Mining translations from the web of open linked data

John Philip M^cCrae
University of Bielefeld
jmccrae@cit-ec.uni-bielefeld.de

Philipp Cimiano
University of Bielefeld
cimiano@cit-ec.uni-bielefeld.de

Abstract

In this paper we consider the prospect of extracting translations for words from the web of linked data. By searching for entities that have labels in both English and German we extract 665,000 translations. We then also consider a linguistic linked data resource, lemonUby, from which we extract a further 115,000 translations. We combine these translations with the Moses statistical machine translation, and we show that the translations extracted from the linked data can be used to improve the translation of unknown words.

1 Introduction

In recent years there has been a massive explosion in the amount and quality of data available as linked data on the web. This data frequently describes entities in multiple languages and as such can be used as a source of translations. In particular, the web contains much data about named entities, such as locations, films, people and so forth, and often these named entities have translations in many languages. In this paper, we address the question of what can be achieved by using the large amounts of data available as multilingual Linked Open Data (LOD). As state-of-the-art statistical machine translation systems (Koehn, 2010) are typically trained on outdated or out-of-domain parallel corpora such as on the transcripts of the European Parliament (Koehn, 2005), we expect to increase the coverage of domain-specific terminology.

In addition, there has recently been a move towards the publication of of language resources using linked data principles (Chiarcos et al., 2011), which can be expected to lead to a significant increase the availability of information relevant to NLP on the Web. In particular, the representation of legacy resources such as Wiktionary and OmegaWiki on the web of data should ameliorate the process of harvesting translations from these resources.

We consider two sources for translations from linked data: firstly, we consider mining labels for concepts from the data contained in the 2010 Billion Triple Challenge (BTC) data set, as well as DBpedia (Auer et al., 2007) and FreeBase (Bollacker et al., 2008). Secondly, we mine translations from lemonUby (Eckle-Kohler et al., 2013), a resource that integrates a number of distinct dictionary language resources in the *lemon* (Lexicon Model for Ontologies) format (McCrae et al., 2012), which is a model for representing rich lexical information including forms, sense, morphology and syntax of ontology labels. We then consider the process of including these extra translations into an existing translation system, namely Moses (Koehn et al., 2007). We show that we can extract many translations which are complementary to those found by the statistical machine translation system and that these translations improve the translation performance of the system.

2 Mining translations from the linked open data cloud

Obtaining translations from the Linked Open Data (LOD) cloud is a non-trivial task as there are many different properties used to specify the label or name of resources on the LOD. The standard method of identifying the language of a label is by means of the `xml:lang` annotation, which should be an ISO-639 code, such as “en” or “eng” for English. As noted by Ell et al. (2011), very little of the data found on the web actually has such a language tag. Further, as the labels are typically short it is very difficult to infer the language reliably based on its surface form. As such we are compelled to rely on the language tags, and this means that from the data we can only recover a small amount of what may be available. In par-

Proceedings of the Joint Workshop on NLP&LOD and SWAIE, pages 8–11,
Hissar, Bulgaria, 12 September 2013.

ticular, out of the Billion Triple Challenge data we recover approximately 398 million labels in English but only 144,000 labels in German. Unsurprisingly, given the dominance of English on the web of data (Gracia et al., 2012), we find that there are many more labels in English. We then filter these two sets so that we keep only URIs for which there is both an English and a German label. Among this data is a significant number of long textual descriptions, which are very unlikely to be useful for translation, such that we also filter out all labels whose length is more than 10 characters. This filter was necessary to reduce the amount of noisy "translations". We do not filter according to any particular property, e.g., we do not limit ourselves to the RDFS label property, but in the case where we have multiple labels on the same entity we select the RDFS label property as a preference.

In addition to the BTC data, we also include two large resources for which there are multilingual labels in their entirety. These resources are DBpedia[1] and FreeBase[2] as these resources contain many labels in many languages. As the resources are consistent in the use of the rdfs:label property, we extract translations by looking at the rdfs:label property and the language tag. In Table 1 we see the number of translations that we extract from the three resources. We see that we extract fewer resources from the BTC data but a large and reasonable number of translations from the other two resources

3 Translations mined from linguistic linked data

For finding translations from linguistic linked data, we focus on the lemonUby (Eckle-Kohler et al., 2013) resource, which is a linked data version of the UBY resource (Gurevych et al., 2012). This resource contains *lemon* versions of a number of resources in particular:

