OWL DL version of RadLex3.9.1 from NCBO BioPortal. This version contains 42,321 concepts, which all have an assigned (English) preferred name and few additionally synonyms. The German translations are represented as synonyms. Most of the German labels were added in 2009, when a first German version was created. Even though the number of concepts is growing significantly (RadLex3.9 contained 34,899), the number of concepts with non-English labels is not evolving the same way. Thus, in RadLex3.9.1 less than 25% of the 42,321 concepts have German labels.

Proposed translations for ontology concepts - as output of the described automatic approach - are evaluated by a clinical expert. We restricted the corpus terms translated to those occurring at least two times. The whole process results in a list of 742 German labels proposed for ontology extension. The expert classified these translations as correct or incorrect. In order to assist the expert in better understanding of the ontology concept to be extended, we provide information on the preferred name, synonyms as well as preferred names of the next two super classes.

This list of evaluated translations is analyzed in detail using three dimensions: First, we analyze how the choice of the dictionary influences the translation outcome. Second, we figure out how the term length and the processing of multi-word terms influences the translation results. Third, the correct translations are added to an extended RadLex ontology. We compare the annotation results using the initial and extended RadLex version. We apply *accuracy* as evaluation measure, which is the proportion of correct translations in the system-proposed set.

6.2 Evaluation of the Translation Services

As described in Section 5.1, we use two different dictionaries. As expected, the accuracy of the medical dictionary is significantly higher than the accuracy of the general dictionary (see Table 1(a)). This is because in many cases only the domain-specific dictionary contains the correct lookup entry for the terms. Nevertheless, the general dictionary is necessary, because RadLex contains also general language terms like 'increased' or 'normal'. Combining the two dictionaries accuracy reaches 75.2%.

6.3 Evaluation of the N-Gram Length

If we take a closer look at n-gram distribution of terms, we see that we translate mainly single words (1-grams), while 2-grams and 3-grams are translated less often. However, the accuracy of 3-grams reaches excellent values (see Table 1(b)). Nevertheless, the translation of n-grams is of high importance, as most of the ontology concepts in the biomedical domain have multiword labels. In particular, labels of anatomical entities are multiword terms; in RadLex they can grow to 10-grams. Consider for example 'Organ component of lymphatic tree organ' or 'Tendon of second palmar interosseous of left hand'.

Thus, a more sophisticated multiword translation is needed to enhance the number of translations for n-grams. For us, the improved handling of stopwords is the main focus in future work: While we remove stopwords in the n-grams, ontology concepts that contain stopwords prevent a match.

Table 1: Evaluation of translation outcomes by choice of dictionary and term length. *Proposed* denotes the number of German labels translated and added to the ontology. *Correct* denotes the subset of translations evaluated by the expert as correct.

(a) Evaluation by translation dictionary

	Translations		
	Proposed	**Correct**	**Accuracy**
medical dict	258	240	0.9302
general dict	484	318	0.6570
both dicts	742	558	0.7520

(b) Evaluation by n-gram length

	Translations		
	Proposed	**Correct**	**Accuracy**
1-grams	609	451	0.7406
2-grams	118	92	0.7797
3-grams	15	15	1.0000

Table 2: Comparison of the annotation coverage using RadLex3.9.1 and the extended version. Total number of tokens of the corpus: 346,963.

		extended	
	RadLex3.9.1	**RadLex3.9.1**	
Tokens with annotation	94,914	147,982	+0.5591
Annotation Coverage	27.36 %	42.65 %	+0.5591
Tokens without annotation	252,049	198,981	- 0.2105
Number of annotations	133,156	204,491	+0.5357

6.4 Extension of RadLex and Evaluation of Annotation Coverage

From Table 1(a), one can see that we correctly translated 558 RadLex concept labels using both dictionaries. After the expert review, we added the (German) terms of these correct matches as labels to 433 distinct RadLex concepts. I.e., some concepts were assigned more than one additional German label. We refer to the new ontology as the *extended RadLex*. For the analysis of how the added translations influence the number of annotations, we conducted two annotation processes. Both the original and the extended RadLex versions were used to semantically annotate the corpus using the annotator described in Section 4.3. The measure to indicate the annotation success is *annotation coverage*, which denotes the relative amount of tokens for which at least one annotation exists. Table 2 shows that we are able to enhance the annotation coverage by about 56% by adding only 558 translations. This shows the effectiveness of the approach. A comparison indicator of these numbers deliver English texts: In (Woods and Eng, 2013) an annotation rate of 62 % was observed for English chest radiography reports. Despite the restrictiveness of the comparison, we see that an annotation coverage of 42.65 % is high considering that only about 25 % of the extended RadLex's concepts have a German label.