- FrameNet (Baker et al., 1998)
- OmegaWiki [3]
- VerbNet (Schuler, 2005)
- Wiktionary [4]
- WordNet (Fellbaum, 2010)

Resource	Translations
OmegaWiki (English)	56,077
OmegaWiki (German)	55,990
Wiktionary (English)	34,421
Wiktionary (German)	43,212
All	114,644
lemonUby and cloud	777,173

Table 2: Number of translations extracted for linguistic linked data resource lemonUby

Out of these resources, FrameNet, VerbNet and WordNet are monolingual English resources, so we focus on the OmegaWiki and Wiktionary part of the resources. LemonUby contains direct translations on the lexical senses of many of the resources (see Figure 1 for an example). The total number of translations extracted for each subresource is given in Table 2 as well as the results in combination with the number of translation extracted from labels in the previous section.

4 Exploiting mined translations mined from linked data

In order evaluate the utility of the translations extracted from the linked data cloud, we integrate them into the phrase table of the Moses system (Koehn et al., 2007) trained on Europarl data (Koehn, 2005). We used the system primarily in an 'off-the-shelf' manner in order to focus on the effect of adding the linked data translations.

Moses uses a log-linear model as the baseline for its translations, where translations are generated by a decoder and evaluated according to the following model:

$$p(\mathbf{t}|\mathbf{f}) = \exp(\sum_i \phi_i(\mathbf{t},\mathbf{f}))$$

Where $\mathbf{t}$ is the candidate translation sentence, $\mathbf{f}$ is the input foreign text and ϕ_i are scoring functions. In the phrase-based model the translation is derived compositionally by considering phrases and their translations stored in the so called *phrase table* of the Moses system.

The main challenge in integrating these translations derived from linked data lies in the fact that they lack a probability score. For each translation pair (a, b) derived from the linked data, we distinguish two cases: If the translation is already in the phrase table, we add a new feature that is set to 1.0 to indicate that the translation was found

[1] We use the dump of the 3.5 version

[2] The dump was downloaded on June 21st 2013

[3] http://omegawiki.org

[4] http://www.wiktionary.org

Resource	English Labels	German Labels	Translations
BTC	398,902,866	144,226	51,756
DBpedia	7,332,616	590,381	540,134
FreeBase	41,261,806	1,654,254	259,923
All	447,497,288	2,338,861	665,910

Table 1: The number of labels and translations found in the linked data cloud by resource

```
<OW_eng_LexicalEntry_0#CanonicalForm> lemon:writtenRep "rain"@eng.
<OW_eng_LexicalEntry_0> lemon:canonicalForm <OW_eng_LexicalEntry_0#CanonicalForm> ;
  lemon:sense <OW_eng_Sense_0> .
<OW_eng_Sense_0> uby:equivalent "schiffen"@deu ,
  "regnen"@deu .
```

Figure 1: An example of the relevant data for a *lemon* lexical entry from the OmegaWiki English section of lemonUby

	BLEU	Evaluator 1	Evaluator 2
Baseline	11.80	36%	34%
+LD	11.78	64%	64%

Table 3: The comparative evaluation of the translations with and without linked data

from the Linked Data Cloud. If the translation was not in the phrase table, we add a new entry with probability 1.0 for all scores and the feature for linked data set to 1.0. For all other translations, the feature indicating provenance from the Linked Data Cloud is set to 0.0. The weights for the log-linear model are learned using the MERT system (Och, 2003). As such we do not use the linked data itself to choose between different translation candidates but rely on the methods built into the machine translations system, in particular the language model.

5 Results

We extracted the baseline phrase table, reordering and language model from version 7 of the EuroParl corpus translating from English to German. In order to evaluate the impact of Linked Data translations on translation quality, we rely on the News Commentary 2011 corpus provided as part of the WMT-12 translation task (Callison-Burch et al., 2012). We found that 25,688 translations from the linked data were relevant to this corpus of which 22,291 (87%) were out of the vocabulary of EuroParl. We used MERT to learn the parameters of the model and observed that the weighting for the linked data feature was negative, indicating that the translations from the linked data improved the translation quality almost exclusively in the case that the machine translation system did not have an existing candidate. We then generated all translations for the baseline system without any linked data translations and for the system augmented with all the linked data translations. Out of 3,003 translations in the test set, we found 346 translations which were changed by the introduction of linked data translation. For each translation, we performed a manual evaluation with two evaluators. They were both presented with 50 translations, one with linked data and one without linked data and asked to choose the best one ("no opinion" was also allowed). The translations were presented in a random order and there was no indication which system they came from so this experiment was performed blind. The evaluators were a native English speaker, who is fluent in German, and a native German speaker, who is fluent in English, and had a Cohen's Kappa Agreement of 0.56. In addition, we calculated BLEU (Papineni et al., 2002) scores. The results are presented in Table 3.

The results show that there is very little change in BLEU scores but the manual evaluation reveals that there was an improvement in quality of the translations. We believe the BLEU scores did not correlate with the manual evaluation due to the fact that many of the translations harvested from the linked data cloud were longer on the German side than the English side, for example the English "RPG" was translated as "Papier-und-Bleistift Rollenspiele", which was not in the reference translation.

6 Conclusion

In this paper we investigated the impact of integrating translations harvested from the Linked Open Data cloud into a state-of-the-art statistical machine translation system. We have shown that it is possible to harvest a large number of translations from the LOD. Furthermore, we found that the task was further enabled by the current growth in linguistic linked data represented in models such as *lemon*. We then integrated these extracted translations into the phrase table of a statistical machine translation system and found that the usage of linked data was most appropriate for terms that were out of the vocabulary of the machine translation system. One of the key challenges in extracting and exploiting such translations is to appropriately capture the context of these translations allow for selecting what kind of linked data may be effective for a given translation task. This will be addressed in future work.

Acknowledgments

This work was funded by the PortDial project under the EU Seventh Framework Programme (FP7), grant number 296170. We would like to thank Sebastian Walter for his help in the experiments.

References

Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. 2007. Dbpedia: A nucleus for a web of open data. In *The Semantic Web, 6th International Semantic Web Conference, 2nd Asian Semantic Web Conference*, pages 722–735. Springer.

Collin F Baker, Charles J Fillmore, and John B Lowe. 1998. The Berkeley FrameNet project. In *Proceedings of the 17th International Conference on Computational Linguistics*, pages 86–90.

Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. 2008. Freebase: a collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, pages 1247–1250.

Chris Callison-Burch, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia. 2012. Findings of the 2012 workshop on statistical machine translation. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 10–51.

Christian Chiarcos, Sebastian Hellmann, and Sebastian Nordhoff. 2011. Towards a linguistic linked open data cloud: The open linguistics working group. *TAL*, 52(3):245–275.

Judith Eckle-Kohler, John McCrae, and Christian Chiarcos. 2013. lemonUby-a large, interlinked, syntactically-rich resource for ontologies. *Semantic Web Journal, submitted. special issue on Multilingual Linked Open Data.*

Basil Ell, Denny Vrandečić, and Elena Simperl. 2011. Labels in the web of data. In *The Semantic Web, 10th International Semantic Web Conference*, pages 162–176.

Christiane Fellbaum. 2010. *WordNet*. Springer.

Jorge Gracia, Elena Montiel-Ponsoda, Philipp Cimiano, Asunción Gómez-Pérez, Paul Buitelaar, and John McCrae. 2012. Challenges for the multilingual web of data. *Web Semantics: Science, Services and Agents on the World Wide Web*, 11:63–71.

Iryna Gurevych, Judith Eckle-Kohler, Silvana Hartmann, Michael Matuschek, Christian M Meyer, and Christian Wirth. 2012. Uby: A large-scale unified lexical-semantic resource based on LMF. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 580–590.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, et al. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the ACL*, pages 177–180.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *MT summit*, volume 5.

Philipp Koehn. 2010. *Statistical Machine Translation*. Cambridge University Press.

John McCrae, Guadalupe Aguado-de Cea, Paul Buitelaar, Philipp Cimiano, Thierry Declerck, Asunción Gómez-Pérez, Jorge Gracia, Laura Hollink, Elena Montiel-Ponsoda, Dennis Spohr, et al. 2012. Interchanging lexical resources on the semantic web. *Language Resources and Evaluation*, 46(4):701–719.

Franz Josef Och. 2003. Minimum error rate training in statistical machine translation. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1*, pages 160–167.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318.

Karin Kipper Schuler. 2005. *VerbNet: A broad-coverage, comprehensive verb lexicon.* Ph.D. thesis, University of Colorado Boulder.