6.5 Limitations

Due to the characteristics of our approach, the outcome of the increased annotation coverage is specific for the corpus used: Even though the reports come from 27 different readers, the vocabulary of the evaluated corpus is specific to one disease and thus limited to a certain degree. Because the vocabulary differentiates in other corpora, the application of the translation added for texts describing other diseases or reports may not result in increases of the annotation coverage as shown. For other corpora, one has to run our approach a second time using the new corpus and add further concepts to obtain a similar annotation coverage. However, we expect the additional effort needed to get smaller over time.

7 Conclusion

We propose a method to make ontologies usable for multilingual semantic annotation of texts by automatically extending them with translations, without the need to invest much effort in a full translation. We believe that our approach is able to unlock the high potential of existing ontologies also for low resourced languages. We address the key problem of identifying those concepts that are worth translating by defining the increase of annotation coverage for a given corpus as the main target. Although it might seem intuitive to apply an English corpus to identify the most frequent terms and their (source) ontology concepts to translate, we do not pursue this approach. Especially when dealing with a domain-specific language, translations are often ambiguous. As the English corpus does not help picking the correct (target) translation candidate, we decided to start the other way around and facilitate a corpus in target language. We show the high quality and efficiency of the approach by translating medical terms from English to German. According to the evaluation results, a better treatment of n-grams shows the biggest potential for enhancement of the approach. Sophisticated linguistic algorithms for the translation, which incorporate the ontology context, can increase the matching of the multi-word terms. In future work, we plan to evaluate our approach using other ontologies from the BioPortal.

Acknowledgements

This research has been supported in part by the KDI project, which is funded by the German Federal Ministry of Economics and Technology under grant number 01MT14001. We thank Dict.cc for providing us with the dictionaries.

References

Christian Bizer, Jens Lehmann, Georgi Kobilarov, Sren Auer, Christian Becker, Richard Cyganiak, and Sebastian Hellmann. 2009. {DBpedia} - a crystallization point for the web of data. *Web Semantics: Science, Services and Agents on the World Wide Web*, 7(3):154 – 165.

Philipp Cimiano, Elena Montiel-Ponsoda, Paul Buitelaar, Mauricio Espinoza, and Asunción Gómez-Pérez. 2010. A note on ontology localization. *Applied Ontology*, 5(2):127–137.

M Espinoza, A Gómez-Pérez, and E Montiel-Ponsoda. 2009a. Multilingual and Localization Support for Ontologies. *The Semantic Web Research and Applications*, 5554:821–825.

Mauricio Espinoza, Elena Montiel-Ponsoda, and Asunción Gómez-Pérez. 2009b. Ontology localization. In *Proceedings of the Fifth International Conference on Knowledge Capture*, pages 33–40, New York. ACM.

Thomas R Gruber. 1993a. Toward Principles for the Design of Ontologies Used for Knowledge Sharing. *International Journal Human-Computer Studies 43*, pages 907–928.

Thomas R. Gruber. 1993b. A translation approach to portable ontology specifications. *Knowl. Acquis.*, 5(2):199–220, June.

Curtis P. Langlotz. 2006. Radlex: A new method for indexing online educational materials. *RadioGraphics*, 26(6):1595–1597. PMID: 17102038.

S.M. Meystre, G.K. Savova, K.C. Kipper-Schuler, and J.F. Hurdle. 2008. Extracting information from textual documents in the electronic health record: A review of recent research. *Yearbook of Medical Informatics*, pages 128–144.

Natalya F. Noy, Nigam H. Shah, Benjamin Dai, Michael Dorf, Nicholas Griffith, Clement Jonquet, Michael J. Montegut, Daniel L. Rubin, Cherie Youn, and Mark A. Musen. 2008. Bioportal: A web repository for biomedical ontologies and data resources [demonstration].

M. F. Porter. 1997. Readings in information retrieval. chapter An Algorithm for Suffix Stripping, pages 313–316. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

Mari Carmen Suárez-figueroa and Asunción Gómez-Pérez. 2008. First Attempt towards a Standard Glossary of Ontology Engineering Terminology. In *Proceedings of the 8th International Conference on Terminology and Knowledge Engineering (TKE2008).*

Michael Tanenblatt, Anni Coden, and Igor Sominsky. 2010. The conceptmapper approach to named entity recognition. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Ryan W. Woods and John Eng. 2013. Evaluating the Completeness of RadLex in the Chest Radiography Domain. *Academic Radiology*, 20(11):1329–1333.